Table 6 Cause-specific mortality fraction estimates when the first four symptoms are removed

		sequentially removed				
cause of death	all 51 symptoms	fever	pale	confused	wheezing	true
HIV	0.146	0.177	0.164	0.185	0.190	0.227
Malaria	0.073	0.082	0.084	0.091	0.101	0.089
Tuberculosis	0.063	0.071	0.086	0.073	0.077	0.035
Infectious diseases	0.058	0.047	0.050	0.044	0.044	0.028
Circulatory diseases	0.225	0.215	0.159	0.163	0.160	0.163
Maternal diseases	0.035	0.026	0.030	0.033	0.032	0.035
Cancer	0.042	0.033	0.033	0.039	0.041	0.092
Respiratory diseases	0.070	0.079	0.070	0.073	0.053	0.046
Injuries	0.039	0.034	0.028	0.023	0.031	0.050
Diabetes	0.113	0.108	0.150	0.133	0.139	0.053
Other diseases I*	0.023	0.031	0.030	0.027	0.027	0.032
Other diseases II	0.170	0.164	0.173	0.159	0.162	0.149

*Note: "other disease I" includes diseases in residual category that are related to internal organs

Table 5 indicates that there was a substantial difference in the prevalence of wheezing between community and hospital deaths (8.3% and 21.3%, respectively). This would mean that the wheezing observed in the community may be highly biased due to the respondents' difficulty in understanding the symptom correctly. Other symptoms which can be considered unbiased, such as vomiting, difficulty in swallowing, and chest pain, are all clinically useful in distinguishing the 12 causes of death, and the difference in prevalence between community and hospital deaths was much smaller.

0.4 Adjusting for Sample Differences

Suppose the key assumption of the King-Lu method is violated because of known differences in the hospital and community samples. For example, it may be that through outreach efforts of hospital personnel, or because of sampling choices of the investigators, children are overrepresented in the hospital sample. Even if the key assumption applies within each age group, diseases will present differently on average in the two samples because of the different age compositions. When it is not feasible to avoid this problem by desiging a proper sampling strategy, we can still adjust ex post to avoid bias, assuming the sample is large enough. The procedure is to estimate the distribution of symptoms for each cause of death within each age group separately, instead of once overall, and then to weight the results by the age distribution in the community.

As Appendix A shows in more detail, this procedure can also be applied to any other variables with known distributions in the community, such as sex or education. Since variables like these are routinely collected in verbal autopsy interviews, this adjustment should be easy and inexpensive, and can be powerful.

Reducing Inefficiencies

We now suppose that the analyst has chosen symptom questions as best as possible to minimize bias, and study what can be done to improve statistical efficiency. Efficiency translates directly into our uncertainty estimates about the CSMR, such as the standard error or confidence interval, and improving efficiency translates directly into research costs saved. We study efficiency by simulating a large number of data sets from a fixed population and measure how much estimates vary. We then study how efficiency responds to changes in the number of symptom questions (10, 20, 30, and 50), size of the hospital sample (500, 1,000, 2,000, 3,000, and 5,000), size of the community sample (1,000, 2,000, 3,000, 5,000, and 10,000), and number of chosen causes of death (5, 10, and 15). Causes of death in the hospital are assumed to be constructed via a case-control methods, with uniform CSMR across causes. The CSMR in the community for each cause of death has some prevalent causes and some rarer causes.

For each combination of symptoms, hospital and community samples, and number of causes of death, we randomly draw 80 data sets. For each, we compute the absolute error between the King-Lu estimate and the truth per cause of death, and finally average over the 80 simulations. This *mean absolute error* appears on the vertical axis of the three graphs (for 5, 10, and 15 causes of death reading from left to right) in Figure 3. The horizontal axis codes both the hospital sample size and, within categories of hospital sample size, the community sample size. Each of the top four lines represent different numbers of symptoms.

The lower line, labeled "direct sampling," is based on an infeasible baseline estimator, where the cause of each randomly sampled community death is directly ascertained and the results are tabulated. The error for this direct sampling approach is solely due to sampling variability and so serves as a useful lower error bound to our method, which includes both sampling variability and error due to extrapolation between the hospital and population samples. No method with the same amount of available information could ever be expected to do better than this baseline.

Figure 3 illustrates five key results. First, the mean absolute error of the King-Lu method is never very large. Even in the top left corner of the figures, with 500 deaths in the hospital, 1,000 in the community, and only 10 symptom

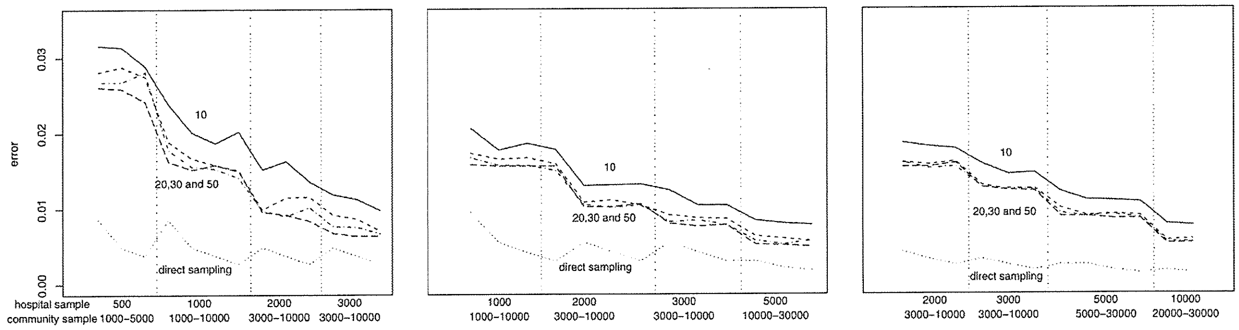

Figure 3 Simulation Results for 5 (left graph), 10 (middle), and 15 (right) causes of death.

questions, the average absolute error in the cause-of-death categories is only about 2 percentage points, and of this 0.86, of a percentage point is due to pure irreducible sampling variability (see the direct sampling line).

Second, increasing the number of symptom questions, regardless of the hospital and community sample sizes, reduces the mean absolute error. So more questions are better. However, the advantage in asking more than about 20 questions is fairly minor. Asking five times more questions (going from 10 to 50) reduces the mean absolute error only by between 15% and 50%. Our simulation is using symptom questions that are statistically independent, and so more questions would be needed if different symptoms are closely related; however, additional benefit from more symptoms would remain small.

Third, the mean absolute error drops with the number of deaths observed in the community, and can be seen for each panel separated by vertical dotted lines in each graph. Within each panel, the slope of each line drops at first and then continues to drop but at a slower rate. This pattern reflects the diminishing returns of collecting more community observations, so increasing from 5,000 to 10,000 deaths does not help as much as increasing the sample size from 1,000 to 3,000.

Fourth, mean absolute error is also reduced by increasing the hospital sample size. Assuming since data collection costs will usually keep the community sample larger than the hospital sample, increasing the hospital sample size will always help reduce bias. Moreover, within these constraints, the error reduction for collecting an extra hospital death is greater than that in the community. The reason for this is that the method estimates only the marginal distribution of symptom profile prevalences from the community data, whereas it estimates this distribution within each category of deaths in the hospital data. It is also true that the marginal gain in the community sample is constrained to a degree by the sample size in the hospital; the reason is that combinations of symptoms in the community can only be analyzed if examples of them are found in the hospital.

And finally, looking across the three graphs in Figure 3, we see that the mean absolute error per cause drops slightly or stays the same as we increase the number of causes of death. Of course, this also means that with more causes, the total error is increasing as the number of causes increase. This is as it should be because estimating more causes is a more difficult inferential task.

The results in this section provide detailed information on how to design verbal autopsy studies to reduce error. Researchers can now pick an acceptable threshold level for the mean absolute error rate (which will depend on the needs of their study and what they want to know) and then choose a set of values for the design features that meets this level. Since the figure indicates that different combinations of design features can produce the same mean absolute error level, researchers have leeway to choose from among these based on convenience and cost-effectiveness. For example, although the advantage of asking many more symptom questions is modest, it may be that in some areas extra time spent with a respondent is less costly than locating and traveling to another family for more interviews, in which case it may be more cost-effective to ask 50 or more symptom questions and instead reduce the number of interviews. Figure 3 provides many other possibilities for optimizing scarce resources.

Discussion

Despite some earlier attempts at promoting standard tools [20], which have now been adopted by various users including demographic surveillance sites under the INDEPTH Network [19,21,22], little consensus existed for some time on core verbal autopsy questions and methods. In order to derive a set of standards and to achieve a high degree of consistency and comparability across VA data sets, a recent WHO-led expert group recently systematically reviewed, debated, and condensed the accumulated experience and evidence from the most widely-used and validated procedures. This process resulted in somewhat more standardized tools, which are now adopted by various users, including demographic surveillance sites.

Verbal autopsy methodologies are still evolving: several key areas of active and important research remain. A research priority must be to carry out state-of-the-art validation studies of the new survey instruments in multiple countries with varying mortality burden profiles, using the methods discussed and proposed here. Such a validation process would contribute to the other areas of research, including further optimization of items included in questionnaires; replicable and valid automated methods for assigning causes of death from VA that remove human bias from the cause-of-death assignment process; and such important operational issues as sampling methods and sizes for implementing VA tools in research demographic surveillance sites, sample or sentinel registration, censuses, and household surveys.

With the advice we offer here for writing symptom questions, weeding out biased questions, and choosing appropriate hospital and community sample sizes, researchers should be able to greatly improve their analyses, reducing bias and research costs. We encourage other researchers and practitioners to use these tools and methods, to refine them, and to develop others. With time, this guidance and experience ought to better inform the VA users, and enhance the quality, comparability, and consistency of causespecific mortality rates throughout the developing world.

