Disclosure risk of cause of death related to basic demographic and epidemiological statistics

Daan G Uitenbroek
* consultant at www.quantitativeskills.com

Abstract. In this article the disclosure risk of mortality statistics such as the Standardized Mortality Ratio (SMR) and the Comparative Mortality Figure (CMF) is discussed. There are methods to reverse calculate these statistics back to the table data on which they are based. This might lead to the disclosure of the cause of death of a deceased person. The conclusion of this paper is that the SMR can be published without concern, except for small communities where the diversity of causes of death is limited. The CMF can be published without concern when the details of the original calculation are unknown. When the details are known the CMF can be published when the total numbers of deaths of a certain cause are sufficiently high, a relatively large number of age categories are used in the original calculation, and the statistic which is used has relatively few meaningful digits.

Keywords. disclosure limitation, disclosure control, confidentiality, privacy, mortality statistics.

1 Introduction

Publication of statistics and other quantitative data is a balance between the privacy of the people on which the data is based and the need to have detailed knowledge as a basis for research and policy development [1]. In this paper we will discuss privacy considerations in the publication of basic mortality statistics, however, the discussion will in general apply to all basic demographic and epidemiological statistics. An important concern in the publication of mortality statistics is that the identity of the deceased is not disclosed and that the cause of death of a person, possibly a cause of a sensitive nature, is made public. Tabulating mortality data by aggregation of individuals into groups with the same cause of death, age, place of residence, ethnicity, and other variables of substantive interest, is most often used. By suppressing cells with only a few cases [2,3], or by rounding the cell data to numbers such as three, five or ten [3,4,5], the information is considered safe for the possible identification of cause of death of deceased persons. Besides cell suppression a minimum number of causes of death are required to ensure that the presence in a table is not easily associated with a particular cause of death. Besides cell suppression and substantive diversity there is variety of other more or less sophisticated methods to prevent disclosure of sensitive personal information in mortality statistics [6].

Another solution to the disclosure problem would be to publish statistics based on tables which include rare events and sparsely filled cells, but not publish the tables. To enable this for external researchers organizations as the CDC/NCHS [7] and Statistics Netherlands [8] allow controlled on-site analysis of sensitive data. The UK-ONS has an accreditation system which gives selected researchers access to sensitive data followed by an output control procedure [3]. However, there is a concern with regard to the publication of statistics which are produced in a controlled environment [8]. Statistics might be reverse calculated to tables which do not meet the statistics bureau's stringent criteria. This has been researched to a certain extent for regression analysis [9] and has led to limitations in the disclosure of the results of complex analysis. These limitations can influence the quality of the outcome regression analysis and other statistics [10]

A similar logic of output control and limitation is now being applied with regard to the most basic demographic and epidemiological statistics, such as the Standardized Mortality Ratio (SMR), Comparative Mortality Figure (CMF), the Years of Potential Life Lost (YPLL or PYLL), the life expectancy and directly standardized rates. These statistics are at the basis of health policy development on the regional, national and international level. These statistics are attractive because they are intuitive and can be easily understood by policy makers and the general public and therefore these statistics form an important input in public debate and decision making. Limiting the number of available statistics, or confusing statistics by applying disclosure limitation or masking technologies, will have serious repercussions for practice.

In this paper we discuss some of the arguments which are important in the publication of the most commonly used mortality statistics in demography and epidemiology and estimate the risk of an intruder reverse calculating such statistics. In this paper we will estimate the risk of identifying a person's cause of death and we will make some subjective recommendation regarding the "safe" use of epidemiological statistics. Note, this is a paper about risks and not about risk-free, but in the context of mortality statistics we will show that the risks of disclosure in relation to mortality statistics are mostly infinitesimally small.

2 Results

2.1 General remark

The likelihood of reverse calculating statistics to disclose information about individuals is the issue discussed in this paper. Although this seems trivial it is important to remember that to recalculate something a calculation has to have taken place in the past on identifiable data, that the details of the calculation are known, and that the person doing the recalculating is knowledgeable about the procedure. For example, it is pointless to recalculate a statistic to find age specific death rates if the age specific death rates one seeks have not been used in the original calculation.

2.2 Indirectly standardized statistics

Indirectly standardized measures, of which the SMR is the most often used, are based on the application of the age specific death rates of a larger standard population on the age structure of a smaller index population. For example, there is a question how a community is doing for a certain cause of death compared with the national figure, corrected for differences in demography. It is not possible to obtain age specific death rates for the community by recalculating the SMR as they have not been used. Recalculating the SMR can only be used to obtain the total number of deaths in a community for the causes for which the SMR is known. In the Netherlands age specific death rates by age and gender for the country are available with a good level of detail for many causes, while for even small communities and demographic groups the age structure is readily available. It is easy to calculate the expected number of death for a certain cause of death for a certain community, and by multiplying this number by the SMR for this community the exact number of deaths can be obtained. It doesn't even seem necessary to go through all this recalculating. Applying the nationwide crude death rate to the community without considering age differences already gives an impression about the number of deaths to be expected, multiplying this crude number with the SMR makes this estimate a lot better, while recalculating the SMR makes it precise.

