RESEARCH COMMUNICATION

New Methods of Handling Cases of Unknown Age in Cancer Registry Data

Mahdi Fallah*, Elham Kharazmi

Abstract

Objective: The essential assumption of random missing age behind the "conventional method" of handling cancer patients of unknown age does not often hold. This article is to introduce four alternative methods based on more acceptable assumptions. Methods: More cases with unknown age are allocated to the older age-groups in all the new methods. In the "weighting method," cases of unknown age are distributed according to distribution of cases of known age, whereas in the "last-group method," all of them are added to the oldest age-group. In the "progressive method," unknown-age cases are added to the age-groups above 60 progressively (weighting=1/63, 2/63, 4/63, 8/63, 16/63, and 32/63), whereas in the "additive method," they are allocated to the age-groups above 60 additively (weighting=1/21, 2/21, 3/21, 4/21, 5/21, and 6/21). Data were from the Cancer in Five Continent database, vol. VIII. Results: Age-standardized rates for "All sites" in Zaragoza (Spain), Cali (Colombia), Algiers (Algeria), and Gambia showed that results by all the methods differed, the magnitude ranging from 0.1 to 3.1% depending on the method, registry, sex, and the defined last age-group. Conclusion: Conventional and weighting methods are not based on acceptable assumptions. The last-group method is not stable because it depends on the defined age-group as last (65+, 75+ or 85+). Both progressive and additive methods have more acceptable assumptions. The progressive method is preferable above all others because it can produce an age-specific curve with the expected exponential increase.

Key Words: Neoplasms - incidence rate - unknown age - old age correction

Asian Pacific J Cancer Prev, **9**, 259-262

Introduction

Despite all the efforts by cancer registries, still there may be some cases of unknown age in the registry data. So far, there are two methods of handling unknown-age cases: 1) Simply exclude them, which naturally results in the under-estimation of age-standardized and cumulative rates; 2) Distributing them equally into all age-groups (this method will be named the "conventional method" in this manuscript). The correction using the conventional method relies on the assumption that the cases with missing age are randomly distributed, so that the probability that the age of a case is unknown does not depend on the age of the case. Although this assumption probably does not hold (it is more likely that age is not recorded in older cases), it is nevertheless important that all registered cases are accounted for, so that the summary statistics are not under-estimated (Jensen et al., 1991; Parkin et al., 2002). The conventional method is widely used although this assumption is often violated. There is no alternative method in the current literature. This article was compiled to introduce four alternative methods to deal with cancer cases with unknown age and compare them with the conventional one.

Materials and Methods

Conventional method

In the conventional method, the procedure involves simply multiplying either summary measure based on known age (such as age-standardized rate) by total number of cases of cancer of the same type in persons of the same sex divided by the number occurring in persons of known age (Jensen et al., 1991; Parkin et al., 2002).

Four other methods to calculate summary statistics with four different assumptions than the assumption in conventional method are: "weighting method", "last-group method", "progressive method" and "additive method".

Weighting method

In the "weighting method," cases of unknown age are distributed based on the distribution of cases of known age but not equally for each group. The formula for this method is as follows:

$$c_i = k_i + [u \times (k_i \Big/ \sum_{1}^{18} k_i)]$$

, where c is the corrected number of cases, i is the age-

Tampere School of Public Health, FIN-33014, University of Tampere, Tampere, Finland. Email: Mahd.Fallah@gmail.com

group, k is the number of cases with known age, u is the total number of cases with unknown age, and

$$\sum_{1}^{18} k_i$$

is the summation of all cases with known age in 18 age-groups (0-4, 5-9, …, 80-84 and 85+). If for any reason, instead of 18 age-groups, less age-groups are used (like using 65+ as the last age-group), the weighting method can be adjusted accordingly and the formula will be

$$c_i = k_i + [u \times (k_i \Big/ \sum_{1}^{14} k_i)]$$

, where $\sum_{1}^{14} k_i$ is the summation of all cases with known age in 14 age-groups (0-4, 5-9, …, 60-64, and 65+).