Appendix A

The King-Lu Method and Extensions

We describe here the method of estimating the CSMR offered in [10]. We give some notation, the quantities of interest, a simple decomposition of the data, the estimation strategy, and a procedure for making individual classifications when useful.

Notation

Hospital deaths may be selected randomly, but the method also works without modification if they are selected via case-control methods whereby, for each cause, a fixed number of deaths are chosen. Case-control selection can greatly increase the efficiency of data collection. Deaths in the community need to be chosen randomly or in some representative fashion. Define an index i (i = 1,..., n) for each death in a hospital and ℓ (ℓ =1,..., L) for each death in the community. Then define a set of mutually exclusive and exhaustive causes of death, 1,..., J, and denote D_ℓ as the observed cause for a death in the hospital and D_i as the unobserved cause for a death in the community. The verbal autopsy survey instrument typically includes a set of K symptom questions with dichotomous (0/1) responses, which we summarize for each decedent in the hospital as a $K \times 1$ vector $\boldsymbol{S}_i = \{S_{i1},..., S_{iK}\}$ and in the community as $\boldsymbol{S}_\ell = \{S_{\ell 1},..., S_{\ell K}\}$.

Quantity of Interest

For most public health purposes, primary interest lies not in the cause of any individual death in the community but rather the aggregate proportion of community deaths that fall into each category: $P(D) = \{P(D = 1),..., P(D = J)\}$, where $P(D)$ is a $J \times 1$ vector and each element of which is a proportion: $P(D=j)=\frac{1}{L}\sum_{\ell=1}^{L} 1(D_\ell = j)$, where $1(a) = 1$ if a is true and 0 otherwise. This is an important distinction because King-Lu gives approximately unbiased estimates of $P(D)$ even if the percent of individual deaths correctly classified is very low. (We also describe below how to use this method to produce individual classifications, which may of course be of interest to clinicians.)

Decomposition

For any death, the symptom questions contain 2^K possible responses, each of which we call a symptom profile. We stack up each of these profiles as the $2^K \times 1$ vector $\boldsymbol{S}$ and write $P(\boldsymbol{S})$ as the probability of each profile occurring (e.g., with K = 3 questions $P(\boldsymbol{S})$ would contain the probabilities of each of these (2^3 = 8) patterns occurring in the survey responses: 000, 001, 010, 011, 100, 101, 110, and 111). $P(\boldsymbol{S}|D)$ as the probability of each of the symptom profiles occurring within for a given cause of death D (columns of $P(\boldsymbol{S}|D)$ corresponding to values of D). Then, by the law of total probability, we write

$$P(S=s)=\sum_{j=1}^{J} P(S=s \mid D=j)P(D=j). \tag{1}$$

and, to simplify, we rewrite Equation 1 as an equivalent matrix expression:

$$\underset{2^K\times 1}{P(S)} = \underset{2^K\times J}{P(S\mid D)}\,\underset{J\times 1}{P(D)}. \tag{2}$$

where $P(D)$ is a $J \times 1$ vector of the proportion of community deaths in each category, our quantity of interest. Equations 1 and 2 hold exactly, without approximations.

Estimation

To estimate $\mathbf{P}(D)$ in Equation 2, we only need to estimate the other two factors and then solve algebraically. We can estimate $P(S)$ without modeling assumptions by direct tabulation of deaths in the community (using the proportion of deaths observed to have each symptom profile). The key issue then is estimating $P(S|D)$, which is unobserved in the community. We do this by assuming it is the same as in the hospital sample, $P^h(S|D)$:

$$P^h(S\mid D) = P(S\mid D). \tag{3}$$

This assumption is considerably less demanding than other data-derived methods, which require the full joint distribution of symptoms and death proportions to be the same: $P^h(\mathbf{S}, D) = P(\mathbf{S}, D)$. In particular, if either the symptom profiles (how illnesses that lead to death present to caregivers) or the prevalence of the causes of death differ between the hospital and community – $P^h(\mathbf{S}) \neq P(\mathbf{S})$ or $P^h(D) \neq P(D)$– then other dataderived methods will fail, but this method can still yield unbiased estimates. Of course, if Equation 3 is wrong, estimates from the King-Lu method can be biased, and so finding ways of validating it can be crucial, which is the motivation for the methods offered in the text. (Several technical estimation issues are also resolved in [10]: Because 2^K can be very large, they use the fact that (2) also holds for subsets of symptoms to draw multiple random subsets, solve (2) in each, and average. They also solve (2) for $P(D)$ by constrained least squares to ensure that the death proportions are constrained to the simplex.)

Adjusting for Known Differences Between Hospital and Community

Let a be an exogneous variable measured in both samples, such as age or sex. To adjust for differences in a between the two samples, we replace our usual estimate of $P^h(\mathbf{S}|D)$ with a weighted average of the same estimator applied within unique values of a, $P^h(\mathbf{S}_a|D_a)$, times the community distribution, $f(a) : P_a^h(S \mid D) = \sum_a P^h(S_a | D_a) f(a)$, and where the summation is over all possible values of a.

Individual Classification

Although the quantity of primary interest in verbal autopsy studies is the CSMR, researchers are often interested in classifications of some individual deaths. As shown in [[10], Section 8], this can be done with an application of Bayes theorem, if one makes an additional assumption not necessary for estimating the CSMR. Thus, if the goal is $P(D_\ell = j|\mathbf{S}_\ell = s)$, the probability that individual ℓ died of cause j given that he or she suffered from symptom profile s, we can calculate it by filling in the three quantities on the right side of this expression:

$$P(D_\ell = j \mid S_\ell = s) = \frac{P(S_\ell = s|D_\ell = j)P(D_\ell = j)}{P(S_\ell = s)}. \qquad (4)$$

First is $P(D_\ell = j)$, the optimal estimate of the CSMR, given by the basic King-Lu procedure. The quantity $P(\mathbf{S}_\ell = s|D_\ell = j)$ can be estimated from the training set, usually with the addition of a conditional independence assumption, and $P(\mathbf{S}_\ell = \mathbf{s}_\ell)$ may be computed without assumptions from the test set by direct tabulation. (Bayes theorem has also been used in this field for other creative purposes [23].)

A Reaggregation Estimator

A recent article [12] attempts to build on King-Lu by estimating the CSMR with a particular interpretation of Equation 4 to produce individual classifications which they then reaggregate back into a "corrected" CSMR estimate. Unfortunately, the proposed correction is in general biased, since it requires two unnecessary and substantively untenable statistical assumptions. First, it uses the conditional independence assumption for estimating $P(\mathbf{S}_\ell = s|D_\ell = j)$ –useful for individual classification but unnecessary for estimating the CSMR. And second, it estimates $P(\mathbf{S}_\ell = \mathbf{s})$ from the training set and so must assume that it is the same as that in the test set, an assumption which is verifiably false and unnecessary since it can be computed directly from the test set without error [[10], Section 8]. To avoid the bias in this reaggregation procedure to estimate the CSMR, one can use the original King-Lu estimator described above. Reaggregation of appropriate individual classifications will reproduce the same aggregate estimates.

Appendix B

A Test for Detecting Biased Symptoms

If symptom k, S_k, is overreported in the community relative to the hospital for a given cause of death, then we should expect the predicted prevalence $\hat{P}(S_k)$ – which can be produced by but is not needed in the King-Lu procedure–to be lower than the observed prevalence $P(S_k)$. Thus, we can view $P(S_k)$ as data point in regression analysis and the misreported prevalence of the kth symptom, $P(S_k)$ as an outlier. This means that we can detect symptoms that might bias the analysis by examining model fit. We describe this procedure here and then give evidence and examples.

Test Procedure

Let $\hat{P}(D)$ be the estimated community CSMRs via the King-Lu procedure, and then fit the marginal prevalence of the kth symptom in the community $\hat{P}(S_k)$ (calculated as $\hat{P}(S_k) = \sum_j P^h(S_k \mid D_j)\hat{P}(D_j)$; see Appendix A). Then define the prediction error as the residual in a regression as $e_k \equiv \hat{P}(S_k) - P(S_k)\cdot$

Under King-Lu, $\hat{P}(D)$ is unbiased, and each e_k is mean 0 with variance $V(e) = \sum_{k=1}^{K} (e_k - \overline{e})^2 / (K-1)\cdot$ Moreover,

$$t_k = \frac{e_k - \overline{e}}{\sqrt{\Sigma(e_k - \overline{e})^2} / \sqrt{K-1}}$$

is approximately t distributed with K - 1 degree freedom. If, on the other hand, $P^h(S|D) \neq P(S|D)$, we would expect t_k will have an expected value that deviates from zero.

Based on above observation, we propose a simple iterative symptom selection procedure. This procedure avoids the classic the "multiple testing problem" by applying a Bonferroni adjustment to symptom selection at each iteration. At the chosen significance level α, we can then assess whether a calculated value of t_k indicates that the k^{th} symptom violates our key assumption and so is biased. Thus:

1. Begin with the set of all symptoms $S = (S_1, ..., S_K)$, and those to be deleted, B. Initialize B as the null set, and the number of symptoms in the estimation, K_0, as $K_0 = K$.
2. Estimate $P(D)$ using the King-Lu method.
3. For symptom k ($k = 1, ..., K_0$), estimate $\hat{P}(S_k)$, and calculate t_k.
4. Find the critical value associated with level α under the t distribution with K_0 - 1 degree freedom, $C_{[\alpha/(\#(B)+1)],(K_0-1)}$. To ensure that the overall significance of the all symptoms that belong to B is less than α, use the Bonferroni adjustment for the significance level (the set of t_k test statistics associated with the symptoms in set B are stochastically independent of each other since the models are run sequentially). The number of the multiple independent tests is counted as the number of elements already in the set B plus the one that is being tested. (Since the maximum $|t_k|$ at each step of symptom selection tends to decrease as more "bad" symptoms are removed, it is sufficient to check whether the maximum $|t_k|$ of the current model is greater or less than the critical value, $C_{\alpha/(\#(B)+1),(K_0-1)}$.)
5. If the largest value of $|t_k|$, namely $|t_{k'}|$, is greater than the critical value $C_{[\alpha/(\#(B)+1)],K_0-1}$, remove the corresponding k' th symptom. Then set $K_0 = K_0 - 1$ and add symptom k' to set B.
6. Repeat step 2-5 until no new symptoms are moved from S to B.

