American Journal of Epidemiology

Vol. 160, No. 1
Printed in U.S.A.
DOI: 10.1093/aje/kwh175

Imputations of Missing Values in Practice: Results from Imputations of Serum Cholesterol in 28 Cohort Studies

Federica Barzi and Mark Woodward

From The George Institute for International Health, Camperdown, New South Wales, Australia.

Received for publication March 24, 2003; accepted for publication January 29, 2004.

Missing values, common in epidemiologic studies, are a major issue in obtaining valid estimates. Simulation studies have suggested that multiple imputation is an attractive method for imputing missing values, but it is relatively complex and requires specialized software. For each of 28 studies in the Asia Pacific Cohort Studies Collaboration, a comparison of eight imputation procedures (unconditional and conditional mean, multiple hot deck, expectation maximization, and four different approaches to multiple imputation) and the naive, complete participant analysis are presented in this paper. Criteria used for comparison were the mean and standard deviation of total cholesterol and the estimated coronary mortality hazard ratio for a one-unit increase in cholesterol. Further sensitivity analyses allowed for systematic over- or underestimation of cholesterol. For 22 studies for which less than 10% of the values for cholesterol were missing, and for the pooled Asia Pacific Cohort Studies Collaboration, all methods gave similar results. For studies with roughly 10–60% missing values, clear differences existed between the methods, in which case past research suggests that multiple imputation is the method of choice. For two studies with over 60% missing values, no imputation method seemed to be satisfactory.

bias; cholesterol; coronary disease; hazard rate; imputation; meta-analysis; missing data; mortality

Abbreviations: APCSC, Asia Pacific Cohort Studies Collaboration; CHD, coronary heart disease; DBP, diastolic blood pressure; EM, expectation maximization; MI, multiple imputation; SBP, systolic blood pressure.

Researchers modeling medical data often encounter the problem of missingness regarding one or more of the variables under investigation. The most common approach is to delete those observations with missing values, leading to a complete participant analysis. This approach not only wastes data and reduces power but also produces biased estimates when the group excluded is a selective subsample from the study population, that is, when the values are not missing completely at random (1, 2). An alternative is to use one of the many methods available for imputing the missing values. Of these, the method of multiple imputation (MI) (3) is attractive since theoretical and simulation studies have shown that it yields estimates with good statistical properties, such as efficiency and validity, when a correct model is specified for the imputation (1). However, MI is not well understood in the medical community and requires advanced software typically implementing the algorithms of Shafer (4, 5). Consequently, alternative, simpler methods of imputation are more commonly adopted at present.

To our knowledge, systematic comparisons of methods of imputation using real-life meta-data are lacking. In this article, we compare the naive, complete participant method and several imputation methods, including different implementations of MI, for dealing with missing values of total cholesterol in 28 cohorts within the Asia Pacific Cohort Studies Collaboration (APCSC) (6, 7). Most importantly, we compare these methods in an investigation of association between cholesterol and coronary heart disease (CHD) mortality. Since the APCSC studies have a wide variation in cholesterol distributions as well as very different rates of missingness for cholesterol, they provide an opportunity to apply different imputation methods to diverse epidemiologic data.

All of the imputation methods considered here assume that data are missing at random (1, 8). That is, the probability of

Correspondence to Federica Barzi, Asia Pacific Cohort Studies Collaboration Secretariat, The George Institute for International Health, PO Box M201, Missenden Road, Camperdown, NSW 2050, Australia (e-mail: fbarzi@thegeorgeinstitute.org).

a value being missing does not depend on the unobserved data, although it may depend on observed data, which can thus provide information about the missing values and a basis for imputation. We further explore the potential effects of a mechanism for handling missingness that does not satisfy the assumption of missing at random (9, 10) by allowing for systematic over- or underestimation of cholesterol.

THE APCSC

APCSC collects individual participant data from studies in the Asia Pacific region, with the major aims of providing reliable information about the determinants of cardiovascular diseases and of comparing the different population groups within the region (6). When this article was written, the collaboration included 37 studies with a total of 419,639 male and female participants aged 20 years or older followed up for a median of 5.5 years. To address the aims of APCSC, the participants were grouped into four geographic areas: Australia and New Zealand, China-Taiwan-Singapore (China), Japan, and Korea.

As a result of the collaboration's inclusion criteria, all studies have complete participant information on age, sex, systolic blood pressure (SBP), and diastolic blood pressure (DBP); after data collection, we additionally excluded subjects for whom information on survival time or causes of death were missing (7). Information was also collected on body mass index (weight (kg)/height $(m)^2$), total cholesterol, and the binary variables smoking and history of diabetes. Some studies did not collect information on all of these variables. When they did, some values were missing for unknown but probably diverse reasons, such as unwillingness to answer certain questions.

Table 1 shows that cholesterol was measured at baseline in 28 studies (numbered in the table in rank order of cholesterol missingness) involving 385,975 persons. Altogether, cholesterol values were missing for 33,664 (9 percent) subjects, but the percentage varied considerably across studies, ranging from 0 percent to 69 percent (study 28). Although cholesterol was almost always measured in Korea, values were missing for 2 percent, 3 percent, and 32 percent of participants in Australia and New Zealand, Japan, and China, respectively. Overall, 5 percent of body mass index values, 51 percent of diabetes values, and 22 percent of smoking values were missing. Excluding the Korean Medical Insurance Corporation study, which contributed by far the largest number of subjects (48 percent) to the collaboration, cholesterol values were missing for 17 percent, diabetes values were missing for 7 percent, body mass index values were missing for 8 percent, and smoking values were missing for 0.5 percent of the participants.

Overall, 90 cases (5 percent) of CHD death were lost by analyzing only those subjects for whom cholesterol was measured (complete participant analyses). Not surprisingly, the percentage lost per study tended to increase with increasing rate of nonresponse, although the relation was not strictly monotonic.

IMPUTATION METHODS AND PREDICTORS

Table 2 summarizes the imputation methods used in this article to "fill in" the missing values for cholesterol. Each will be described in detail subsequently. All but one uses other variables to predict the missing values of cholesterol. Some methods require assumptions to be made about the distribution of cholesterol and its predictor variables. Furthermore, some (the "multiple" methods) proceed by imputing each missing value several (m) times, therefore generating several independent, completed data sets. Each completed data set is analyzed by using standard methods, leading to m sets of estimates and standard errors that are then combined by using Rubin's equations (1, 3, 4). The combined standard errors will include both between- and within-imputation components. Usually, m is taken to be between three and five (4), but data sets with a high rate of missingness need both more iterations and more imputations, so we chose $m = 10$ in the current analyses.