The assumption behind this method is that cases of unknown age are distributed according to the distribution of cases with known age, so not randomly distributed. As the number of cases of known age often increases by age, more cases with unknown age are added to the older age-groups and less to the younger ones.

Last-group method

In the "last-group method," all the cases of unknown age are added to the last age-group (85+). So the formula is $c_{18}=k_{18}+u$ where c_{18} is corrected number of cases in the age-group 85 or more, k_{18} is the number of cases with known age in the age-group 85 or more and u is the total number of cases of unknown age. The assumption for this method is that all cases with missing age are from the oldest age-group. If for any reason, instead of 18 age-groups, less age-groups are used (i.e. using 65+ as the last age-group), the last-group method can be adjusted accordingly and the formula for last-group method will then be $c_{14}=k_{14}+u$, where c_{14} is corrected number of cases in the age-group 65 or more, k_{14} is the number of cases of known age in the age-group 65 or more and u is the total number of cases with unknown age.

Progressive method

In the "progressive method," unknown-age cases are added to the age-groups above 60 progressively. The formula for this method is as follows:

$$c_i = k_i + (u \times w_i)$$

, where c is the corrected number of cases, i is the age-group, k is the number of cases of known age and u is the total number of cases of unknown age, and w_i is the corresponding progressive weight for elderly age-groups (w_{13}=1/63 for age-group 60-64, w_{14}=2/63 for 65-69, w_{15}=4/63 for 70-74, w_{16}=8/63 for 75-79, w_{17}=16/63 for 80-84, and w_{18}=32/63 for 85+; 63 is result of summing 1, 2, 4, 8, 16, and 32). If for any reason, instead of 18 age-groups, less age-groups are used (i.e. using 65+ as the last age-group), the progressive method can be adjusted accordingly and the formula will be the same but weights are different, so that w_{13}=1/63 for 60-64, and w_{14}=61/63 for 65+.

Additive method

In the "additive method," unknown-age cases are allocated to the age-groups above 60 additively. The formula for this method is again as follows:

$$c_i = k_i + (u \times w_i)$$

, where c is the corrected number of cases, i is the age-group, k is the number of cases of known age, u is the total number of cases of unknown age, and w_i is the corresponding progressive weight for elderly age-groups (w_{13}=1/21 for age-group 60-64, w_{14}=2/21 for 65-69, w_{15}=3/21 for 70-74, w_{16}=4/21 for 75-79, w_{17}=5/21 for 80-84 and w_{18}=6/21 for 85+; 21 is result of summing 1, 2, 3, 4, 5, and 6). If for any reason, instead of 18 age-groups, less age-groups are used (i.e. using 65+ as the last age-group), the additive method can be adjusted accordingly and the formula will be the same, but weights are different, so that w_{13}=1/21 for 60-64, and w_{14}=20/21 for 65+.

To compare these methods, data from cancer registries of Zaragoza (Spain), Cali (Colombia), Algiers (Algeria) and Gambia with highest percentage of cases of unknown age in the computerized database enclosed in the book Cancer Incidence in Five Continents, vol. VIII (Parkin et al., 2002) were used as examples (Table 1). Authors had no conflict of interest in the selection of sample registries.

Since validation of the new methods was not possible by finding age of cases of unknown age through individual case-tracing or capture-recapture method, shape of age-specific incidence curve was used as indicator of validity

Table 1. Comparison Between Age-standardized Rates Calculated with Five Different Methods (Conventional, Weighting, Last-group, Progressive and Additive) of Correction for Cases of Unknown Age