Acknowledgements
Open source software that implements all the methods described herein is available at http://gking.harvard.edu/va. Our thanks to Alan Lopez and Shanon Peter for data access. The Tanzania data was provided by the University of Newcastle upon Tyne (UK) as an output of the Adult Morbidity and Mortality Project (AMMP). AMMP was a project of the Tanzanian Ministry of Health, funded by the UK Department for International Development (DFID), and implemented in partnership with the University of Newcastle upon Tyne. Additional funding for the preparation of the data was provided through MEASURE Evaluation, Phase 2, a USAID Cooperative Agreement (GPO-A-00-03-00003-00) implemented by the Carolina Population Center, University of North Carolina at Chapel Hill. This publication was also supported by Grant P10462-109/9903GLOB-2, The Global Burden of Disease 2000 in Aging Populations (P01 AG17625-01), from the United States National Institutes of Health (NIH) National Institute on Aging (NIA) and from the National Science Foundation (SES-0318275, IIS-9874747, DMS-0631652).

Author details
[1]Institute for Quantitative Social Science, Harvard University, Cambridge MA 02138, USA. [2]Department of Humanities and Social Sciences in the Professions, Steinhardt School of Culture, Education and Human Development, New York University, USA. [3]Department of Global Health Policy, Graduate School of Medicine, University of Tokyo, Japan.

Authors' contributions
GK, YL, and KS participated in the conception and design of the study and analysis of the results. YL wrote the computer code and developed the statistical test and simulations; GK drafted the paper. All authors read, contributed to, and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 8 January 2010 Accepted: 23 June 2010
Published: 23 June 2010

References

1. Mathers CD, Fat DM, Inoue M, Rao C, Lopez A: **Counting the dead and what they died from: an assessment of the global status of cause of death data.** *Bulletin of the World Health Organization* 2005, 83:171-177.
2. Lopez A, Ahmed O, Guillot M, Ferguson B, Salomon J, *et al*: **World Mortality in 2000: Life Tables for 191 Countries.** Geneva: World Health Organization 2000.
3. Sibai A, Fletcher A, Hills M, Campbell O: **Non-communicable disease mortality rates using the verbal autopsy in a cohort of middle aged and older populations in beirut during wartime, 1983-93.** *Journal of Epidemiology and Community Health* 2001, 55:271-276.
4. Setel PW, Sankoh O, Velkoff V, Mathers CYG, *et al*: **Sample registration of vital events with verbal autopsy: a renewed commitment to measuring and monitoring vital statistics.** *Bulletin of the World Health Organization* 2005, 83:611-617.
5. Soleman N, Chandramohan D, Shibuya K: **WHO Technical Consultation on Verbal Autopsy Tools.** *Geneva* 2005 [http://www.who.int/healthinfo/statistics/mort verbalautopsy.pdf].
6. Thatte N, Kalter HD, Baqui AH, Williams EM, Darmstadt GL: **Ascertaining causes of neonatal deaths using verbal autopsy: current methods and challenges.** *Journal of Perinatology* 2008, 1-8.
7. Anker M: **Investigating Cause of Death During an Outbreak of Ebola Virus Haemorrhagic Fever: Draft Verbal Autopsy Instrument.** Geneva: World Health Organization 2003.
8. Pacque-Margolis S, Pacque M, Dukuly Z, Boateng J, Taylor HR: **Application of the verbal autopsy during a clinical trial.** *Social Science Medicine* 1990, 31:585-591.
9. Soleman N, Chandramohan D, Shibuya K: **Verbal autopsy: current practices and challenges.** *Bulletin of the World Health Organization* 2006, 84:239-245.
10. King G, Lu Y: **Verbal autopsy methods with multiple causes of death.** *Statistical Science* 2008, 23:78-91 [http://gking.harvard.edu/files/abs/vamc-abs.shtml].
11. Todd J, Francisco AD, O'Dempsey T, Greenwood B: **The limitations of verbal autopsy in a malaria-endemic region.** *Annals of Tropical Paediatrics* 1994, 14:31-36.
12. Murray CJ, Lopez AD, Feean DM, Peter ST, Yang G: **Validation of the symptom pattern method for analyzing verbal autopsy data.** *PLOS Medicine* 2007, 4:1739-1753.
13. Chandramohan D, Rodriques LC, Maude GH, Hayes R: **The validy of verbal autopsies for assessing the causes of institutional maternal death.** *Studies in Family Planning* 1998, 29:414-422.
14. Coldham C, Ross D, Quigley M, Segura Z, Chandramohan D: **Prospective validation of a standardized questionnaire for estimating childhood mortality and morbidity due to pneumonia and diarrhoea.** *Tropical Medicine & International Health* 2000, 5:134-144.
15. Quigley M, Schellenberg JA, Snow R: **Algorithms for verbal autopsies: a validation study in kenyan children.** *Bulletin of the World Health Organization* 1996, 74:147-154.
16. Hand DJ: **Classifier technology and the illusion of progress.** *Statistical Science* 2006, 21:1-14.

17. Kalter H: **The validation of interviews for estimating morbidity.** *Health Policy and Planning* 1992, 7:30-39.
18. Levy P, Kass EH: **A three population model for sequential screening for bacteriuria.** *American Journal of Epidemiology* 1970, **91**:148-154.
19. Setel P, Rao C, Hemed Y, Whiting DYG, *et al*: **Core verbal autopsy procedures with comparative validation results from two countries.** *PLoS Medicine* 2006, **3**:e268, Doi:10.1371/journal.pmed.0030268.
20. World Health Organization: **Verbal Autopsy Standards: Ascertaining and Attributing Causes of Death.** Geneva: World Health Organization 2007.
21. Anker M, Black RE, Coldham C, Kalter HD, Quigley MA, *et al*: **A standard verbal autopsy method for investigating causes of death in infants and children.** *World Health Organization, Department of communicable Disease Surveillance and Response* 1999.
22. INDEPTH Network: *Standardised verbal autopsy questionnaire* 2003 [http://indepth-network.org].
23. Byass P, Fottrell E, Huong DL, Berhane Y, Corrah T, *et al*: **(34) Refining a probabilistic model for interpreting verbal autopsy data, volume 2006.** *Scandinavian Journal of Public Health* 26-31.

doi:10.1186/1478-7954-8-19
Cite this article as: King *et al.*: Designing verbal autopsy studies. *Population Health Metrics* 2010 **8**:19.

添付資料　7

Tracking China’s health reform

China's blood-banking system, as detailed by Xuerong Yu and colleagues.[6] Production of China's human resources depends on professional medical education. As described by Dong Xu and colleagues,[7] higher education in China has several purposes: global academic excellence, training professionals to service diverse Chinese populations, and re-engineering the professions to grapple with the onslaught of non-communicable diseases. How China harmonises these diverse aims will determine not only the contributions of its next generation of professionals but also very probably the success and sustainability of its ambitious reforms towards universal health coverage.

For these and other studies, more and more Chinese health data are increasingly accessible, as illustrated by the wealth of data on child mortality from the internet in Chinese that was analysed by Rudan and colleagues.[4] As reviewed by Yan Guo and colleagues,[8] the recent round of health reforms will be monitored by key indicators, just as the reforms were stimulated in part by results from earlier health surveys. While it is too early to draw conclusions, China has rapidly expanded medical insurance so that the recent national household health-survey found a dramatic turnaround: 87% of urban and rural populations reached by 2008.[9] In view of these developments, it seems likely that China will achieve universal insurance coverage well before 2020, albeit providing only limited benefits.

The Lancet's two China collections, in 2008 and 2010, illustrate the deepening engagement of Chinese academics in the global health sciences. The recent appointment of a *Lancet* Asia editor based in Beijing augers well for the journal's future publications on health in China. Just as China has much to learn from international science, China also has much to share for the mutual benefit of the entire global health community.

*Qide Han, *Lincoln Chen, Tim G Evans, William Summerskill*
Peking University Health Sciences Center, Beijing, China (QH); China Medical Board, Harvard University, Cambridge, MA 02138, USA (LC); WHO, Geneva, Switzerland (TGE); and *The Lancet*, London, UK (WS)
lincoln_chen@harvard.edu

We declare that we have no conflicts of interest.

1 Han QD, Chen L, Evans T, Horton R. China and global health. *Lancet* 2008; **372:** 1439–41.

2 CPC Central Committee and State Council. Opinions of the CPC Central Committee and the State Council on Deepening the Health Care System Reform. 2009. http://shs.ndrc.gov.cn/ygjd/ygwj/t20090408_271138.htm (accessed March 21, 2010).

3 Zhang J, Mauzerall DL, Zhu T, Liang S, Ezzati M, Remais JV. Environmental health in China: progress towards clean air and safe water. *Lancet* 2010; **375:** 1110–19.