How serious is it that the number of deaths for certain causes can be precisely obtained for a community? It is not the number of people who die of particular causes which is important, but the diversity of causes of death in the group. If we know that one person died from a certain cause, while in this same group people also died from 5 other causes, the only thing we can say about one particular deceased is that the person died from any one of six causes. Some more, some less probable.

2.3 Directly standardized statistics

Directly standardized statistics include the directly standardized rates, the CMF and the YPLL. This discussion applies to all these directly standardized measures. In calculating directly standardized measures the age structure of the larger standard population is applied to the age specific death rates of the smaller community. If directly standardized statistics can be recalculated detailed information is obtained within age and gender groups in the community of the causes of death. Often in the age groups there will be insufficient diversity, or no diversity, of causes of death. This will particularly be true for younger age groups with a low mortality. In which case one only needs to know the age or age band of a deceased person to determine the cause of death on the basis of recalculated knowledge.

If the total number of deaths for a certain cause is known for a community, and we know the age structure of the community and of the larger standard population, it is possible by permutating the number of deaths over the age structure to produce a large number of combinations of deaths over the age categories which can be used to calculate expected CMF or YPLL statistics. The expected statistic which matches the observed CMF or YPLL for the community, digit by digit, shows the combination of deaths over the age categories which produced the observed statistic. A large number of the generated combinations will be theoretically unlikely, for example a combination which states that all babies died of breast cancer. By introducing rules the number of combinations can be reduced and the likelihood of a match which discloses the deaths by age improved. The number of unique combinations C for n death over k categories is given by [11]:

$$C = x!/[n!(x-n)!], \text{ whereby } x=n+k-1 \quad (1)$$

With two deaths over ten age groups the number of unique combinations equals 55. By 5 deaths over 20 age groups there are 42,504 combinations, by 10 deaths over 20 age group there are 20,030,010 combinations. Another factor in finding age specific death rates is the precision of the standardized measure with which the combinations are compared. If the standardized measure consists of four meaningful digits, thus any value between 0 and 999.9, a percentage CMF with one decimal number for example, then 42,504 expected combinations have to be rounded to on average 42,504/10,000=4.25 possible matches for an observed CMF or YPLL.

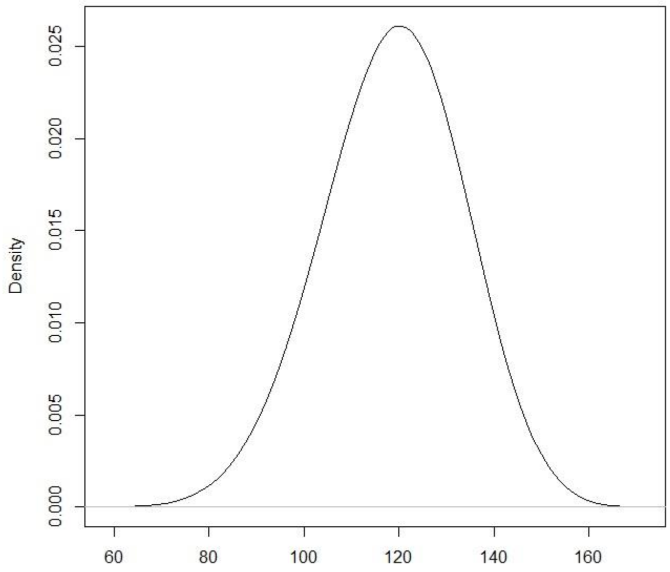

Figure 1. Density of expected number death.

Some expected outcomes are more likely than other outcomes. In the case of the CMF values near unity (1 or 100%) are more likely. Figure 1 shows the density of the expected number of deaths after permutating 5 deaths over 20 age categories of the population of Amsterdam and applying the age specific mortality ratios to the population of the Netherlands. The expected number of deaths in the Netherlands equals:

PopulationNL/PopulationAms*5=16829289/811184*5=103.7.

Lastly, recalculating directly standardized measures presumes prior knowledge. If the age structure used in the original calculation is not known then recalculating becomes nearly impossible because there is no knowledge of the age structure over which to permutate the number death. If the working hypothesis is that some categorization of the original single age year mortality table has been used, and this is considered in the recalculation procedure, the number of categorizations to investigate equals a Bell number, the sum of sterling number of the second kind [12]. For example, a 105 row single year mortality table can be categorized in 1.17E+122 different ways. If only age years which are adjacent are combined (11 and 12 are adjacent, 11 and 13 are not adjacent), the number of possible categorizations to investigate is given by:

$$Cr = 2^{(r-1)} \text{ ; Cr is the n of combinations of adjacent rows in r rows} \qquad (2)$$

In which case a 105 row table can be categorized in 2.03E+31 different ways. If the number of combinations is to be further reduced formulae 3, based on the binomial coefficient, can be used to reduce the number of adjacent rows to between minc and maxc categories.

$$Cr = \sum_{k=minc}^{maxc} \binom{r-1}{k-1} \qquad (3)$$

Say that there is a reason to believe that a 105 rows mortality table was reduced to between 10 and 20 adjacent rows, that can be done in 8.34E+19 ways. Some form of prior knowledge is always required in reverse calculating statistics. If for example age of death could also be measured in months, days, or three-quarter years since birth, and that varies, the number of possibilities becomes infinite.