All of the imputation methods used assume that the data are missing at random, a hypothesis that cannot be verified since there is no knowledge of the unobserved data. Nevertheless, the more predictors the imputation model includes, the more the assumption of missing at random is likely to hold because the uncertainty associated with missingness is reduced (4). Imputation models should ideally include all covariates that are related to the missing data mechanism, have distributions that differ between the respondents and nonrespondents, are associated with cholesterol, and will be included in the analyses of the final complete data sets (1, 3, 4, 11). Predictors with a high rate of missingness are probably best omitted because they would increase the variability about the estimates of interest rather than provide information.

The candidate predictors in the APCSC studies were age, sex, SBP, DBP, body mass index, smoking, diabetes, survival time, CHD death, and death from any cause. Preliminary descriptive analyses indicated that cholesterol was associated with all of them, although in a different fashion across studies. Predictor variables were also found to have different distributions by missingness status, and, overall, participants for whom cholesterol values were missing were at increased risk of mortality compared with those for whom cholesterol had been measured. These differences indicate that complete participant analysis might well have generated biased results and that all of the predictors considered may be important for specifying a proper imputation model. We therefore considered models with all predictors as well as models restricted to predictors without missing values.

Mean imputations

With unconditional mean imputation ($\overline{U}$), study-specific mean cholesterol is substituted for each missing cholesterol value. Since all imputations are the same, this method will underestimate the variance for cholesterol (12) and result in test statistics that are too often significant. Correlations with other variables could also be underestimated (13, 14) because of the bias in the variance and because values are imputed independently of any predictor.

TABLE 1. Total number of subjects and total number of coronary heart disease deaths per study, by geographic area and overall, Asia Pacific Cohort Studies Collaboration*

Study	Geographic area†	Study name	Total no. of subjects	Data missing (% of total no. of subjects)				Coronary heart disease deaths	
				Cholesterol	Body mass index	Smoking	Diabetes	Total no.	% excluded with CP†
1	J	Shirakawa	4,638	0.0	100.0	0.1	100.0	44	0.0
2	C	Xi'an	1,687	0.0	100.0	0.0	100.0	35	0.0
3	K	KMIC†	183,600	0.0	0.1	46.3	100.0	114	0.0
4	J	Saitama	3,624	0.0	0.7	0.2	0.0	24	0.0
5	C	Singapore 92	3,332	0.1	0.4	0.0	0.0	22	0.0
6	J	Akabane	1,828	0.1	0.2	0.1	0.0	7	0.0
7	J	Civil Service Workers	9,318	0.1	0.1	0.8	0.0	1	0.0
8	J	Shigaraki	3,757	0.2	0.7	0.7	0.0	3	0.0
9	C	Anzhen 02	4,152	0.3	0.0	0.0	0.0	1	0.0
10	ANZ	Melbourne Cancer	41,286	0.4	0.1	0.0	0.0	161	0.0
11	J	Tanno/Soubetsu	1,977	0.4	0.1	0.3	1.0	23	0.0
12	J	Kounan Town	1,226	0.5	2.8	0.0	0.0	3	0.0
13	J	Shibata	2,350	0.9	0.9	0.0	0.0	67	0.0
14	ANZ	Fletcher Challenge	10,326	1.0	0.4	0.9	0.0	70	2.9
15	J	Aito Town	1,718	2.7	0.5	34.2	0.0	16	6.3
16	C	Huashan	1,648	3.2	1.0	0.1	0.0	3	0.0
17	J	Hisayama	1,595	3.2	5.8	0.8	0.0	50	0.0
18	C	CVDFACTS	5,729	3.3	0.5	0.0	0.0	13	7.7
19	C	Shanghai Factory Workers	9,334	3.5	100.0	0.0	0.0	86	5.8
20	ANZ	Perth	10,227	4.9	0.1	0.0	0.1	194	4.1
21	ANZ	Busselton	7,881	6.1	5.4	1.0	33.0	688	2.9
22	C	Capital Iron and Steel Company	5,255	9.1	4.2	1.1	100.0	41	2.4
23	J	Ohasama	2,240	14.9	2.0	0.0	0.0	7	42.9
24	C	Six Chinese	19,384	25.2	0.3	0.0	0.0	37	48.6
25	C	Seven Cities	37,619	52.1	0.5	0.0	0.7	79	21.5
26	C	Yunnan Tin Miner	6,570	60.1	0.2	0.0	0.0	17	70.6
27	J	Miyama	1,077	61.6	4.4	0.5	0.6	2	100.0
28	C	Fangshan	2,597	68.6	0.8	0.2	0.0	0	
Area totals									
K			183,600	0.0	0.1	46.3	100.0	114	0.0
ANZ			69,720	1.8	0.7	0.2	3.7	1,113	2.7
J			35,348	3.3	14.0	2.1	13.2	247	2.4
C			97,307	32.1	11.9	0.1	7.4	334	16.2
Total			385,975	8.7	4.5	22.3	51.3	1,808	5.0

* The percentages of missing cholesterol, body mass index, smoking, and diabetes values are also shown, as are the percentages of coronary heart disease deaths excluded from survival analyses by using cholesterol as the explanatory variable with the complete participant analysis method.

† J, Japan; C, China (China-Taiwan-Singapore); K, Korea; ANZ, Australia and New Zealand; CP, complete participant analysis; KMIC, Korean Medical Insurance Corporation.

With conditional mean imputation ($\overline{C}$), also called "cold deck," the population is cross-classified according to levels of the predictor variables, and the imputed value for anyone for whom cholesterol information is missing is taken as the observed mean cholesterol within that person's cross-class. Here, cross-classification occurred by sex, seven age catego-

TABLE 2. Summary of the imputation methods used to "fill in" the missing values for cholesterol, Asia Pacific Cohort Studies Collaboration

Abbreviation used	Imputation approach	m*	Predictors of cholesterol	Distributional assumptions†	Software used
$\overline{U}$	Unconditional mean	1		None	SAS‡ 8.2
$\overline{C}$	Conditional mean	1	Categorical: age (<30, 30–39, …, ≥80 years) and SBP‡ (<120, 120–139, …, ≥180 mmHg); binary: sex	None	SAS 8.2
MHD	Multiple hot deck	10	Categorical: age (<30, 30–39, …, ≥80 years) and SBP (<120, 120–139, …, ≥180 mmHg); binary: sex	None	STATA§ Hotdeck command
EM	Expectation maximization	1	Continuous: age, SBP, DBP‡, survival time; binary: sex; categorical: outcome (no death/CHD‡/other cause)	All normal	SAS 8.2 PROC MI
SI	Single imputation	1	Continuous: age, SBP, DBP, survival time; binary: sex; categorical: outcome (no death/CHD/other cause)	All normal	SAS 8.2 PROC MI
MI	Multiple imputation	10	Continuous: age, SBP, DBP, survival time; binary: sex; categorical: outcome (no death/CHD/other cause)	All normal	SAS 8.2 PROC MI
MI+	Multiple imputation	10	Continuous: age, SBP, DBP, BMI‡, survival time; binary: sex, diabetes, smoking status; categorical: outcome (no death/CHD/other cause)	All normal	SAS 8.2 PROC MI
$MI+_{MIX}$	Multiple imputation	10	Continuous: age, SBP, DBP, BMI, survival time; binary: sex, diabetes, smoking status; categorical: outcome (no death/CHD/other cause)	Normal for continuous; log-linear for binary and categorical	S-PLUS¶ MIX

* m, no. of complete data sets imputed.

