Overview of Data Associated With Lancet Surveys of Mortality in Iraq

David Kane*

May 29, 2007

Introduction

The *Lancet* published two controversial articles about mortality in Iraq: Roberts et al. (2004) and Burnham et al. (2006). This article and the accompanying **R** software package have three purposes. First, I provide a data frame of the publicly available data distributed from Roberts et al. (2004).[1] Second, since the underlying data from Burnham et al. (2006) has been made available to some researchers but not others, I provide a summary of the restricted data. Third, for those with access, I provide some basic functions for working with the data.

Data from Roberts et al. (2004)

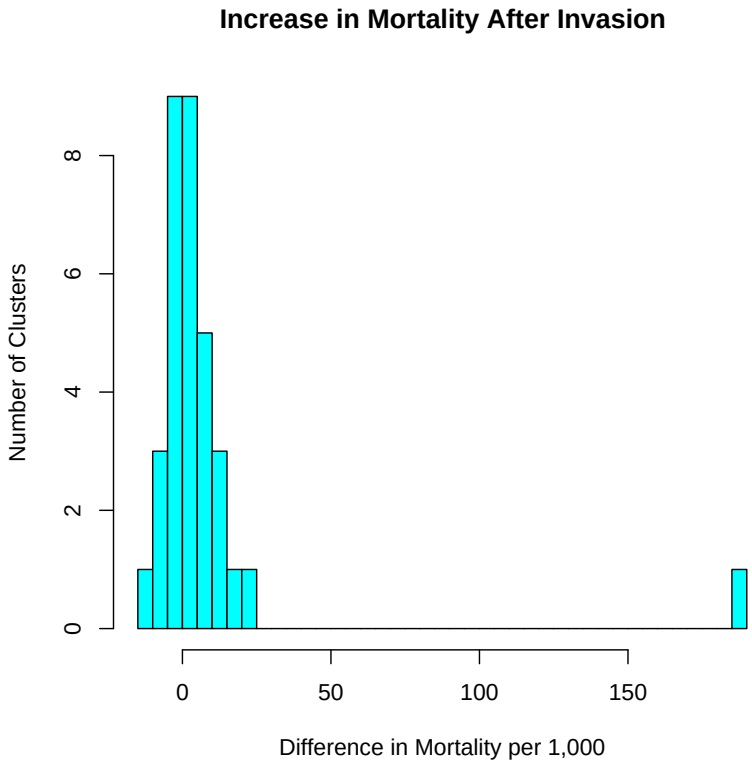
Typing `data(lancet1)` after installing the package loads the data frame. **lancet1** contains 20 variables and 33 rows, one for each of the clusters. The variables split broadly into two categories: pre-invasion and post-invasion.

*Institute Fellow, IQSS, Harvard University, Cambridge, MA 02138. dkane@iq.harvard.edu. Thanks to Arjun Ravi Narayan for excellent research assistance.

[1]I thank Tim Lambert of Deltoid for posting the summary data associated with Roberts et al. (2004). I thank the authors of Burnham et al. (2006) for making their data available to me. I urge them to make the data available to all. I also thank Shannon Doocy for patiently answering my questions.

Statistics for each cluster include the number of births, deaths (infant and total), and persons (in different age groups) alive.

The person-months in each period is provided, although the exact calculation is unavailable to us, as the dataset does not include the timings of the births and deaths in each cluster. The mortality rate pre-invasion is calculated as $\frac{pre.deaths*12000}{pre.person.months}$. This same calculation is done for post-invasion mortality. The difference is used to calculate diff.mort.rate.



Summary statistics for diff.mort.rate are:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -14.5    -2.4     2.4     8.4     8.0   187.0
```

The Falluja cluster is the outlier, with an increase in mortality of 187 per 1,000. Removing this cluster yields:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -14.5    -2.5     2.4     2.8     7.8    23.2
```

Since all the (public) data is available in the package, I do not provide further analysis here.