3 Conclusions

Our article is an attempt to look at the disclosure risks inherent in basic mortality statistics by calculating disclosure probabilities. Generally these probabilities will be very small in the case of mortality statistics as long as a few minimum precautions are observed. Indirectly standardized measures such as the SMR which are calculated on the basis of detailed mortality tables can in principle be published without concern for disclosing the cause of death of an individual in smaller communities. The exception is very small communities were the diversity of different causes of death is small. ONS [3] allows publication of mortality statistics when the size of a community is larger than 5,000. In the context of the argument of this paper that seems to be a very reasonable precaution. For males in such a community one would expect on the basis of ONS statistics [13] about 58 deaths annually, age standardized, with an expectation of at least one death in each of the 9 leading causes of death. For females one would expect 43 deaths, with at least one death in each of the 8 leading causes of death.

Directly standardized statistics can be published in any case where there is no knowledge of how the original calculation has been done. It is important that the original calculation is not implicitly known, for example by using five or ten year age categories. If some kind of random or otherwise unexpected age structure is used in the original calculation the method of recalculating the data becomes very complex and the number of combinations extremely large. If there is knowledge of the original calculation directly standardized statistics can be published without concern when the total number of death of a certain cause is a sufficient number, a relatively large number of age categories is used in the original calculation, and the directly standardized statistic which is published has relatively few meaningful digits. The number of combination given a certain age structure and number of death divided by the number of significant digits in the statistic of interest gives the number of solutions which can be expected for each observed CMF or YPLL. When this is above a certain number (say 5 or 10) the CMF or YPLL can be published. For extreme values of a statistics however this number must be higher, for values near unity it can be lower, considering a density graph as published in this paper.

Lastly, although we are in full support of disclosure control we feel that in the case of mortality statistics disclosure prevention techniques are applied to prevent extremely small disclosure risks which can in practice only be realized by a very knowledgeable intruder doing a great technical and time investment. The price to pay for the prevention of this small risk is that great and easy to understand statistics which are an important input in health policy development and the public discussion become unavailable or unreliable.

Disclaimer and acknowledgements

None.

References

[1] Unknown. (2001) Introduction. In: Doyle, P., Lane, J. I., Theeuwes, J. J., & Zayatz, L.V. Confidentiality, disclosure, and data acces: theory and practical applications for statistical agencies.

[2] Nabar, S.U. Mishra, N. (2010) Releasing Private Contingency Tables. Journal of Privacy and Confidentiality 2:109–140.

[3] Office for National Statistics (ONS) (2014) Disclosure control guidance for birth and death statistics. http://www.ons.gov.uk/ons/guide-method/best-practice/disclosure-control-policy-for-birth-and-death-statistics/index.html

[4] Fellegi, I.P. Phillips, J.J. (1974) Statistical Confidentiality Some Theory and Application to Data Dissemination. Annals of Economic and Social Measurement 3:399-409. http://www.nber.org/chapters/c10117.pdf

[5] Gonzalez, J.F. (2009) Disclosure Limitation Techniques for Tabular Data. Section on Statistics in Defense and National Security – JSM 2009:5181-5190.

[6] Matthews, G., Harel, O. (2011) Data confidentiality A review of methods for statistical disclosure limitation and methods for assessing privacy. Statistics Surveys 5:1–29.
https://projecteuclid.org/download/pdfview_1/euclid.ssu/1296828958.

[7] A Centers for Disease Control and Prevention (CDC) (2005) CDC/ATSDR Policy on Releasing and Sharing Data, CDC-GA-2005-14. http://www.cdc.gov/maso/policy/releasingdata.pdf

[8] Centraal Bureau voor de Statistiek (CBS) (2011) Richtlijnen voor On Site/ Remote Access output. http://www.cbs.nl/NR/rdonlyres/8057A6F0-B9C7-46E8-A294-16E61CAAEFCC/0/2011richtlijnenonsiteremoteaccessoutput.pdf

[9] Gomatam, Shanti, et al. "Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers." Statistical Science (2005): 163-177.

[10] Reiter, Jerome P. "Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences." Public

opinion quarterly 76.1 (2012): 163-181. (het effect van bewerkte tabellen op statistieken)

[11] Ruskey, R. (2001) Combinatorial Generation. Victoria University of Victoria http://www.1stworks.com/ref/RuskeyCombGen.pdf

[12] Wikipedia. (201*) Stirling numbers of the second kind http://en.wikipedia.org/wiki/Stirling_numbers_of_the_second_kind.

[13] Office for National Statistics (ONS) (2014) Deaths Registered in England and Wales http://www.ons.gov.uk/ons/dcp171778_381807.pdf