† Conditional distribution for the predictors of cholesterol. The conditional distribution of cholesterol was always assumed to be normal.

‡ SAS, Statistical Analysis System (SAS Institute, Inc., Cary, North Carolina); SBP, systolic blood pressure; DBP, diastolic blood pressure; CHD, coronary heart disease; BMI, body mass index.

§ Stata Corporation, College Station, Texas.

¶ Insightful Corporation, Seattle, Washington.

ries, and five SBP categories. Compared with $\overline{U}$, $\overline{C}$ should improve estimation of the variance and maintain associations with the predictors used. The variance estimates will still be underestimates because no account is taken of residual (error) variance, a disadvantage also encountered with other deterministic imputation methods such as median imputation and linear regression imputation.

Multiple hot deck imputation

Hot deck imputation uses the same setup as $\overline{C}$, but persons for whom values are missing in any cross-class receive a value from a donor selected from all those without missing values in that cross-class (rather than the mean). In each cross-class, selection proceeds by first taking a random sample, with replacement, of the same size as the population without missing values; then, the required donors are randomly taken with replacement (14, 15) from the sample. In multiple hot deck imputation, the whole process is repeated several times; we used STATA (16) software to generate 10 completed data sets. This method has the advantage of introducing variability into the analysis consistent with the range of values observed. It shares with other multiple imputation approaches (such as MI) the advantage of taking account of the variability of the imputed cholesterol values across the imputations, resulting in larger standard errors reflecting the uncertainty about the missing values. In contrast, single imputation methods treat the imputed values as true observations, therefore yielding artificially small standard errors. A disadvantage of both types of deck imputations is that they require categorical, completely observed, predictor variables.

Expectation maximization (EM) imputation

EM (and also MI) requires specifying a joint probability distribution for the variable to be imputed and the predictor variables (1, 4). It provides maximum likelihood estimates in the presence of missing data. The first E step fills in the missing values based on the observed values and on initial values of the parameters of the imputation model. Then, the first M step reestimates the parameters by using the observed and imputed values. The algorithm iterates from E to M steps until the log-likelihood converges to a stationary point. The number of cycles required depends on the fraction of nonresponse data. The covariance matrix takes into account the residual variance from the regression on the E step, thereby correcting for the underestimation of variance typical of mean imputations. Implementation of EM in the current application took age, sex, SBP, DBP, survival time, and death by cause (no death/CHD/other cause) as predictor variables and assumed a normal distribution for each, condi-

tional on all other variables. The imputation model for cholesterol, conditional on the set of predictors (P), is Cholesterol$|P = \alpha + \beta_1$age $+ \beta_2$SBP $+ \beta_3$DBP $+ \beta_4$survival_time $+ \beta_5$sex $+ \beta_6$CHD_death $+ \beta_7$other_death $+ \varepsilon$, where α is an intercept, the βs are regression coefficients, ε is a normal error term, and all of the predictors are assumed to be normally distributed. The categorical variable for death by cause has been replaced with two dummy variables. Use of the normal distribution for the binary variables (death by cause and sex) represents a crude approximation.

MI

MI follows from EM, introducing a random component to the imputation process. Variance and covariance are adjusted by adding to the imputed values a random draw from the residual distribution of each imputed variable. In this application, the random draws were achieved by using data augmentation, an algorithm suitable for arbitrary missing data patterns (4). Very similar to EM, data augmentation is a Bayesian technique in which the deterministic iterative E and M steps are replaced by stochastic equivalents. For a sufficient number of iterations, k, the algorithm converges to the Bayesian posterior distribution of the parameters. EM provides data augmentation with a good starting point for the parameters' values, and, when prior information about their distribution is not available (as in this application), a noninformative prior distribution is used. The number of iterations required for EM to converge gives an approximate value for k. Several software programs are currently available for EM and MI (through data augmentation), including 1) a series of S-Plus functions, each suiting different types of data: NORM, CAT, MIX, PAN (17); 2) PROC MI and PROC MIANALYZE in SAS (18); and 3) SOLAS, software designed specifically for analysis of data sets with missing information (http://www.statsolusa.com). Other software implements MI through the method of sampling importance resampling (3).

We imputed missing values by using four different formulations of MI, the simplest of which was "single imputation": MI ran only once. As noted earlier, for all other implementations of MI, 10 runs were made. In each case, imputations were obtained by running independent chains of $k = 10{,}000$ iterations. Convergence of cholesterol means and standard errors to their posterior distributions was explored through time-series and autocorrelation function plots (4).

The models denoted here as single imputation and MI used the same imputation model for cholesterol as for EM. The extended-predictors model, MI^+, additionally included in the set of predictors the partially missing variables body mass index, smoking, and diabetes, imputed simultaneously to cholesterol (or excluded from the model for the studies in which they are not available). Imputations with these three models were carried out by using PROC MI, which assumes a normal distribution for each variable, conditional on all others. PROC MIANALYZE was used to combine the 10 individual estimates for MI and MI^+. Although deviations from the normal assumption are said to be unimportant, particularly when no values are missing for the predictor variables (4), this issue was addressed through the model $MI^+_{MIX,}$ implemented by using the S-Plus package function MIX and taking the same predictors as MI^+. Here, the conditional distributions of death by cause, sex, diabetes, and smoking were taken to be log-linear; all other distributions were still assumed to be conditional normal. Specification of such a model generally requires estimating a parameter for every cell of the contingency table created by the categorical variables; for the means of the continuous variables, free to vary across cells; and for the covariance matrix, constant across cells. Since the size of some APCSC studies could not support specification of this "saturated" model, we restricted the parameter space, adopting a model in which the categorical variables are marginally independent (no interaction terms are allowed) and the means of the continuous variables vary marginally with each categorical variable.