4 Rudan I, Chan KY, Zhang JSF, et al, on behalf of WHO/UNICEF's Child Health Epidemiology Reference Group (CHERG). Causes of deaths in children younger than 5 years in China in 2008. *Lancet* 2010; **375:** 1083–89.

5 Yip WC-H, Hsiao W, Meng Q, Chen W, Sun X. Realignment of incentives for health-care providers in China. *Lancet* 2010; **375:** 1120–30.

6 Yu X, Huang Y, Qu G, Xu J, Hui S. Safety and current status of blood transfusion in China. *Lancet* 2010; published online March 26. DOI:10.1016/S0140-6736(10)60003-7.

7 Xu D, Sun B, Wan X, Ke Y. Reformation of medical education in China. *Lancet* 2010; published online March 26. DOI:10.1016/S0140-6736(10)60241-3.

8 Guo Y, Shibuya K, Cheng G, Rao K, Lee L, Tang S. Tracking China's health reform. *Lancet* 2010; **375:** 1056–58.

9 Center for Health Statistics and Information, Ministry of Health, Government of People's Republic of China. Summary analysis report of National Health Services Survey in China, 2008. Beijing: Center for Health Statistics and Information, 2009.

Tracking China's health reform

See Comment pages 1055

In response to growing public concerns over widening inequalities in health, the Chinese Government officially launched another round of the national health system reform plan in early 2009 with a commitment of ¥850 billion (US$125 billion) for the next 3 years.[1] The proposed reform has ambitious targets, including 90% health insurance coverage by the end of 2010 and universal coverage of essential health-care by 2020. But how will these reforms be monitored?

Over the past three decades, China's health information system has evolved to include national household surveys of health-care use and expenditures, real-time surveillance of communicable diseases, and periodic disease-specific prevalence surveys.[2] These health metrics were key factors in the current reforms. With use of the national health service surveys in 1993 and 1998, the Development Research Center concluded in 2002–04 that earlier reforms had not been successful.[3] The proportion of health expenditures covered by the government had dropped from 36% to 15%, funding for prevention had relatively declined, health insurance coverage had collapsed to 13% of rural and 40% of urban residents, and 70% of residents from the poorest regions of the country failed to seek inpatient treatment because of cost. In July, 2005, a newspaper article[4] about the study in the *China Youth Daily* provoked widespread public discussion and debate that was sustained by a litany of media stories of human and financial health-system failings.

Three major metrics challenges must be met to ensure that the health reform is more successful than previous efforts. The quality and comprehensiveness of health metrics need to be gradually improved; the analytical capacity for integrating several types of data from national surveys, facility-based routine reporting, and national disease surveillance needs to be strengthened; and the traditional top-down process for policy formulation needs to become increasingly evidence-based, driven by the commitment of decision makers to make use of available information.

At a conference in Beijing in December, 2009, China's Ministry of Health Center for Health Statistics and Information announced the development of 19 indicators (panel) to track six components of the health reform: access, quality, cost, financial protection, patients' satisfaction, and health improvement.[5] These indicators will be constructed by integration of data from population-based sample surveys, vital events data, disease-focused population surveys, and facility-based administrative data. Identification of such indicators is an important first step but there remain serious concerns about the quality, reliability, and validity of some of the data.

Most of the primary data are recorded by government workers who have been trained for other functions, so systematic retraining and supervision to enforce rigorous data collection will be essential. Abundant data from the existing health-information system should be sufficient to monitor changes in insurance coverage, disease prevalence, and health services, but little information is available about supply, pricing, distribution, and use of drugs—the fourth major subsystem being addressed. China's State Food and Drug Administration has not yet established an effective information system and data from the pharmaceutical industry are scant. New data collection systems will be needed here.

A second challenge is analytical capability. Integration of the huge volume of data will require a large network of sophisticated analysts. Although not unique to China, a major constraint in China is the lack of qualified human resources and imbalances in the strength of relevant academic disciplines. Chinese institutions have fairly solid capabilities in statistics, epidemiology, and survey research. Less well-developed areas include health economics, policy analysis, sociobehavioural research, and monitoring and evaluation. Moreover, most of China's institutional strengths are based in government departments and agencies. Academic centres and autonomous research institutes that focus on health metrics are few and underdeveloped.

Panel: 19 indicators for monitoring and evaluation of China's health system reform

Access
- Percentage of population covered by health insurance
- Health expenditure as percentage of total household income
- Percentage of patients referred for hospital admission but not admitted because of cost
- Percentage of households reaching a health provider within 20 min

Quality
- Percentage of patients with hypertension being followed-up every 3 months
- Percentage of pregnant women receiving antenatal care
- Hepatitis B immunisation coverage
- Percentage of patients receiving intravenous drip administration

Cost
- Average expenditure of hospital admission at different levels
- Average expenditure of outpatient visit at county, township, and village levels

Improved health
- Maternal mortality
- Infant mortality
- Tuberculosis morbidity
- Mortality of hypertension-apoplexy or stroke

Patient satisfaction
- Percentage of patients satisfied with medical care
- Percentage of patients satisfied with medical care by income group

Financial risk protection
- Out-of-pocket payment as percentage of total national health expenditure
- Percentage of household with catastrophic expenditure of medical care
- Rate of households living in poverty because of medical care cost

Finally, as in all countries, improvements in the quality, comprehensiveness, and analysis of health data will not necessarily lead to more effective policy making. Health data are mainly gathered and used by different government departments, so there are issues about resolving the inconsistency of data provided by different governmental sources and, more importantly, of providing access to the data to other non-governmental actors. Although generally moving towards greater openness and wider access, routes of disseminating government-collected health data through journals, reports, websites, press briefings, and other media are not well established. Provision of full public access to reports about the progress of the reforms and use of data monitoring to make continuing adjustments in reforms will require changes in the traditional process of decision making by government. The transition will be difficult for health-care policy makers in China, but the transparency encouraged during the development of these reforms must be sustained and expanded to ensure success.

In health metrics, China has learned much from other countries and international agencies.[6] In view of its growing role in global health, China's efforts to track its health reform are likely to generate valuable lessons for others.

*Yan Guo, Kenji Shibuya, Gang Cheng, Keqin Rao, Liming Lee, *Shenglan Tang*
School of Public Health, Peking University Health Science Centre, Beijing, China (YG, GC, LL); University of Tokyo, Tokyo, Japan (KS); Ministry of Health, Beijing, China (KQ); TDR/WHO, 1211 Geneva 27, Switzerland; and Liverpool School of Tropical Medicine, Liverpool, UK (ST)
tangs@who.int

We declare that we have no conflicts of interest.

1 CPC Central Committee and State Council. Opinions of the CPC Central Committee and the State Council on Deepening the Health Care System Reform. 2009. http://shs.ndrc.gov.cn/ygjd/ygwj/t20090408_271138.htm (accessed March 21, 2010).

2 Ministry of Health. The national health statistics and survey systems. Beijing: Peking Union Medical College Press, 2007.

3 The State Council Research Group. Evaluations and advice on health reforms in China. *China Development Comments* 2005 (supp 1): 1–159.

4 Wang J. China's health reform not successful (in Chinese). *China Youth Daily* July 29, 2005. http://zqb.cyol.com/content/2005-07/29/content_1150962.htm (accessed March 11, 2010).

5 Xu L. The Seinor Level Seminar on National Health Services Survey and Health Reform M & E of China. Beijing, China. Dec 2–3, 2009.

6 Murray CJL, Frenk J. Health metrics and evaluation: strengthening the science. *Lancet* 2008; **371:** 1191–99.

Unravelling the enigma of health statistics in China

See Comment pages 1055
See Articles page 1083

China accounts for 19·6% of the global population, and is now the power house of the global economic recovery. However, much health-related information about China is based on official reporting systems and few, if any, externally verified surveys. Several of the UN estimates for health indicators from China are also based on projections from a limited number of datasets, with correction factors for under-reporting.[1] The available information for China and its health and nutrition indicators thus varies greatly, with not much subnational data on inequity and social determinants of health.[2]

In *The Lancet* this week, Igor Rudan and colleagues[3] from the Child Health and Epidemiology Reference Group (CHERG) present information from a comprehensive analysis of hitherto unexplored Chinese scientific publications and data from the national surveillance systems. Rudan and colleagues uncovered a treasure trove of Chinese language publications from representative community settings in China, and analysed the information with standard CHERG methods to generate information about trends and causes of neonatal and child mortality rates. The results are striking and provide robust information about cause-specific child mortality estimates, trends, and differentials. The information is important, and although the findings are broadly similar to those obtained from a recent review based on the national Maternal and Child Mortality Surveillance System (MCMS), some aspects differ from official data.[4,5] The mortality denominators are also quite different. For example, the number of deaths caused by rotavirus infections in China was estimated to be about 10 400 per year in 2004.[6] Today's analysis clearly suggests that, with fewer than 10 000 deaths per year associated with all-cause diarrhoea in China, the previous estimates were probably three times higher.

Several limitations need to be emphasised in the approach and analysis by Rudan and colleagues. The data quality could not be thoroughly assessed, other than the congruence of the final estimates with available data and plausibility. Despite peer-reviewed scientific publications from China increasing by 60-fold since 1981,[7] opinions differ about the scientific quality and vetting of such publications.[8,9] The mortality rates were adjusted from available MCMS and Gapminder data, but could well be underestimates because the official neonatal mortality rate is only calculated for infants after gestations of 28 weeks or longer. Additionally, the mortality data also

添付資料 8

Deaths from heart failure: Using coarsened exact matching to correct cause of death statistics

Stevens *et al. Population Health Metrics* 2010, **8**:6
http://www.pophealthmetrics.com/content/8/1/6

POPULATION HEALTH METRICS

RESEARCH Open Access

Deaths from heart failure: using coarsened exact matching to correct cause-of-death statistics

Gretchen A Stevens*[1], Gary King[2] and Kenji Shibuya[3]

Abstract

Background: Incomplete information on death certificates makes recorded cause-of-death data less useful for public health monitoring and planning. Certifying physicians sometimes list only the mode of death without indicating the underlying disease or diseases that led to the death. Inconsistent cause-of-death assignment among cardiovascular causes of death is of particular concern. This can prevent valid epidemiologic comparisons across countries and over time.

