

If your local weatherman would predict that tomorrow's temperature would be somewhere between 50 degrees below zero and 50 degrees above, you would call your tv station and demand to replace him since the interval is too big and the weatherman's advice to just take the average – 'It will be zero, I guess' – beyond any realistic reasoning.

With nothing more than this weatherman's estimates, two years ago, on October 29, 2004, the renowned British scientific journal *The Lancet* shocked the world with an article about the death rate after the March 18, 2003, Iraq invasion by American-led forces.

The journal published a survey that concluded that the Iraq war had caused between 8.000 and 194.000 deaths, excluding the province Falluja. The survey took the average of these two numbers and concluded that there probably had been 98.000 deadly victims (outside Falluja) during a period of 18 months after the US-led invasion. This is not how reliable statistical estimates usually work. A reliable estimate functions within a workable bandwidth, not within an extremely wide 'confidence interval' such as 8.000-194.000.

There were only a handful of journalists asking the right questions. Most media gleefully repeated *The Lancet*'s claims; it was exactly what most of them wanted to hear, so they kept quiet about the dubious methodology used by the authors of *The Lancet*'s article.

The researchers had divided Iraq in 18 'Governorates'. In these Governorates they randomly selected 33 clusters (areas of research), and in each cluster they randomly selected 30 households to be interviewed.

The number of clusters in a certain Governorate was proportionate to the population size of this Governorate, meaning: in Governorates with a larger population such as Baghdad they projected more clusters, but the clusters itself were more or less equal in size.

Every cluster represented 739.000 people, or 30 households, containing on average 238 individuals (7.9 per household). The survey did not make clear how many individuals were interviewed. The article suggested 1 or 2 per household.

A total of 988 households in 33 clusters were questioned by groups of researchers. The research teams registered 21 deaths in the period after the invasion. Of these victims, 9 were attributed to violence on the part of the Coalition forces.

How the research teams went from household to household was not made clear in the published survey. The survey stated that the teams randomly visited neighboring households after randomly selecting a starting household in a certain cluster.

In order to judge the essential condition that neighboring households (to the first household) with violent deaths had the same chance of being selected as neighboring households without violent deaths, the survey should have given the precise method – judged by independent observers – of how the teams proceeded in the field. There is a serious danger that households that had experienced personal losses led the teams to other

households with the same experience – this would distort the general statistical conclusion. Another potential danger is that the households which refused to be interviewed represented a biased sample, such as households with relatively few violent deaths. It is also important to know how the field researchers were selected and screened, particularly their political biases should have been a case of concern.

The small amount of clusters raises the question whether it was possible to extrapolate the findings of a relatively small amount of households to the Iraqi population in general.

It is easier to estimate how many houses have water damage in a specific area hit by bad weather; the area can be defined, the amount of rain per square feet, and a random sample can reveal with a reasonable level of reliability how many houses have water damage.

But in the case of Iraq, the violence of the invasion and of the post-invasion months was not equally spread over the 18 Governorates. After the fall of the Saddam Hussein regime, terrorist and guerrilla warfare flared up, but most of the violence was concentrated in specific areas and towns.

The unjustified belief of the researchers in their predicted death counts, in spite of the immensely wide confidence intervals, becomes apparent in the following sentence taken from the October, 2006, article discussed below: ‘The likelihood that another number is the *correct* [emphasis added] number decreases very rapidly as one moves up or down from the figure [98.000]’.

This is a wrong statement: the fact that the predicted death count is subject to so much uncertainty as reflected by the wide confidence interval means that it is unreliable. This unreliability should have been the main obstacle for publication in a scientific magazine.

This confidence the researchers put in their predicted death counts is even more remarkable in light of the fact that a much larger Iraq Living Conditions 2004 survey - sampling 22.000 households instead of less than 1000, reported a predicted death count of 24.000 with a 95 percent confidence interval from 18.000 till 29.000.

That is, a much more reliable study yields a predicted death count a factor 4 smaller than the predicted death count in The Lancet’s 2004 article.

The 2004 survey lacked the list with the exact numbers of the violent deaths per cluster. It was not possible to identify the clusters where violent deaths had been registered by the research teams.

Early this year we asked one of the writers of the article, Dr. Les Roberts: ‘In which clusters did you observe violent deaths?’

Dr. Roberts replied: ‘We cannot release that information due to the confidentiality constraints imposed by Johns Hopkins. With that information someone could go to a neighborhood or town and ask where those doctors interviewed a year ago and then figure out who said what to us.’

We tend to disagree with Dr. Roberts. Each cluster represents on average three quarter of a million of inhabitants, and it was impossible to identify the interviewees. Because this list of numbers per cluster was not part of the article, it was not possible to judge

whether the victims were spread over all the clusters or were significantly limited to specific clusters.