The imputation model for cholesterol, conditional on the set of predictors (P), is Cholesterol$|P = \alpha + \beta_1$age $+ \beta_2$SBP $+ \beta_3$DBP $+ \beta_4$body mass index $+ \beta_5$survival_time $+ \delta_1$sex $+ \delta_2$diabetes $+ \delta_3$smoking $+ \delta_4$CHD_death $+ \delta_5$other_death $+ \varepsilon$, where α is an intercept; the βs are the regression coefficients for the continuous variables, normally distributed; the δs are the regression coefficients for the binary variables, log-normally distributed; and ε is a normal error term.

CRITERIA FOR COMPARISON

For each study, differences between the imputation methods and complete participant analysis were explored by comparing the mean and standard deviation of cholesterol and the CHD mortality hazard ratios, and corresponding 95 percent confidence intervals, for a unit increase in cholesterol. The hazard ratios for each study were obtained from time-dependent Cox proportional hazards models, adjusted for age at risk and stratified by sex, given by $S(t \mid R) = [S_0^{(\text{sex})}]^{\exp(\beta_1 \text{cholesterol} + \beta_2 \text{SBP} + \beta_3 \text{age}(t))}$, where $S(t \mid R)$ is the probability of surviving without death from CHD to time t given the set of risk factors R; $S_0^{(\text{sex})}$ is the baseline survivor function, different for men and women; the βs are the regression coefficients for the risk factors; and age(t) is age at time t.

Random effects meta-analyses of the cholesterol-CHD relation were performed by geographic area and for the total APCSC (19). The imputations of cholesterol were carried out for each study separately because of heterogeneity in relations between predictors and cholesterol missingness.

SENSITIVITY ANALYSIS

The validity of the missing at random assumption cannot be tested empirically. However, the variability of the results across different imputation methods and different MI models describes the sensitivity of the results to the mechanism of missingness (10, 20). We also assessed how the results would be biased if nonresponders had systematically higher or lower cholesterol values than those imputed by assuming that they were missing at random by increasing or decreasing the values imputed with MI^+_{MIX} by a perturbation factor, δ ($\delta = 0$ is equivalent to missing at random) (9, 10). Since the overall standard deviation for cholesterol was 1 mmol/liter (7), we considered perturbation factors, δ, of –1, –0.5, 0, 0.5, and +1 mmol/liter.

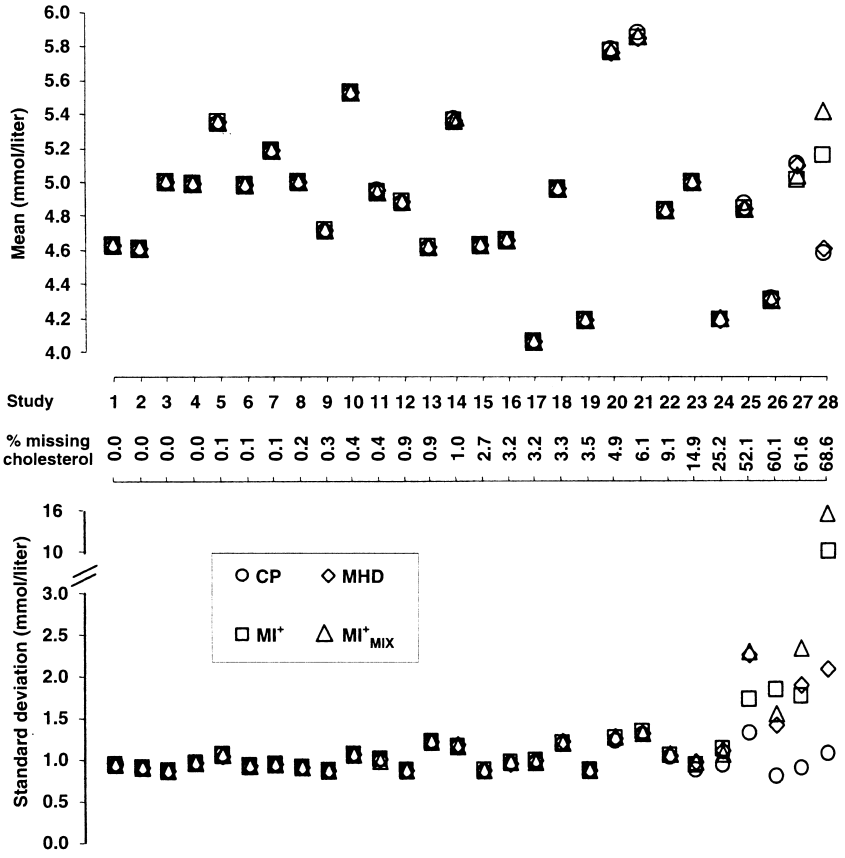

FIGURE 1. Study-specific cholesterol mean and standard deviation from the complete participant (CP) analysis and after imputation with multiple hot deck (MHD) and multiple imputation models MI$^+$ and MI$^+_{MIX}$, Asia Pacific Cohort Studies Collaboration.

RESULTS

Estimates of mean cholesterol

With the exception of those for the two studies in which 61 percent or more values were missing, the cholesterol means estimated by using the different imputation methods and complete participant analysis were very similar. However, compared with complete participant analysis, $\overline{U}$ markedly underestimated the standard deviation when more than 10 percent of cholesterol values were missing, and the underestimation deteriorated as the rate of missingness increased (results not shown). Standard deviations obtained from $\overline{C}$ were equivalent to those from complete participant analysis; those from EM and single imputation were larger but still less than those from multiple hot deck imputation and the MI models when more than 10 percent of cholesterol values were missing. Figure 1 displays the results for each study's cholesterol mean and standard deviation for complete participant analysis and after imputing cholesterol by using multiple hot deck and the MI models MI$^+$ and MI$^+_{MIX}$. For studies 27 and 28, the high rate of nonresponse created extreme uncertainty about the imputed values that resulted in very high standard deviations; with the data augmentation algorithm, there were problems in reaching convergence to the posterior distribution of cholesterol. Therefore, for these two studies, results obtained from any analysis involving cholesterol should be treated with caution.