Methods: We propose that coarsened exact matching be used to infer the underlying causes of death where only the mode of death is known. We focus on the case of heart failure in US, Mexican, and Brazilian death records.

Results: Redistribution algorithms derived using this method assign the largest proportion of heart failure deaths to ischemic heart disease in all three countries (53%, 26%, and 22% respectively), with larger proportions assigned to hypertensive heart disease and diabetes in Mexico and Brazil (16% and 23% vs. 7% for hypertensive heart disease, and 13% and 9% vs. 6% for diabetes). Reassigning these heart failure deaths increases the US ischemic heart disease mortality rate by 6%.

Conclusions: The frequency with which physicians list heart failure in the causal chain for various underlying causes of death allows for inference about how physicians use heart failure on the death certificate in different settings. This easy-to-use method has the potential to reduce bias and increase comparability in cause-of-death data, thereby improving the public health utility of death records.

Background

Effective national and international public health planning and policymaking requires accurate information on population health, especially about deaths and their causes. Death statistics can provide evidence to evaluate health reforms and to identify poorly served populations or diseases. In countries with complete or nearly complete vital registration, including most high-income and some middle-income countries, death statistics are compiled from death certificates. However, inaccurately or incompletely completed death certificates may compromise cause-of-death data in these countries. Physician practice in filling death certificates may vary over place and time [1]. This may result in death rates calculated from death certificate data that are biased or are not comparable across regions, countries, or over time. Inconsistent cause-of-death assignment among cardiovascular causes of death is particularly important, as cardiovascular causes are the leading cause of death, causing 29% of deaths worldwide [2].

Certifying physicians sometimes complete death certificates incorrectly for cardiovascular deaths. Causes such as heart failure and cardiac arrest are routinely used in ways that violate standard protocols. For public health purposes, the underlying cause of death (UCD), as defined by the World Health Organization, should be "the disease or injury which initiated the train of morbid events leading directly to death, or the circumstances of the accident or violence which produced the fatal injury," a definition that is also the most useful for public health monitoring and planning [3]. The UCD listed on the death certificate may be incorrect because of 1) an incorrect diagnosis, or 2) incomplete cause-of-death information. In the second case, the certifying physician often lists only the mode of dying, such as cardiac or respiratory arrest, shock, or heart failure. The World Health

* Correspondence: stevensg@who.int
[1] Information, Evidence and Research, World Health Organization, 20 Avenue Appia, 1211 Geneva, Switzerland
Full list of author information is available at the end of the article

Organization's International Statistical Classification of Diseases and Related Health Problems (ICD) specifies that the mode of death should never be designated as the UCD if another plausible cause is listed on the death certificate [3]. Yet certifying physicians regularly list only the mode of dying due to uncertainty about the UCD or lack of knowledge or interest in correct procedures for completing a death certificate [4,5]. Among cardiovascular deaths in the US, 6% are certified to heart failure and 2% to cardiac arrest (Table 1). In Mexico, a middle-income country with a high-quality death registration system, 8% of cardiovascular deaths are assigned to heart failure; in Brazil, 10%. Our goal is to redistribute these deaths into the categories to which they belong.

One way to learn how to redistribute these deaths is to compare hospital records or autopsy findings to the cause of death listed on death certificates. Such studies, which have been carried out in the US [6,7] and elsewhere [8,9], often find substantial discrepancies between the death certificate, physician review of hospital records, and autopsy findings. However, these studies are limited by financial and practical constraints. Deaths that occur in hospitals and those selected for autopsy are likely to systematically differ from deaths that do not occur in hospitals and those that are not autopsied. Autopsies, which have been declining in the US and elsewhere [9], may be more likely in difficult-to-diagnose deaths [7], and therefore could be more likely to find less common diseases as the underlying cause of death.

Statistical methods provide an alternative to the autopsy for correcting cause-of-death statistics. Researchers have developed algorithms to redistribute deaths certified to causes that are unspecified or that cannot be underlying causes of death (hereafter referred to as ill-defined causes). Deaths can be reassigned based on expert knowledge of disease etiology, using an empirical basis, or by some combination of the two. A variety of approaches have been taken, including pro-rata redistribution [10], ecological regression analysis [11], and multinomial logistic regression of individual-level data [12]. Aside from pro-rata redistribution, each of these methods requires expert judgment to select the causes to which deaths are redistributed (or target causes of death).

Table 1: Frequency of selected cardiovascular and other causes as the underlying cause of death, US, Mexico, and Brazil.

Cause	ICD-10 codes	US (%)	Mexico (%)	Brazil (%)
Total number of death records		**14,500,497**	**920,517**	**3,033,240**
Lower respiratory infections	J10-J18, J20-J22	2.6	3.6	3.6
Cancers	C00-C97	22.9	13.4	13.8
Diabetes	E10-E14	3.0	13.4	3.9
All Cardiovascular diseases	I00-I99	38.1	23.3	27.8
Ischemic heart disease	I20-I25	20.5	10.5	8.4
Cerebrovascular disease	I60-I69	6.7	5.5	8.9
Hypertensive heart disease	I10-I13	2.0	3.0	3.0
Cardiomyopathy	I42-I43	1.1	0.2	1.3
Heart failure	I50	2.3	1.9	2.7
Cardiac arrest	I46, I47.2, I49.0	0.8	0.1	0.1
Other cardiovascular diseases	balance of I00-I99	4.6	2.1	3.3
Chronic obstructive pulmonary disease (COPD)	J40-J44	4.9	3.9	3.4
Digestive diseases	K20-K92	3.5	9.8	4.8
Other diseases		25.0	32.7	42.8

Data compiled from death certificates usually contain both the sequence of conditions that lead to the death and other contributing conditions, called multiple causes of death (MCDs). Unlike the previous approaches, which rely only on underlying cause-of-death data, MCD data allow for an empirical basis to select redistribution targets, and have been used to improve geographic comparability in the use of diabetes as an underlying cause of death using multinomial logistic regression [13]. An empirical redistribution algorithm may result in targets that are not expected based on pathophysiology, but may reflect how modes of death such as heart failure are used in practice.

Though multinomial regression has been used in the past, nonparametric methods are ideal for death certificate data. Multinomial regression requires strong assumptions about how variables are related, which are often violated. It also requires that target causes be broad and distinct (i.e., it limits detailed information about causes to which ill-defined deaths are redistributed). In contrast, nonparametric methods require weaker assumptions and allow for detailed information on target causes. We propose that coarsened exact matching [14], a nonparametric method, be used with MCD data to generate a redistribution algorithm for deaths certified to heart failure (ICD-10 cause I50). The method is demonstrated and validated using death records from two middle-income countries and one high-income country -- Brazil, Mexico, and the US.

Heart failure is a leading ill-defined cardiovascular cause of death in the US and in many other countries [11]. Coronary heart disease is the primary cause of heart failure in the US, but in developing countries, infections such as Chagas disease can play an important role [15]. Hypertension, diabetes, and overweight increase the risk of developing heart failure [16]. Determining heart failure etiology is often complicated by the presence of multiple co-morbid conditions [15].

Methods

In part 1 of the standard international death certificate, certifying physicians are asked to indicate the sequence of conditions leading directly to the death, listing the UCD last. Part 2 of the death certificate allows the certifier to list other contributing conditions. The underlying cause of death is then selected according to ICD-10 selection rules, typically using linkage tables (in Mexico prior to 2007 and in Brazil) or the automated coding system developed by the US National Center for Health Statistics in the US [17]. Heart failure is only selected as the UCD when no plausible underlying cause is listed in part 1 of the death certificate (ICD rules consider cancers plausible UCDs for this purpose), and neither ischemic heart disease (IHD) nor Chagas disease are listed on the death certificate. Therefore, we treated deaths certified to heart failure as records for which the UCD is missing and drew inferences about possible UCDs for these deaths by using deaths for which heart failure is listed in the causal chain leading to death (i.e., within part 1 of the death certificate). For public health purposes, it is not necessary to assign a unique UCD for each death certified to heart failure. Instead, each death can be distributed among several UCDs, as is the practice in the literature [10-13]. Distributing deaths certified to heart failure among several causes reflects uncertainty about the true UCD. After redistribution, new cause-specific death rates were calculated.

We used coarsened exact matching to generate a distribution of likely causes of death for each death certified to heart failure. Coarsened exact matching is a powerful algorithm but simple to use: the variables on which the match is made are first coarsened (divided into discrete categories), and then all exact matches are made (Figure 1). Thus, each death record certified to heart failure (treatment observations) was matched to all death records where heart failure appeared in part 1 of the death certificate and also had the same value for sex, age, and the other variables listed in Table 2 (control observations). Because the algorithm was not sensitive to how the variables in Table 2 were coarsened, we coarsened the variables until most treatment deaths could be matched to at least one control death. In order to avoid reassigning deaths certified to heart failure to other modes of death or ill-defined causes, we eliminated deaths with these UCDs from the potential control records [18]. In addition, we assumed that physicians would not miscertify injuries to heart failure and eliminated injury deaths from potential matches. Essentially, we matched incomplete death certificates to properly completed death certificates (their controls). Finally, we generated redistribution algorithms by assigning each heart failure death proportionally to the underlying causes of death of all controls. We tested the sensitivity of the method to varying match variables (Table 1), and show two alternate match algorithms:

> Demographic specification: match on age, sex, death location, region, and urban/rural, and restrict controls to non-Hispanic whites (US only) with the highest education level and health insurance (Mexico only);
>
> Autopsy specification: match on age, sex, death location, region, urban/rural, and restrict controls to those deaths that were autopsied (autopsy specification).