In a scientific journal, the only sound conclusion of a study reporting a confidence interval of 8.000-194.000 would be that the study has not been able to report anything at all of interest with an acceptable level of statistical significance.

The field survey of the 2004 article took place in September 2004. Already on October 29, 2004, The Lancet published its results.

Usually it takes many months of 'peer review' by neutral referees before a scientific magazine decides to publish - beyond the time needed to properly analyze the data by the research team itself. Not in this case. It must have been a coincidence that four days after The Lancet took the world's headlines with this survey, the American people would vote for its new president.

Again the American people go to the polls, and accidentally The Lancet publishes another article about the Iraq War. This second survey is called "Mortality after the 2003 invasion of Iraq: a cross-sectional cluster sample survey". On October 12 this article raged through the world media.

The design of this new survey involves again dividing up Iraq in 18 Governorates. In these 18 areas the researchers randomly projected 47 clusters. In every cluster it was aimed to question 40 households. Every cluster is supposed to represent 577.000 people based upon an estimated population size of 27.139.583 persons. In other words: every household represents on average 14.452 people. The survey does not mention the use of a standardized questionnaire.

Each of these 18 Governorates can be expected to have its own probability distribution of number of deaths in a randomly sampled cluster. In order to estimate the variability of an estimate of the number of deaths (by cause) within a particular Governorate one needs to sample several clusters in this Governorate – it speaks for itself that the error in the estimate of the number of violent deaths and in the estimate of the variability of this estimate is large if the number of sampled clusters in the Governorate is small, and the opposite speaks for itself too.

Unfortunately, just like in the first survey, in this second survey the number of clusters per Governorate is very small (0,1,2 or 3), except in Baghdad (12). Also the number of households per cluster varied between 160 and 320, which also adds to the widening of a confidence interval (the bandwidth in which a certain estimate is projected).

This means that the data do neither allow the construction of a reliable estimate of the count of deaths per Governorates for each of the 18 Governorates, nor the construction of a reliable estimate of the precision/standard error of this estimate.

If one would try to obtain a more or less reliable 'confidence interval' (the bandwidth in which a certain estimate is projected), a so-called 'nonparametric bootstrap method' would be an acceptable model in order to estimate the variability of the reported Governorates' specific estimates. A 'nonparametric bootstrap method' is a technical

standard procedure involving the random re-sampling of the results of the clusters in each Governorate and then recalculating the estimates. The accuracy of these estimates relies on the 47 clusters being a good representation of the total Iraqi population. The low sample size of 47 clusters and the large variability of violent death counts across Iraq limits its reliability.

A statistical analysis - calculated by us and under the assumption that the numbers in every cluster have been collected objectively and truly represent unbiased samples - with the cluster data of the new 2006 survey, at our request kindly provided to us by the authors of the survey - shows that the true confidence intervals will need to be much wider.

We found that a data driven confidence interval for the true number of post-invasion violent deaths, is not 426.369-793.663 (as given by the survey) but is given by the approximate range of 100.000-1.000.000.

That is, our statistical analysis could at most conclude that the total number of violent deaths is more than 100.000 with a 0.95 confidence - but this takes not into account various other potential biases in the original data.

Are there potential biases in the underlying data? We have to go back to these data: the collected numbers in the 47 clusters in Iraq. Are there potential biases in those numbers? To name a few concerns: 1) What proportion of the reported violent deaths were verified with a death certificate? 2) Why is the fact that overall 20% of the death reports had a missing death certificate (as the survey states) not incorporated in the statistical analysis to quantify this extra uncertainty in the reported predicted death counts - or are we absolutely sure that each non-verified violent death is correctly labeled? 3) Why does the article not make it very explicit that the teams of interviewers were in no way influenced by their visits to previous households in selecting the next household among the cluster of 40 households?

For example, a violation of the latter assumption would have dramatic consequences for the validity of the extrapolated death counts and intervals. To make the point, if a biased 40 households cluster corresponds in essence with an unbiased 80 household cluster, then the confidence interval above will be shrunk by a factor of 2 which reduces the statistically significant statement to 'there are more than 50,000 violent deaths' - that is, if this assumption (the randomly sampling of 40 households) is violated one needs to conclude that the data cannot be interpreted on a sane scientific basis.

Another element of the design of this study exposes its structural weakness. The authors state: 'One team could typically complete a cluster of 40 households in 1 day.'