Estimates of hazard ratios for CHD death

Figure 2 shows the hazard ratios for CHD death under complete participant analysis and after imputing cholesterol with multiple hot deck and the MI models MI$^+$ and MI$^+_{MIX}$ (results from other methods not shown). Study 28 was omitted because no CHD deaths were recorded. Differences were apparent for only those studies in which more than 10 percent of cholesterol values were missing and then depended on a combination of several study-specific factors, including the proportion of cholesterol values missing, the total number of CHD deaths, the proportion of CHD deaths omitted by the complete participant analysis, and relations between cholesterol and the mechanism of missingness of

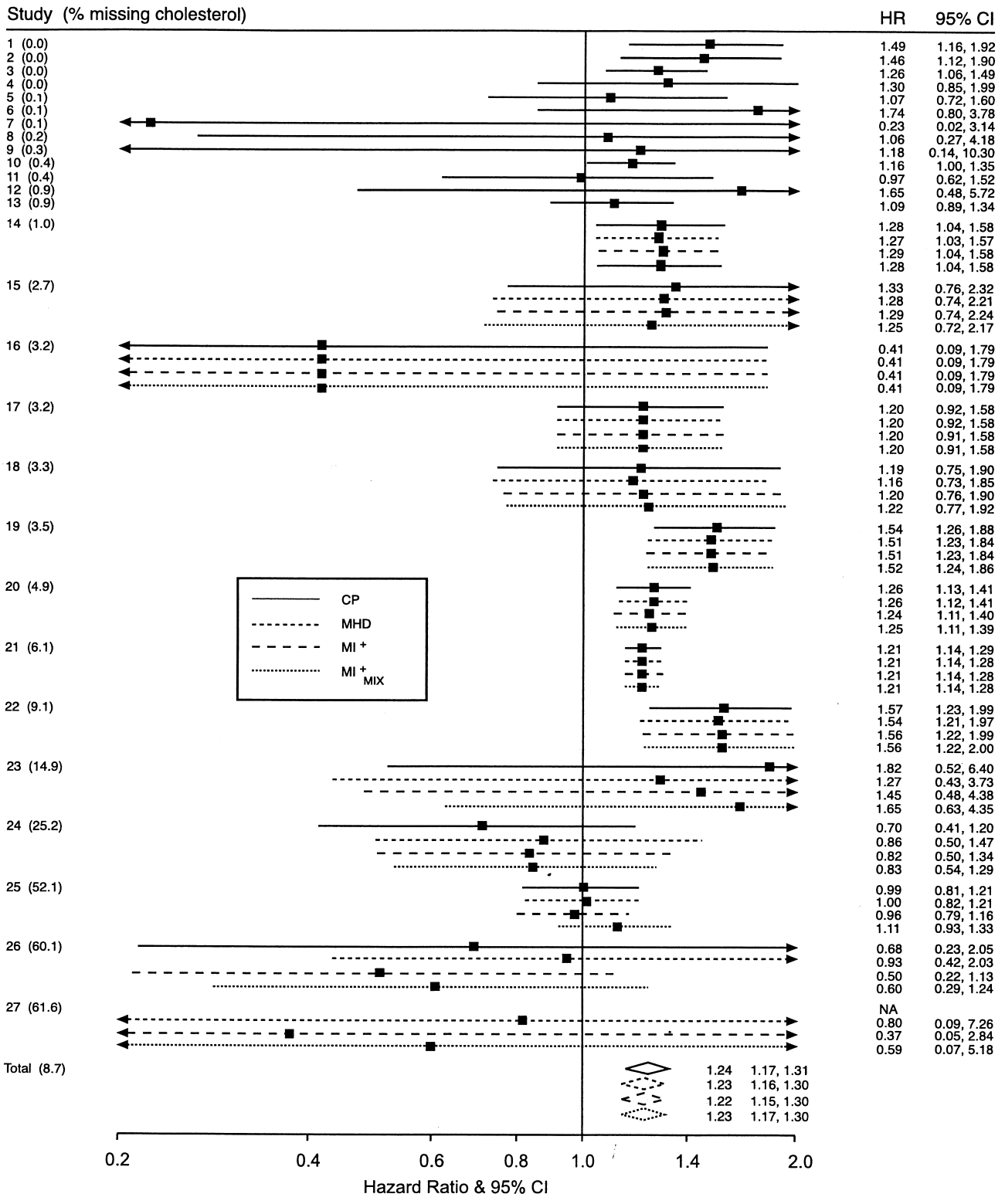

FIGURE 2. Hazard ratios and 95% confidence intervals for coronary heart disease death for a 1-mmol/liter increase in cholesterol, adjusted for age at risk and sex, from the complete participant (CP) analysis and after imputation with multiple hot deck (MHD) and multiple imputation models MI^{+} and MI^{+}_{MIX}, Asia Pacific Cohort Studies Collaboration. Studies in which less than 1 percent of cholesterol values were missing (studies 1–13) have only one hazard ratio and 95% confidence interval since these were invariant to the method used. Arrows indicate that the confidence intervals exceed the range 0.2–2. NA, not available; HR, hazard ratio; CI, confidence interval.

cholesterol and its predictors. Thus, differences between imputation methods were more accentuated for study 23 (15 percent missing cholesterol values) than for studies 24 and 25 (substantially more values missing). Furthermore, although study 24 had both more missing cholesterol values and more CHD deaths omitted by complete participant analysis than study 23 did, the confidence intervals nevertheless expanded much more when missing values were not imputed (i.e., complete participant analysis) for study 23 than for study 24. These differences are explained by the fact that study 23 recorded many fewer CHD deaths than either study 24 or 25 did (table 1).

Figure 3 shows the results for the meta-analyses in which complete participant analysis and all of the imputation methods were used. Within geographic areas, hazard ratios and their confidence intervals did not change appreciably with different approaches to dealing with missingness. As expected from the high rate of missingness, the largest differences were found for China.

Sensitivity analyses

Figure 4 shows the sensitivity of the hazard ratios and their confidence intervals to over- or underestimation of cholesterol when MI^{+}_{MIX} was used. Again, there were important differences for only those studies in which more than 10 percent of the cholesterol values were missing, where the anticipated effects were sometimes substantial. There was no effect on the overall meta-analysis and little effect for the area-specific meta-analyses (not shown).

DISCUSSION

APCSC provides an unusual opportunity to assess the possible bias induced regarding point and variance estimates by using complete participant analyses and to compare the results obtained with different imputation techniques in the presence of different levels of missing information. Although missingness for cholesterol was clearly not completely random in APCSC, the complete participant analysis was not importantly different for studies with low rates of missingness (below 10 percent), and results derived from different imputation approaches were also very similar for these studies. This finding is consistent with previous reports (9, 21). In contrast, for those studies in which more than 10 percent of the values were missing, differences were more evident, particularly for the standard deviation of cholesterol, and became increasingly more important as missingness became more frequent.