Registry	Year	Last group	Sex	Unknown-age cases %	Age-standardized rate (ASR, per 100,000) Not corrected O	Corrected for cases of unknown age C	W	L	P	A	Increase in ASR (%) Compared to C W	L	P	A
Gambia	1997-8	65+	Male	7.8	77.9	84.5	84.0	86.0	86.0	86.0	-0.6	1.8	1.8	1.8
			Female	11.1	75.2	84.6	83.6	87.0	87.0	87.0	-1.2	2.8	2.8	2.8
Algiers	1993-7	75+	Male	6.6	87.4	93.6	93.2	93.5	93.7	93.9	-0.4	-0.1	0.1	0.3
			Female	4.9	85.0	89.4	89.2	88.9	89.1	89.4	-0.2	-0.5	-0.3	0.0
Cali	1992-6	85+	Male	4.0	184.9	192.6	192.3	194.3	198.5	193.1	-0.2	0.9	3.1	0.3
			Female	4.3	190.9	199.4	199.1	198.7	198.5	199.0	-0.2	-0.4	-0.5	-0.2
Zaragoza	1991-5	85+	Male	3.4	303.3	313.9	313.5	311.9	311.0	311.1	-0.1	-0.6	-0.9	-0.9
			Female	3.3	184.1	190.5	190.2	187.0	187.1	187.7	-0.1	-1.8	-1.7	-1.5

C, Conventional; W, Weighting; L, Last-group; P, Progressive; A, Additive

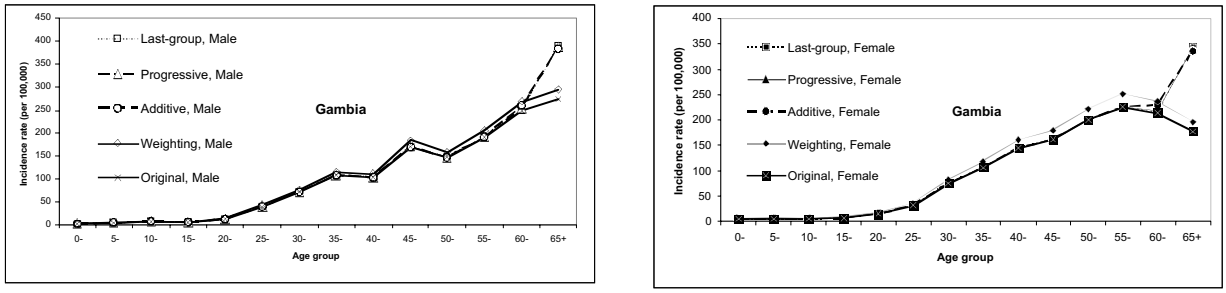

Figure 1. Age-specific Incidence Curves for the Gambia, Before and After Correction for Cases of Unknown Age with Four New Methods, 1997-8

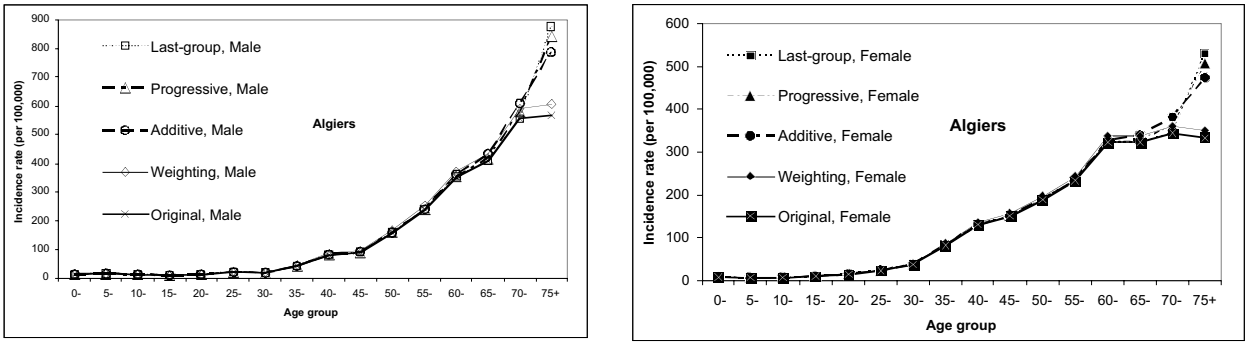

Figure 2. Age-specific Incidence Curves for Algiers, Algeria, Before and After Correction for Cases of Unknown Age with Four New Methods, 1993-7