Because congestive heart failure (ICD-10 I50.0) and left ventricular heart failure (I50.1) may be used differently than unspecified heart failure (I50.9), we generated redistribution algorithms considering these causes separately,

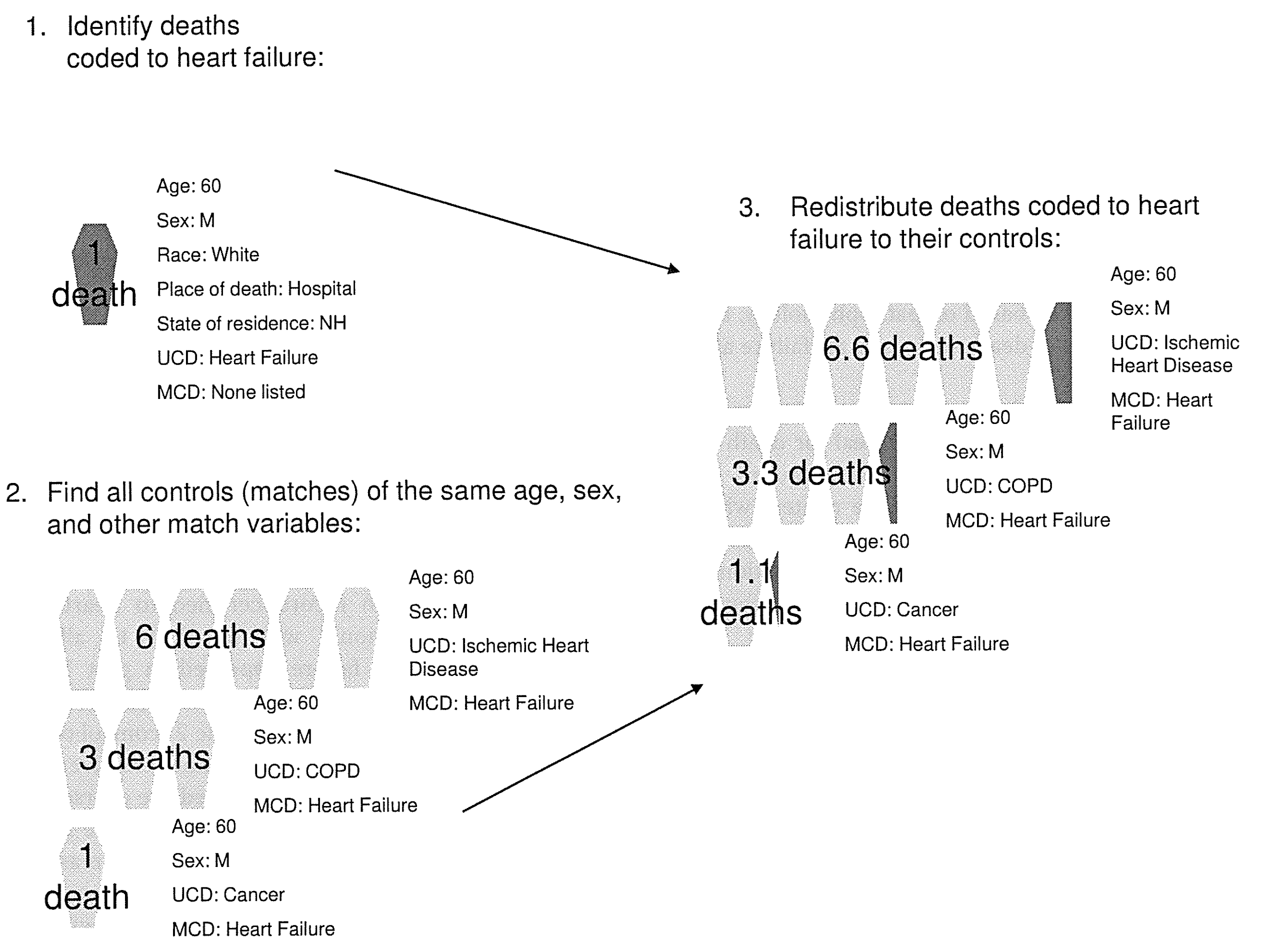

Figure 1 An example of coarsened exact matching. UCD: Underlying cause of death; MCD: multiple cause of death; COPD: chronic obstructive pulmonary disease. (1) All deaths with heart failure as the underlying cause of death are identified. In this example, a 60-year-old white male who died in a New Hampshire hospital is identified. (2) All deaths that match the treatment (heart failure) deaths are identified. In this case, all deaths of the same age and sex, with heart failure listed in the causal chain, are identified. (3) The treatment death identified in step 1 is redistributed to the UCDs of the control deaths identified in step 2, proportionally to the number of times each UCD appears among the control deaths. Thus, because 60% of the control deaths identified in step 2 have a UCD of IHD, 60% of the deaths in step 1 are assigned a UCD. The new total of IHD deaths among 60-year-old men is 6.6.

in addition to generating a redistribution algorithm for all deaths certified to heart failure.

In this paper, the method was applied to individual death records from three datasets: US vital registration death records for the years 1999-2004; Brazilian death records for the years 2003-2005, as provided to the Pan-American Health Organization; and Mexican vital registration data collected by the Health Ministry for 2004-2005 [19]. Mahapatra et al. classified both Mexico and the US as collecting high-quality cause-of-death data, with full coverage and less than 10% use of ill-defined codes [1]. Brazil was classified as collecting medium- to low-quality death statistics due to coverage of approximately 80%, with more than 15% of death records indicating ill-defined causes of death.

Validation

We also tested the performance of this method by dropping the underlying cause of death for specific groups of US death records that list heart failure among multiple causes of death. We then used matching to predict UCDs. Predicted underlying causes were compared to actual underlying causes using the average relative error (ARE), calculated as follows:

$$ARE = \frac{\sum_{j=1}^{n} \left| \frac{C\hat{S}D}{CSD} - 1 \right|}{n}$$

Table 2: Variables on which death records were matched, base specification.

Variable	US	Mexico	Brazil
Age	10-year intervals from age 20 to 49	10-year intervals from age 20 to 49	10-year intervals from age 20 to 49
	5-year intervals from 50-84	5-year intervals from 50-84	5-year intervals from 50-84
	Over 85	Over 85	Over 85
Sex	Male/Female	Male/Female	Male/Female
Death Location	In a clinic or hospital	In a clinic or hospital	In a clinic or hospital
	All other locations	All other locations	All other locations
Region	9 regions	5 regions	5 regions
Urban/rural	Urban/Rural	Urban/Rural	Urban/Rural
Education	Less than high school	Less than primary	None
	At least high school	At least primary	1-7 years
	4-year college or more	Secondary or more	More than 7 years
Race	White		
	Other		
	Black		
Hispanic	Hispanic		
	Non-Hispanic		
Occupation		Professional/technical	
		Informal economy	
Health Insurance System		IMSS	
		Other public or private	
		Seguro popular/none	

IMSS: Mexican Social Security Institute, which provides health insurance to formal-sector workers. Seguro popular: government-subsidized health insurance scheme for the uninsured.

where n is the number of causes considered, $C\hat{S}D$ is the predicted cause-specific number of deaths, and CSD is the actual cause-specific number of deaths [12]. Because few studies like this one have been carried out, there is little information by which an acceptable ARE can be set a priori. Therefore, it should be considered a descriptive indicator only.

We tested the method in the demographic groups for which heart failure and other ill-defined causes are most frequently used (i.e., where cause-of-death assignment is poor): 1) a region (the Southeastern US, which consists of Alabama, Kentucky, Mississippi, and Kentucky); 2) a racial/ethnic group (all blacks and Hispanics); 3) all deaths on which an autopsy was not performed; and 4) all deaths that occurred out-of-hospital.

The authors had full access to the data and take responsibility for their integrity. All authors have read and agree to the manuscript as written.

Results

Heart failure (ICD I50) is listed as the underlying cause of death in 2.3% of US death records, 1.9% of Mexican records, and 2.7% of Brazilian records. Of those deaths, 32%, 13%, and 33%, respectively, did not contain any other information on the MCDs; when other causes were listed, they were primarily other ill-defined causes (i.e., cardiac arrest or respiratory failure).

Prior to matching heart failure deaths to other deaths, records with ill-defined or incomplete cause-of-death information -- 7% of total matches -- were eliminated. Ill-defined deaths certified to renal failure (N17-N19), essential hypertension (I10), and general/unspecified atherosclerosis (I70.9) occurred frequently among the potential matches that were eliminated (35%, 5.9%, and 15.5% respectively). A sensitivity analysis was performed, where potential matches were restricted to records with at least three causes listed (i.e., more detailed cause-of-death information was provided), but this had little effect on the results.

In the base analysis, US heart failure deaths were matched to an average of 2,888 death records with mention of heart failure but with another disease as the UCD; Mexican deaths, to 251 records; and Brazilian deaths, to 985 records. Overall, 0.1% of heart failure deaths were not matched to any non-heart-failure death. The aggregate percentage of heart failure deaths redistributed to each underlying cause is shown in Table 3. In all three countries, the largest proportion of heart failure deaths were redistributed to IHD (53%, 26%, and 22% respectively in the US, Mexico, and Brazil). However, a larger proportion of deaths were redistributed to chronic obstructive pulmonary disease, diabetes, and hypertensive heart disease in Mexico and Brazil than in the US. The largest proportion of deaths assigned to cardiomyopathy was in Brazil (9% vs. 4% in the US and 1% in Mexico). In Brazil, an additional 3.7% of heart failure deaths were reassigned to Chagas disease (ICD-10 B57). Because few deaths are certified to Chagas disease, this resulted in a 20% increase in the number of deaths certified to Chagas disease.