Compared with the other imputation methods used here, MI is the most appealing because it allows for any type and number of variables and should give the most reliable variance estimates. MI is known to yield estimates with theoretical properties that the other imputation methods do not provide when the missing at random assumption is satisfied. In addition, it is robust to departures from the data model, unless large amounts of data are imputed. Several authors (22–24) who have tested the performance of various imputation techniques through simulation have recommended MI. Similar to MI, multiple hot deck is expected to be more reliable than the simpler imputation approaches used here (25, 26), but its ability to deal with continuous predictors is limited. APCSC results suggest that MI would be the method of choice when missingness is above 10 percent but that complete participant analysis is otherwise acceptable.

At present, MI implementation in SAS software is limited to assuming normally distributed data, clearly not a valid assumption for the categorical and binary variables used in some prediction models here. However, only for those studies with many missing values and for the result in China (32 percent missing) did there appear to be a difference between MI obtained through specifying a multivariate normal model (MI^{+}) and through a model more suitable to mixed types of data (MI^{+}_{MIX}). These results agree with simulation studies (4) showing that MI is generally robust to departures from normality and generally to model misspecification when the amounts of missing data are not large. For studies where it was possible, we drew imputations under a model for mixed data that allowed for all interactions between covariates; again, the results did not change. When predictors of moderate importance are left out of MI models, inferences should still remain valid, although the between-imputations variance will increase (13). We examined this issue by comparing MI and MI^{+}, the latter having the extra predictors body mass index, smoking, and diabetes; important differences were found only when rates of missingness were greater than 10 percent.

All of the models assume that cholesterol was missing at random. There was, once again, evidence that the association between cholesterol and CHD death was sensitive to systematic error in this assumption only for those studies in which more than 10 percent of cholesterol values were missing. In such cases, special models for mechanisms of missingness that are not missing at random (27) might be used. Since we lacked specific information about the process of missingness in APCSC, we could not fit such models. Without ancillary information, use of such models would lead to results that can hardly be considered less biased than those obtained by using models that assume missing at random (14).

Differences between imputation methods depend not only on the amount of missing cholesterol data but also on a number of factors including the relations between missingness of cholesterol and the predictors used. We found no systematic relation, from study to study, between missingness of cholesterol and any of the predictor variables described in this article (results not shown). This lack of consistency is an important finding, suggesting that there is no standard missing value mechanism and thus no simple recipe for handling missingness of cholesterol. In every situation, preliminary descriptive analyses are essential for identifying the type of missingness and of important predictors and therefore to specifying a correct imputation model. Analyses such as those shown in figures 2 and 4 indicate to what extent the results depend on the missing values, and they are useful for diagnosing specific problems. Monitoring of time-series and autocorrelation function plots and warnings of lack of convergence, available in SAS PROC MI, are very helpful for indicating when estimates are likely to be unreliable.

With each imputation method, we observed between-study heterogeneity for the estimated association between

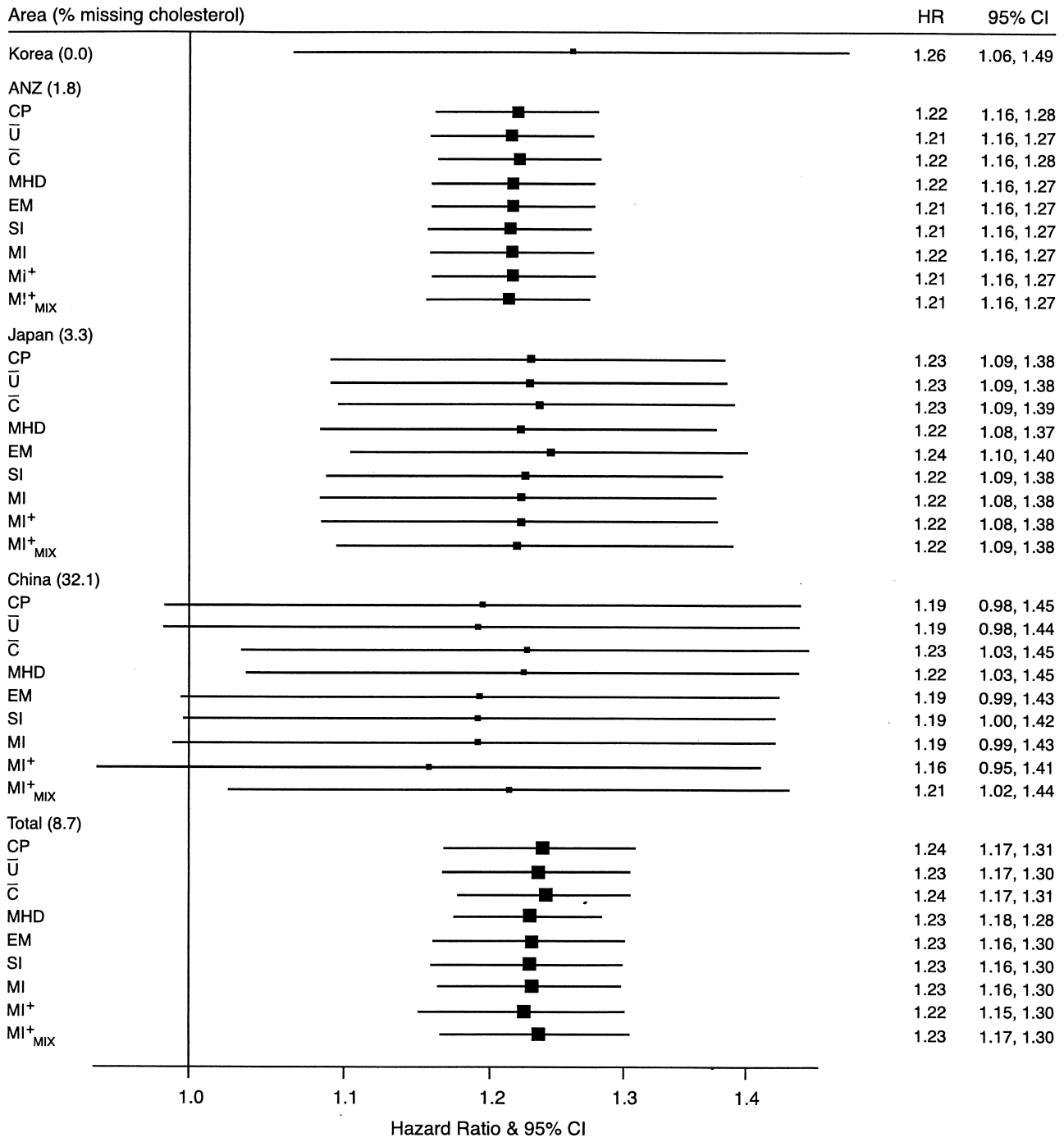

FIGURE 3. Hazard ratios and 95% confidence intervals for coronary heart disease death for a 1-mmol/liter increase in cholesterol, adjusted for age at risk, sex, and study, Asia Pacific Cohort Studies Collaboration. Results are shown by geographic area and overall for complete participant analysis and for each imputation method used. The size of the black squares is proportional to the standard error of the log hazard ratio. HR, hazard ratio; CI, confidence interval; ANZ, Australia and New Zealand; CP, complete participant; $\bar{U}$, unconditional mean imputation; $\bar{C}$, conditional mean imputation; MHD, multiple hot deck; EM, expectation maximization; SI, single imputation; MI, MI+, and MI+$_{MIX}$, multiple imputation models; China, China-Taiwan-Singapore.