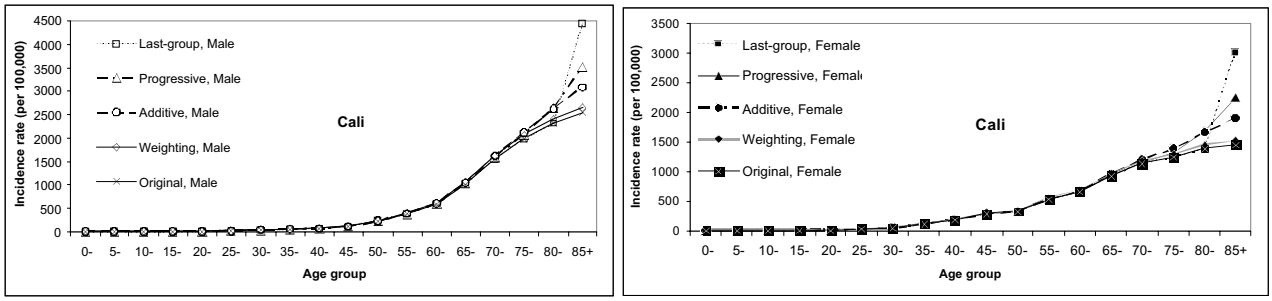

Figure 3. Age-specific Incidence Curves for Cali, Colombia, Before and After Correction for Cases of Unknown Age with Four New Methods, 1992-6

of correction methods. The method that could successfully compensate the under-ascertainment in the elderly ages was indicated as the most valid method.

Results

Age-standardized rates for "All sites" in Zaragoza (Spain), Cali (Colombia), Algiers (Algeria) and Gambia showed that results by all the methods differed (Table 1). Last-group, progressive and additive methods resulted in more acceptable shape of age-specific curves for the oldest age-groups, however, progressive method resulted in the most satisfactory shape of age-specific curve (Figures 1 to 4). The magnitude of the difference between age-standardized rates calculated by different methods ranged from 0.1 to 3.1% depending on the defined last age-group, registry, sex, and method (Table 1).

Discussion

Comparing age-standardized rates of five different methods showed that results by the weighting method were quite similar to the conventional one, but with trivial under-estimation. Results by last-group, progressive and additive methods differed when the last age-group was

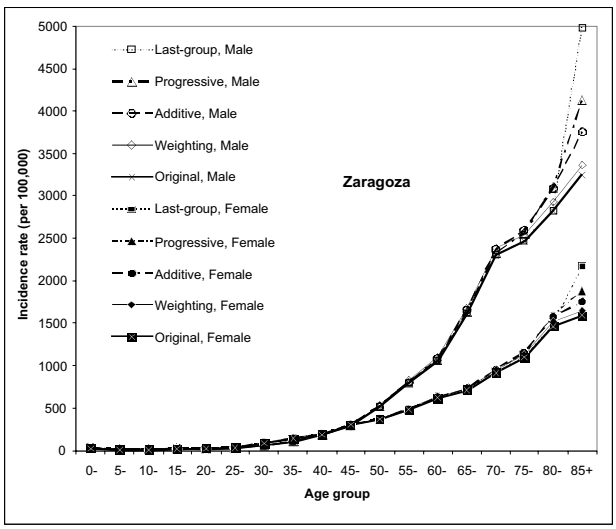

Figure 4. Age-specific Incidence Curves for Zaragosa, Spain, Before and After Correction for Cases of Unknown Age with Four New Methods, 1993-7