Redistribution algorithms were generated by sex, age, and other demographic characteristics. For the US, redistribution algorithms are quite similar across demographic characteristics (See additional file 1: Table S1), with some exceptions for race and ethnicity. Heart failure deaths among blacks were 50% more likely to be redistributed to diabetes; among Hispanics, they were nearly twice as likely. A larger proportion of deaths among blacks was also redistributed to hypertensive heart disease and cardiomyopathy (14% vs. 6% among whites for hypertensive heart disease, and 8% vs. 4% for cardiomyopathy). In Mexico, there was a clear socioeconomic gradient in the proportion of heart failure deaths redistributed to hypertensive heart disease: the proportion was largest among women, deaths occurring outside of a hospital, those with less than primary school completed, and those without health coverage through their employer (Additional file 1: Table S2). A similar pattern was apparent in Brazil (Additional file 1: Table S3).

After redistributing heart failure deaths, IHD death rates among US adults over age 30 increased from 3.95 per 1,000 to 4.19 per 1,000; hypertensive heart disease rates increased from 0.38 to 0.41 per 1,000; and cardiomyopathy death rates increased from 0.20 per 1,000 to 0.22 per 1,000. Both absolute and proportional increases in death rates were greater for older age groups, when deaths are more likely to be assigned to heart failure (for example, adjusted IHD death rates for adults over age 85 were 9.5% higher than unadjusted rates vs. 2.3% for adults age 60-64). Changes in death rates varied little between 1999 and 2004 (Figure 2).

Several different specifications of the matching algorithm were tested to determine the effect on the resulting redistribution algorithm (Table 3). For the US data, varying the matching algorithm did not have a major effect on the results. When matched only to deaths for which an autopsy was performed, the percentage of deaths redistributed to digestive diseases and cardiomyopathy increased, and those to diabetes and stroke decreased. However, in these cases, the autopsy results are often not incorporated into the death records, and it is unclear what role selection bias (in terms of the characteristics of deaths that are autopsied) plays.

Results were more sensitive to the specification of the matching algorithm for Mexico and Brazil. In Mexico, when matching to autopsied deaths, the proportion of deaths redistributed to IHD increases (33% vs. 26% in the base specification), and the proportion redistributed to cancers decreases (4% vs. 6%). Likewise, for Brazil, matching only to autopsied deaths results in a substantial increase in the percentage of deaths redistributed to IHD (41% vs. 22%); however, unlike in Mexico, it also doubles the percentage of deaths redistributed to cardiomyopathy (17% vs. 9%). This may reflect more variable epidemiology or quality in cause-of-death assignment in Mexico and Brazil, where patterns in causes of death recorded vary more across population subgroups than in the US.

Congestive heart failure and left ventricular heart failure (ICD-10 I50.0 and I50.1) are used far more frequently than unspecified heart failure (I50.9) in the US, with the pattern reversed in Brazil and Mexico (Additional file 1: Table S4). However, specified heart failures (congestive and left ventricular) were associated with the same underlying causes as unspecified heart failure, and the redistribution algorithm varied little by heart failure type in all three countries.

When tested by dropping underlying cause-of-death information for specific population subgroups in the US, the method performed well (ARE of 19% when underlying causes of death in Southeastern states were predicted; ARE of 22% when causes for all non-white deaths were predicted). However, when the method was used to predict the cause-of-death distribution for all out-of-hospital deaths and for all nonautopsied deaths, it performed less well (ARE of 31% and 35%, respectively).

Discussion

In this paper, we proposed using coarsened exact matching to predict the likely UCD when heart failure was assigned as the UCD on death certificates. This method requires individual death certificates with multiple cause-of-death data. This method assumes that for all causes of death that a certifying physician lists as heart failure, he or she is equally likely to omit the underlying cause of death from the death certificate (regardless of whether the underlying cause is known). We performed a preliminary validation of the method. The validation indicated that even if the underlying cause of death is more likely to be omitted for certain demographic groups, the method would work well.

Using a nonparametric method such as matching to correct cause-of-death data has a number of advantages over multinomial logistic regression, which has been used

Table 3: Redistribution algorithm derived under alternate matching algorithms.

	USA			Mexico			Brazil		
	Base (%)	Demographic (%)	Autopsy (%)	Base (%)	Demographic (%)	Autopsy (%)	Base (%)	Demographic (%)	Autopsy (%)
Lower respiratory infections	1	1	2	3	2	2	3	3	1
Diabetes	6	5	2	13	15	12	9	9	3
Cancers	4	4	4	6	8	4	3	3	2
Ischemic heart disease	53	54	54	26	28	33	22	24	41
Cerebrovascular disease	2	2	0	3	3	2	4	4	1
Hypertensive heart disease	7	6	9	16	16	14	23	22	19
Cardiomyopathy	4	4	6	1	1	1	9	9	17
Other cardiovascular diseases	10	11	12	8	8	11	5	6	6
Chronic obstructive pulmonary disease (COPD)	5	5	3	11	8	8	9	8	3
Digestive diseases	1	1	3	4	3	4	2	2	2
Other diseases	7	7	6	9	8	10	11	9	6

Match variables are shown in Table 2. In addition, potential matches were restricted as follows: Demographic: matches (controls) were selected from demographic groups that have the best access to health care (US: non-Hispanic white college graduates; Mexico: secondary school graduates covered by a formal health insurance system; Brazil: individuals with at least seven years of schooling). Autopsy: matches (controls) were selected only from deaths that were autopsied.

elsewhere [12,13]. First, this method is fast compared to multinomial regression. Second, it does not impose assumptions about the functional form and therefore, unlike regression, is unaffected if those assumptions are wrong. Matching is equivalent to a fully saturated multinomial model, including all pairwise and higher order interactions, but without assuming that treatment effects are constant. Using a matching algorithm results in an algorithm that is insensitive to analysts' choices about whether to include interactions and higher order terms [20]. Third, we do not assume parameter constancy (that all of the predictor variables mean the same thing for all observations). This assumption may not hold if the variation in the parameters is related to the relatively small number of available covariates. If this is the case, the results would very likely be biased. Fourth, logistic regression can be biased if its crucial "independence of irrelevant alternatives" assumption is violated; coarsened exact matching is not biased whether or not this assumption holds. An implication of this is that the outcome categories need not be broad and distinct when coarsened exact matching is used. Finally, an important related advantage of matching is that it does not require the analyst to select the underlying causes of death to which ill-defined deaths are reassigned. In fact, it identifies the causes of death with which specific ill-defined causes of death are associated. For example, it is implausible that heart failure is in the causal chain for cancers, yet certifying physicians frequently list heart failure and cancers together on the death certificate. This method identifies that association and redistributes heart failure deaths accordingly. A multinomial regression using the match variables in the base case and the outcome categories identified using matching yields quite similar results to the matching algorithm -- but arriving at the model using multinomial regression alone would have required more stringent assumptions, as well as fitting a larger number of models, and therefore more time and computational resources.

The method described here could be applied to other intermediate cause-of-death codes that are frequently recorded on death certificates, such as septicemia (ICD-10 A40-A41). It could also be applied to underlying causes of death that are used inconsistently for different demographic groups, such as diabetes [13], or liver cirrhosis and liver cancer. Death records for a demographic group for whom certification is expected to be of poor quality can be matched to records for a reference demographic group for whom certification is of high quality.

This method has several limitations. First, the validation presented did not test the assumption that the probability of omitting the underlying cause of death is equal across causes for which heart failure is listed. To validate that assumption, a review of medical records and/or autopsies of a random sample of deaths certified to heart failure, and a tally of the revised underlying causes of death, would be needed. Second, death records can only be matched on recorded covariates. The results could be improved by measuring and including additional covariates (such as additional indicators of socioeconomic status or additional signs and symptoms not recorded on the death certificate) and assessing the results. Finally, the redistribution algorithm may not be transferable to other countries. Even if the assumptions of how heart failure is