cholesterol and CHD mortality. For example, unexpected hazard ratio estimates of less than one were found, for complete participant analysis and all of the imputation approaches, in four studies. This variability is likely to be due to sampling variation or to the high rate of missingness in some studies rather than to any real differences. Only

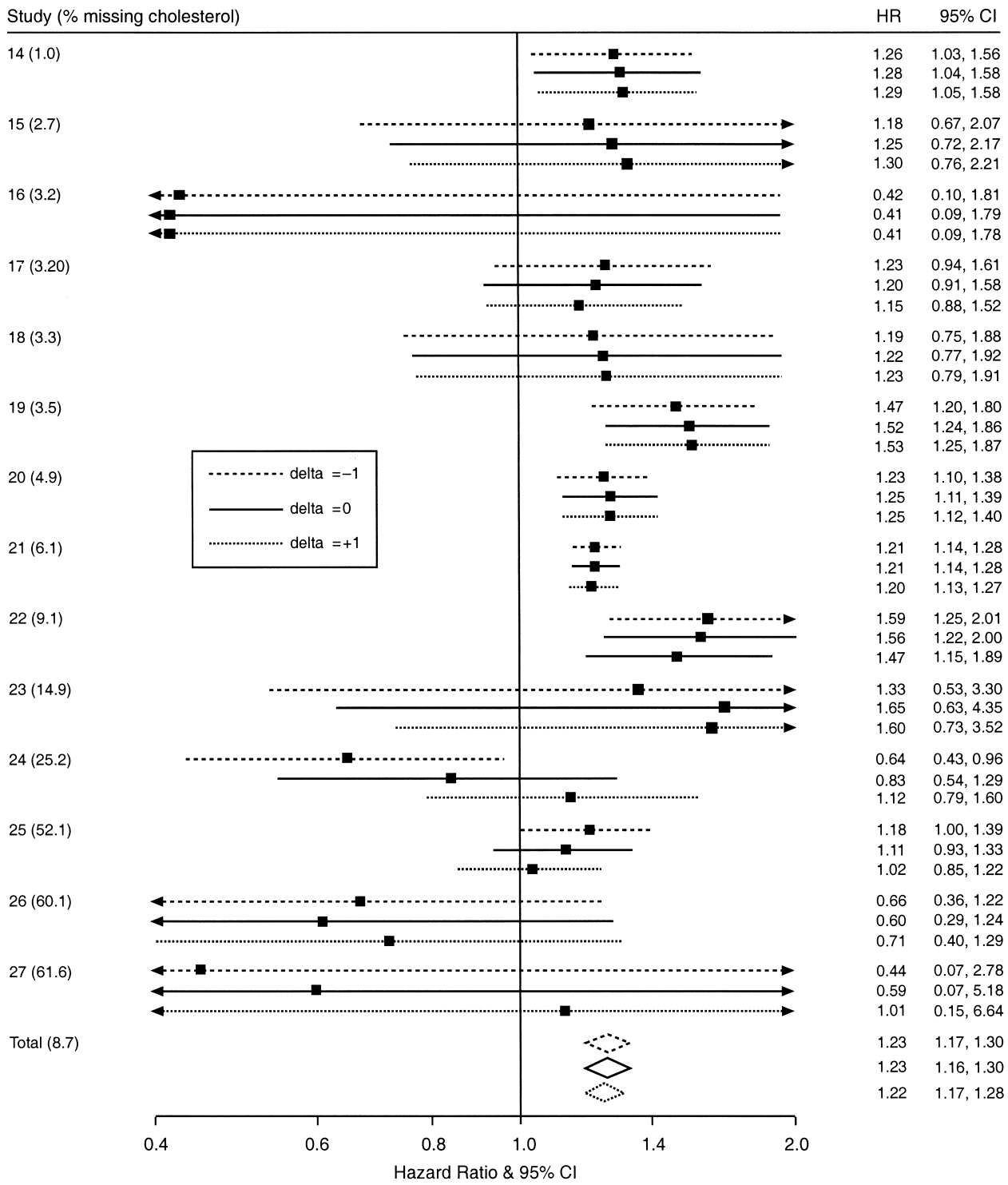

FIGURE 4. Sensitivity analyses showing the effects of a perturbation, delta, equal to –1 or +1 mmol/liter on cholesterol imputed with MI^{+}_{MIX} (a multiple imputation model): hazard ratios and 95 percent confidence intervals for coronary heart disease death for a 1-mmol/liter increase in cholesterol, adjusted for age at risk and sex, Asia Pacific Cohort Studies Collaboration. Arrows indicate that the confidence intervals exceed the range 0.4–2. HR, hazard ratio; CI, confidence interval.

one study had both a hazard ratio of less than one and reasonably tight confidence limits (even then, unity was within these limits).

An important and reassuring finding, key to APCSC, was the similarity of the area-specific and overall meta-analyses results obtained with complete participant anal-

ysis and the different imputation models (figure 3), presumably due to averaging out the different mechanisms leading to missingness as well as diluting the proportional representation of missing values. This finding suggests an extra advantage of meta-analysis in observational epidemiology, where missing values are common. In addition to the well-known advantage of reducing random noise, pooled analyses may be less prone to bias due to missing values.

ACKNOWLEDGMENTS

The authors thank Derrick Bennet, Sarah Lewington, Rachel Huxley, Anushka Patel, and Gary Whitlock for helpful comments on the manuscript; Varsha Parag for compiling the data; and Beverley Mullane for her contribution in producing the figures.

APCSC Executive Committee: D. Gu, T. H. Lam, C. Lawes, S. MacMahon, W.-H. Pan, A. Rodgers, I. Suh, H. Ueshima, and M. Woodward.