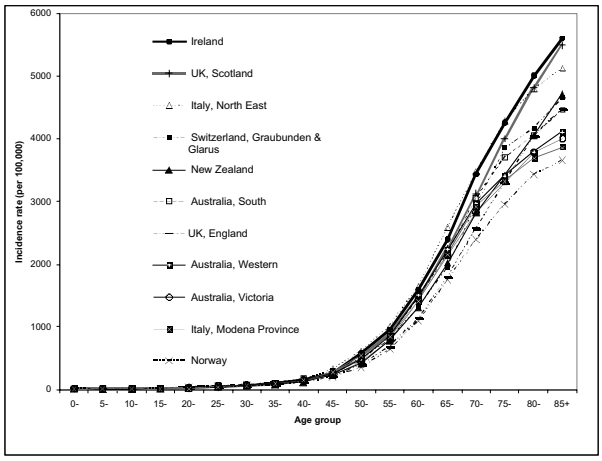

Figure 5. Age-specific Incidence Curves of "All sites excluding skin" Cancer in Some Registries without Fall-off in the Oldest Age-groups, Male, 1993-97

defined differently, but when the last group was 65+, they all gave similar results. However, if 85+ was considered as the last group, results by these three methods were quite different and the judgment on the validity of these methods was made based on the shape of age-specific curve. For most of the epithelial cancers, incidence rate increases as a power of age, and since most cancers are epithelial, this pattern should be observed for "All sites" (at least after the age of 15). It is usual to see some decline in the oldest age-groups (over 70). This is partly due to less efficient case ascertainment, some of which is a consequence of competing causes of mortality in the elderly (so that cancer is not recorded on death certificate). Under-ascertainment must always be considered if there is an actual decline in rates (Parkin et al., 1994). In complete registries, incidence of all cancer sites increases exponentially with age (Figure 5). The best result is then obtained with the progressive method since it provided the most satisfactory shape of age-specific curve (Figures 1 to 4).

The assumptions behind these five methods are different and it seems that generally the assumptions for all four new methods are more rational than the conventional one because having missing age is not independent of age. Those with missing age are more likely to be at older ages (Jensen et al., 1991; Parkin et al., 2002). Even among these four new methods, it seems the weighting method is less acceptable as it keeps the shape of under-ascertainment in the elderly ages the same as the original one (Figures 1 to 4). Age-specific incidence curve is often used as an indicator of quality of cancer registry data in order to detect abnormal fluctuations in the anticipated patterns, including any fall-off in the incidence rate in older subjects (suggestive of under-ascertainment in oldest age-groups) (Parkin et al., 2002). The age-specific curves show that last-group, progressive and additive methods can have another benefit which is accounting for the under-ascertainment in the oldest age-group which occurs in many cancer registries.

The importance of using correct method of accounting for cases of unknown age increases by increase of number of cases with missing age. Cases of unknown age are more common in developing countries. This importance of a valid method of handling unknown-age cases increases while establishment of new cancer registries especially in developing countries is increasing. In countries where a considerable proportion of population is reported as unknown age in the census data, using a similar method to distribute people of unknown age in the population into the age-groups seems necessary.

None of these five methods affect the calculation of crude incidence rate and only the last-group method does not affect the cumulative incidence rates under age 85 (usually 0-64 or 0-74 years is reported) since all cases of unknown age are added to the last age-group (85+).

In conclusion, it is imperative to use a valid method of accounting for cases of unknown age to avoid possible under/over-estimation of summary measures since this study showed that using an incorrect method can result in up to 3% bias in the estimates. The conventional and weighting methods are based on less acceptable assumptions and provide very similar but rather under-estimated results. The last-group method is not stable because it depends on the defined age-group as last (65+, 75+ or 85+). Progressive and additive methods not only are based on more acceptable assumptions, but also they can compensate the under-ascertainment in the very elderly age-group, which can be reflected as a more acceptable shape of age-specific curve. The progressive method is preferable above all others because it can produce an age-specific curve with expected exponential increase.

References

Jensen OM, Parkin DM, Maclennau R, Mair CS, Skeet RG (1991). Cancer registration principles and methods. IARC Press, Lyon

Parkin DM, Chen VW, Ferlay J, Galceran J, Whelan SL (1994). Comparability and Quality Control in Cancer Registration. IARC Press, Lyon

Parkin DM, Whelan SL, Ferlay J, Teppo L, Thomas DB (2002). Cancer Incidence in Five Continents. IARC Press, Lyon