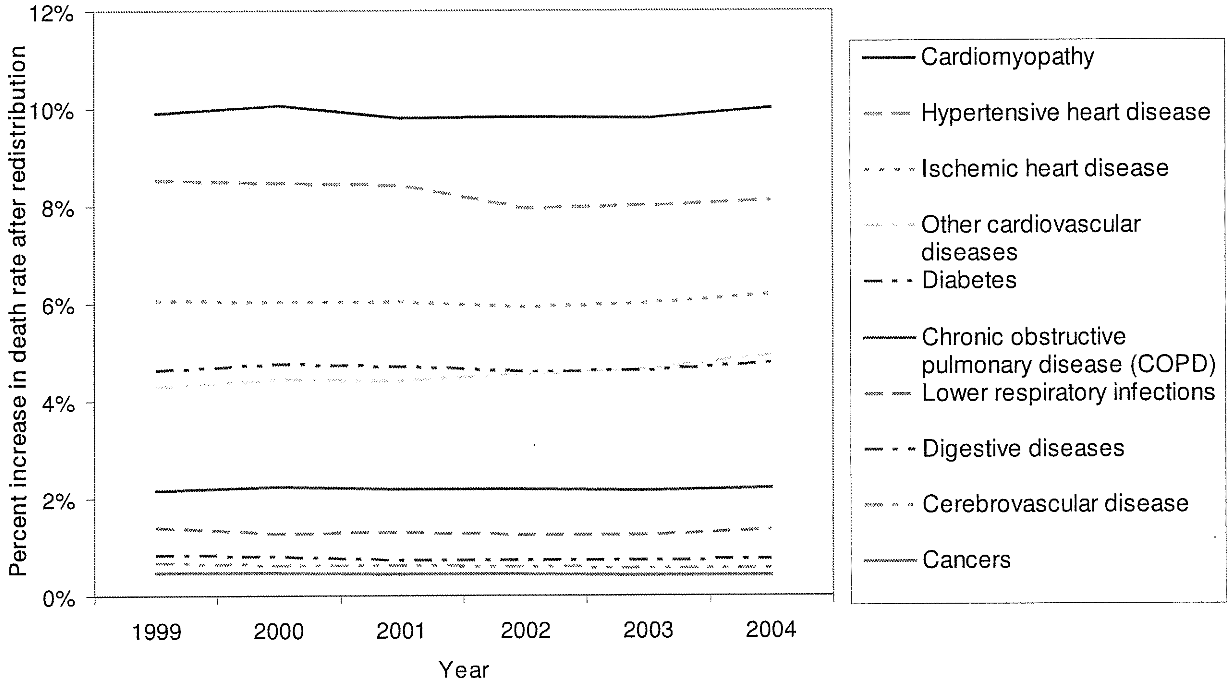

Figure 2 Increase in US adult death rates after redistribution of heart failure deaths. Rates are for adults over age 30 and are age-standardized using the US age distribution in the year 2000.

used hold true within each of the three countries analyzed, physician culture surrounding the use of heart failure is likely to vary from country to country. For example, 18% of recorded deaths were certified to heart failure in Egypt in 2007 [21]; we suspect that physicians commonly use heart failure when the cause of death is unknown. The corresponding correct underlying causes of these deaths likely represent a broader range of underlying causes than in the US, Mexico, or Brazil.

A challenge when interpreting cause-of-death statistics is distinguishing between true epidemiological differences across demographic groups and variations in quality of cause-of-death assignment. For example, there is a clear association both across and within the countries studied between use of hypertensive heart disease as an underlying cause of death (and therefore, redistribution of heart failure deaths to hypertensive heart disease) and indicators of low socioeconomic status. Individuals with low socioeconomic status are less likely to have diagnosed and controlled their hypertension, and therefore would be more likely to die from hypertensive heart disease. However, it is also plausible to argue that hypertensive heart disease is overused as an underlying cause of death, and that overuse is higher among groups with inferior access to health care. Likewise, the high proportion of heart failure deaths reassigned to diabetes in Mexico may represent a true epidemiological difference, or merely physician practice surrounding the certification of deaths to diabetes. As with a complete validation of this method, resolving such doubts would require review of medical records and/or autopsies of a random sample of hypertensive heart disease deaths.

ICD rules for designating the underlying cause of death represent a categorical system of classification; that is, each death is assigned one and only one cause. Categorical classification has the advantage that deaths from each disease sum to the total number of deaths [22]. However, in some cases, including with many heart failure deaths, several diseases contribute to a given death, and the death may have been delayed by removing any one of the disease factors. This can make categorical attribution of the death to a single cause somewhat arbitrary [22]. Relatedly, policymakers may be interested in the entire chain of risks and diseases that lead to a given death so that they can estimate the effect of intervening early in the causal chain (i.e., promoting physical activity to reduce hypertension) or at a later stage (i.e., improving management of patients with heart failure). Nevertheless, there is currently no consensus on an alternate (counterfactual) method for classifying deaths. An important first step is to collect multiple cause-of-death information, and make these data available for analysis, as done by the US. This allows researchers to assign deaths according to their specific research goals. We encourage other national statistical offices to collect and disseminate multiple cause-of-death data to allow for this type of research.

Conclusions

Reassigning ill-defined deaths to plausible underlying causes of death reduces bias in cause-specific mortality rates and increases comparability of mortality statistics over time and across demographic groups. In this paper, we suggest that coarsened exact matching be used to identify causes of death to which deaths should be redistributed and to derive situation-specific redistribution algorithms. We performed a preliminary validation of the method, and suggest that it be validated with a review of medical records or autopsies of deaths certified to heart failure.

Disclaimer

The view presented in this paper does not necessarily represent the view of the World Health Organization.

Additional material

Additional file 1 Additional tables, which contains Tables S1-S4.

Abbreviations

ARE: average relative error; COPD: chronic obstructive pulmonary disease; IHD: ischemic heart disease; MCD: multiple cause of death; UCD: underlying cause of death

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

GK developed the methods described in this paper. GK and GS tailored the methods for this application. GS carried out the calculations and drafted the manuscript. KS conceived of and coordinated the study. All authors helped to write the manuscript, and have read and approved the final manuscript.

Acknowledgements

We would like to thank Rafael Lozano and Christopher Murray for useful discussions on the methods presented in this paper. We would also like to thank John Silvi for assistance obtaining data used in this analysis and Maya Mascarenhas for thoughtful comments on the figures. This research was partially supported by Japan's Ministry of Health, Labour and Welfare, and the Bill and Melinda Gates Foundation. The funding bodies played no role in study design, data collection, analysis or interpretation, writing, or the decision to submit the manuscript.

Author Details

[1]Information, Evidence and Research, World Health Organization, 20 Avenue Appia, 1211 Geneva, Switzerland, [2]Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA and [3]Department of Global Health Policy, University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan

Received: 11 November 2009 Accepted: 13 April 2010
Published: 13 April 2010

References

1. Mahapatra P, Shibuya K, Lopez AD, Coullare F, Notzon FC, Rao C, Szreter S: **Civil registration systems and vital statistics: successes and missed opportunities.** *Lancet* 2007, **370(9599)**:1653-1663.
2. World Health Organization: *The global burden of disease - 2004 update* Geneva: World Health Organization; 2008.
3. World Health Organization: *International Statistical Classification of Disease and Related Health Problems* Tenth edition. Geneva: World Health Organization; 1992.
4. Hanzlick R: **Improving accuracy of death certificates.** *JAMA* 1993, **269**:2850.
5. Yamashita T, Ozawa H, Aono H, Hosokawa H, Saito I, Ikebe T: **Heart disease deaths on death certificates re-evaluated by clinical records in a Japanese city.** *Jpn Circ J* 1997, **61**:331-338.
6. Smith Sehdev AE, Hutchins GM: **Problems with proper completion and accuracy of the cause-of-death statement.** *Arch Intern Med* 2001, **161**:277-284.
7. Ravakhah K: **Death certificates are not reliable: revivification of the autopsy.** *South Med J* 2006, **99**:728-733.
8. Johansson LA, Bjorkenstam C, Westerling R: **Unexplained differences between hospital and mortality data indicated mistakes in death certification: an investigation of 1,094 deaths in Sweden during 1995.** *J Clin Epidemiol* 2009, **62**:1202-1209.
9. Roulson J, Benbow EW, Hasleton PS: **Discrepancies between clinical and autopsy diagnosis and the value of post mortem histology; a meta-analysis and review.** *Histopathology* 2005, **47**:551-559.
10. Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJ: **Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data.** *Lancet* 2006, **367**:1747-1757.
11. Lozano R, Murray CJ, Lopez AD, Satoh T: **Miscoding and misclassification of ischaemic heart disease mortality.** In *Global Program on Evidence for Health Policy Discussion Papers* Geneva: World Health Organization; 2001.
12. Murray CJ, Kulkarni SC, Ezzati M: **Understanding the coronary heart disease versus total cardiovascular mortality paradox: a method to enhance the comparability of cardiovascular death statistics in the United States.** *Circulation* 2006, **113**:2071-2081.
13. Murray C, Dias RH, Kulkarni SC, Lozano R, Stevens GA, Ezzati M: **Improving the comparability of diabetes mortality statistics in the United States and Mexico.** *Diabetes Care* 2007, **31**:451-458.
14. Iacus SM, King G, Porro G: **Matching for causal inference without balance checking: coursened exact matching.** 2008 [http://gking.harvard.edu/files/cem.pdf].
15. Krum H, Abraham WT: **Heart failure.** *Lancet* 2009, **373**:941-955.
16. He J, Ogden LG, Bazzano LA, Vupputuri S, Loria C, Whelton PK: **Risk factors for congestive heart failure in US men and women: NHANES I epidemiologic follow-up study.** *Arch Intern Med* 2001, **161**:996-1002.
17. Lu TH: **Using ACME (Automatic Classification of Medical Entry) software to monitor and improve the quality of cause of death statistics.** *J Epidemiol Community Health* 2003, **57**:470-471.
18. Naghavi M, Ross J, Lozano R: **Taking out the trash: rethinking garbage codes and redistribution methods.** Seattle: Institute for Health Metrics and Evaluation; 2008.
19. Secretaría de Salud de México: **Sistema Estadístico de Defunciones.** [http://www.sinais.salud.gob.mx/].
20. Ho D, Imai K, King G, Stuart E: **Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference.** *Political Analysis* 2007, **15**:199-236.
21. World Health Organization: **WHO Mortality Database.** 2009 edition. World Health Organization; 2009.
22. Mathers CD, Ezzati M, Lopez AD, Murray CJL, Rodgers A: **Causal Decomposition of Summary Measures of Population Health.** In *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications* Edited by: Murray CJL, Salomon J, Mathers CD, Lopez AD. Geneva: World Health Organization; 2002.

doi: 10.1186/1478-7954-8-6
Cite this article as: Stevens *et al.*, Deaths from heart failure: using coarsened exact matching to correct cause-of-death statistics *Population Health Metrics* 2010, **8**:6