APCSC participating studies and principal collaborators in APCSC (the underlined studies provided data used in this paper): *Aito Town*: A. Okayama, H. Ueshima, and H. Maegawa; *Akabane*: N. Aoki, M. Nakamura, N. Kubo, and T. Yamada; *Anzhen*: C. H. Yao and Z. S. Wu; *Anzhen02*: Z. S. Wu; *Beijing Steelworkers*: L. S. Liu and J. X. Xie; *Blood Donors' Health*: R. Norton, S. Ameratunga, S. MacMahon, and G. Whitlock; *Busselton Study*: M. W. Knuiman; *Canberra-Queanbeyan*: H. Christensen; *Capital Iron and Steel Company*: X. G. Wu; *CISCH*: J. Zhou and X. H. Yu; *Civil Service Workers*: A. Tamakoshi; *CVDFACTS*: W. H. Pan; *East Beijing*: Z. L. Wu, L. Q. Chen, and G. L. Shan; *Fangshan Farmers*: D. F. Gu and X. F. Duan; *Fletcher Challenge:* S. MacMahon, R. Norton, G. Whitlock, and R. Jackson; *Guangzhou*: Y. H. Li; *Guangzhou Occupational*: T. H. Lam and C. Q. Jiang; *Hisayama Study*: M. Fujishima, Y. Kiyohara, and H. Iwamoto; *Hong Kong*: J. Woo and S. C. Ho; *Huashan*: Z. Hong, M. S. Huang, and B. Zhou; *Kinmen*: J. L. Fuh; *Korean Medical Insurance Corporation*: I. Suh, S. H. Jee, and I. S. Kim; *Kounan*: H. Ueshima, Y. Kita, and S. R. Choudhury; *Melbourne Cohort*: G. Giles; *Miyama*: T. Hashimoto and K. Sakata; *Newcastle*: A. Dobson; *Ohasama*: Y. Imai, T. Ohkubo, and A. Hozawa; *Perth Cohort*: K. Jamrozik, M. Hobbs, and R. Broadhurst; *Saitama*: K. Nakachi; *Seven Cities study*: X. H. Fang, S. C. Li, and Q. D. Yang; *Shanghai Factory Workers*: Z. M. Chen; *Shibata City*: H. Tanaka; *Shigaraki*: Y. Kita, A. Nozaki, and H. Ueshima; *Shirakawa*: H. Horibe, Y. Matsutani, and M. Kagaya; *Singapore NHS92*: D. Heng and C. S. Kai; *Singapore Thyroid and Heart Study*: K. Hughes and J. Lee; *Six Cohorts*: B. F. Zhou and H. Y. Zhang; *Tanno/Soubetsu*: K. Shimamoto and S. Saitoh; *Tianjin*: Z. Z. Li and H. Y. Zhang; *Xi'an*: Y. He and T. H. Lam; *Yunnan*: S. X. Yao.

REFERENCES

1. Little RJA, Rubin DB. Statistical analysis with missing data. New York, NY: John Wiley & Sons, Inc, 1987.
2. Demissie S, LaValley MP, Horton NJ, et al. Bias due to missing exposure data using complete-case analysis in the proportional hazard regression model. Stat Med 2003;22:545–57.
3. Rubin DB. Multiple imputation for nonresponse in surveys. New York, NY: John Wiley & Sons, Inc, 1987.
4. Shafer JL. Analysis of incomplete multivariate data. London, United Kingdom: Chapman & Hall, 1997.
5. van Buuren S. Multiple imputation online. Leiden, the Netherlands: Department of Statistics, TNO Prevention and Health, 2003. (http://www.multiple-imputation.com).
6. Asia Pacific Cohort Studies Collaboration. Determinants of cardiovascular disease in the Asia Pacific region: protocol for a collaborative overview of cohort studies. CVD Prevention 1999;2:281–9.
7. Asia Pacific Cohort Studies Collaboration. Cholesterol, coronary heart disease and stroke in the Asia Pacific region. Int J Epidemiol 2003;32:563–72.
8. Rubin DB. Inference and missing data. Biometrika 1976;63: 581–92.
9. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med 1999;18:681–94.
10. Longford NT. Handling missing data in diaries of alcohol consumption. J R Stat Soc (A) 2000;163:381–402.
11. Barnard J, Meng XL. Application of multiple imputation in medical studies: from AIDS to NHANES. Stat Methods Med Res 1999;8:17–36.
12. Haitovsky Y. Missing data in regression analysis. J R Stat Soc (B) 1968;30:67–82.
13. Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc 1996;91:473–89.
14. Allison PD. Missing data. Thousand Oaks, CA: Sage Publications, Inc, 2001. (Series: quantitative applications in the social sciences. Vol 136. A Sage University paper).
15. Rubin DB, Schenker N. Multiple imputations in health care databases: an overview and some applications. Stat Med 1991;10:585–98.
16. Mander A, Clayton D. HOTDECK: Stata module to impute missing values using the hotdeck method. Statistical Software Components S366901. Chestnut Hill, MA: Boston College Department of Economics, 1999. (Revised September 12, 2002). (http://ideas.repec.org/s/boc/bocode1.html).
17. Shafer JL. Software for multiple imputation. University Park, PA: The Pennsylvania State University Department of Statistics, 1999. (http://www.stat.psu.edu/~jls/misoftwa.html).
18. Yuan YC. Multiple imputation for missing data: concepts and new development. Rockville, MD: SAS Institute, Inc. (Paper 267-25). (http://support.sas.com/rnd/app/papers/multipleimputation.pdf).
19. Sutton AJ, Abrams KR, Jones DR. et al. Methods for meta-analysis in medical research. Chichester, United Kingdom: John Wiley & Sons, Inc, 2000.
20. Taylor JMG, Cooper KL, Wei JT, et al. Use of multiple imputation to correct for nonresponse bias in a survey of urologic symptoms among African-American men. Am J Epidemiol 2002;156:774–82.
21. Arnold AM, Kronmal RA. Multiple imputation of baseline data in the Cardiovascular Health Study. Am J Epidemiol 2003;157:74–84.
22. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995;142:1255–64.
23. Zhou XH, Eckert GJ, Tierney WM. Multiple imputation in public health research. Stat Med 2001;20:1541–9.
24. Twisk J, de Vente W. Attrition in longitudinal studies: how to

deal with missing data. J Clin Epidemiol 2002;55:329–37.
25. Gmel G. Imputation of missing values in the cases of multiple item instrument measuring alcohol consumption. Stat Med 2001;20:2369–81.
26. Perez A, Dennis RJ, Gil JFA, et al. Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia. Stat Med 2002;21:3885–96.
27. Little RJA. Pattern-mixture models for multivariate incomplete data. J Am Stat Assoc 1993;88:125–34.