Imagine that the field researchers took a 12 hour workday, starting at 8 AM and working until 8 PM. They did not break for lunch or siesta. They did not suffer from the heat. They moved from one household to the other within the context of tribal communities brimming with distrust, explained their mission and succeeded in gaining access to the living quarters, got into a person's confidence, asked for intimate experiences, listened to personal stories of loss and grief - and all this within 18 minutes per household.

Researchers of Oxford University and the London University claimed ‘The study suffers from main street bias by only visiting houses close to side streets of the main streets’’. Many Iraqi families do not satisfy this condition and had zero probability to be sampled. “Main street bias” increases estimates of the number of violent deaths since conflicts such as car bombs, drive-by shootings, explosions on market places more likely to occur in the type of area visited by the researchers.

Regarding the design of the study, it is also remarkable that a priori knowledge about the places in which violence has been prevalent is not used in the design and interpretation of the analysis: a better design would have been to sample frequently in areas in which violence has been prevalent and sample less in areas in which it has been relatively calm.

Instead the authors advertise the selected design of the study as the accepted standard, which would then wrongly imply that it makes sense to sample as many clusters in areas in which violence is non-existent as in violent areas, as long as these areas have the same population size. The truth seems to be that the authors’ methodology does not use crucial and available information, but does use wrong assumptions.

It is easy to see how a priori knowledge could have been used to improve the sampling strategy and thereby obtain more reliable estimates. In this survey, the extrapolation of the predicted violent death count of 600.200 for the total Iraq population would correspond with a violent death count of around 100.000 in the Kurdish Autonomous Regions, under the ‘equal spread of violence assumption’ for which their design would have been rational.

However, the factual death toll of the car bombings there is between 250 and 300 – the Kurdish regions are peaceful and under firm control. This shows the immense violation of the ‘equal spread of violence assumption’ and thereby the equally large degree of irrationality of the design.

One also needs to wonder why one did not focus on counting actual death certificates, given that the survey reported that 80% of the counted deaths were verified with a death certificate. If we accept the predicted number of 650.000 ‘excess deaths’ in Iraq in the post-invasion period, as reported by The Lancet’s survey, then we should also accept that more than 500.000 additional death certificates have been issued by government organizations; in other words, these government organizations are somehow hiding all of these certificates from the public.

In the 2006 survey, the post-invasion period has been cut in three parts.

The first one starts at March, 2003, and ends at April, 2004. Here 45 violent deaths are reported.

The second part ends at May, 2005. Here 90 deaths are reported.

The third part ends at June, 2006. Here, most violent deaths are reported: 165. Total violent deaths: 300.

Let us focus on that third period with 165 violent deaths in the 47 clusters within the timeframe June 2005-June 2006. The study estimates that in this 12 month period 330.000 people have died because of violence. That is a number of 27.500 per month. 6875 per week. About 982 persons per day – every day.

The study projects that about 40.000 people died because of air strikes in the period June 2005-June 2006. Car bombs killed 60.000 people in that same period. ‘Other explosions’ ended the lives of 40.000 people more. ‘Gunshots’ killed 174.000 persons.

These are terribly huge numbers – without the media noticing this? Hundreds of independent radio stations, tv-stations, newspapers and magazines operate in Iraq nowadays, and nobody noticed these massacres? According to this survey, American air strikes must have erased whole neighborhoods: where, how, when?

Here as well, a priori knowledge would have prevented the authors of making ridiculous extrapolations. According to The Lancet, the representatives of the media, most of whom are criticizing the Coalition’s presence, missed 330.000 victims in the last year. It is not a surprise that the anti-war group Iraq Body Count vehemently condemned this survey.

Given all of this, one needs to wonder if this large estimated number of violent deaths is not only due to statistical uncertainty (100.000-1000.000), but possibly also due to one or more of the potential biases mentioned above (and biases not mentioned at all because of a lack of space). Could it be that The Lancet’s survey is juggling with statistics and defies common sense?

We conclude that it is virtually impossible to judge the value of the original data collected in the 47 clusters. We also conclude that the estimates based upon these data are extremely unreliable and cannot stand a decent scientific evaluation.

It may be that the number of violent deaths is much higher than previously reported, but this specific report, just like the October 2004 report, cannot support the estimates that have been flying around the world on October 29, 2006. It is not science. It is propaganda.

Mark van der Laan is the Jiann-Ping Hsu/Karl E. Peace Professor in Biostatistics and Statistics at UC Berkeley. In 2005, Dr. van der Laan received the Presidential Award of the Committee of Presidents of Statistical Societies (COPSS).

Leon de Winter is a novelist and weekly columnist for Elsevier Magazine. His commentaries and essays appear frequently in Dutch and German publications. In 2002, he was awarded the Die Welt Literature Prize in Germany. He is an adjunct-fellow at the Hudson Institute in Washington DC.