Data from Burnham et al. (2006)

I have performed a series of manipulations to the raw data as distributed by the authors. See the functions `prep.deaths()` and `prep.houses()` for details. Each function produces a data frame. Call them **deaths** and **houses**. With 629 rows and 14 variables, **deaths** includes a row for every death recorded in the study. The variables are:

id A number for each household interviewed. Multiple deaths from the same household share the same id. There is no unique id in the **deaths** data frame.

governorate The governorate in which the household of the deceased is located. There are 18 governorates in Iraq, but miscommunication led to only 16 being sampled by the survey team.

cluster The cluster in which the household of the deceased was located. There are 47 clusters. Clusters are numbered from 1 to 51. Numbers 17, 19, 29 and 50 are missing. The authors sampled 50 clusters, but three were discarded.

date The death month. Although the actual date of death was recorded by the interviewers, the authors transform all dates to the first day of the respective months to protect participant identity. There are 57 missing dates.

year The year of death. There are no missing values. The number of deaths by year are:

```
2002 2003 2004 2005 2006
  64   80  141  194  150
```

The survey ended in July 2006, so mortality has increased every year.

sex There were 485 male deaths and 144 female deaths.

age The age of the deceased person. Summary statistics are:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      0      24      40      42      61     100      19
```

age.group The age group of the deceased. Following Table 2 in the article, categories are child (0-14 years), adult (15-59 years), and elderly (60+).

```
        age.group
sex      child adult elderly
  male      49   292     127
  female    31    60      51
```

This ignores the 19 deaths with NA age.

post.invasion A binary variable indicating whether the death occurred before or after the invasion in March 2003. This is provided even when there is no date of death. There are 547 deaths post-invasion and 82 deaths pre-invasion.

impute A TRUE/FALSE flag indicating whether an imputation was done to fill in the year and post.invasion variables even if the date of death was missing. There are 18 TRUE values for impute. In each of these cases, date is NA. There are a total of 57 NA dates. For the 39 observations with NA date and impute equal to FALSE, the interviewers were able to establish the year and whether or not the death happened after the invasion, so no imputation was necessary.

cause.summary A brief summary that describes the cause of death using one of 111 different categories. The five most common are:

```
gunshot from unknown                  MI
                  91                  61
                 CVA   bullet by USA army
                  45                  44
    exploded vehicle
                  38
```

death.nature A binary variable indicating whether the death was due to violent or non-violent causes. There are 302 violent deaths and 327 non-violent deaths.

	child	adult	elderly
violent	26	251	12
non-violent	54	101	166

This ignores the 19 deaths with NA age.

cause.category The authors aggregate cause.summary into 13 higher level categories.

gunshot	carbomb
169	38
other explosion	air strike
43	40
violent, unknown	old age
6	27
accident	cancer
36	48
heart disease or stroke	chronic illness
122	43
infection disease	infant death
4	40
non-violent, other	
13	

has.certificate A binary variable indicating whether a death certificate was presented to the interviewers or not. There are 83 NA entries. Burnham (2007) reports that the NA cases arose when the interviewer "forgot" to ask for a certificate.

The second data frame, **houses**, has 1849 rows and 18 variables, one of which (id) can be used for linking to the **deaths** data frame. The period covers January 1, 2002 through the date of the survey, which ranges from May 20, 2006 through July 10, 2006.

id, governorate, cluster The same as in the **deaths** data frame.

size The number of household members at the time of survey.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      1       5       6       7       9      26      13
```

males, females The number of males/females in the household. There are 55 NA values.

births The number of births in the household.

```
  0   1   2   3   4   5   7
924 531 284  78  23   8   1
```

deaths The number of deaths in the household.

```
   0    1    2    3    4    7
1323  440   74   10    1    1
```

immigration The number of immigrants to the household. There are 1720 houses with zero immigrants. The summary statistics of the remaining households are:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     3.0     5.0     5.9     7.0    21.0
```

emigration The number of emigrants from the household. There are 1697 houses with zero emigrants. The summary statistics of the remaining households are:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     2.0     4.0     4.4     6.0    13.0
```

size.2002 The number of household members on January 1, 2002. This is the number of members at the time of the survey (size) minus births plus deaths. For 6 observations, this does not add up. These mistakes occurred in 3 clusters (34, 36 and 49). As with size, there are 13 NA values, and the NA values for size.2002 are for the same rows as the NA values for size.

size.2002.mig The number of members of the household as of January 1, 2002, including migration: size minus births plus deaths minus immigration plus emigration. There are no further errors in the migration calculations other than those also occurring in size.2002. However, there are 58 more NA values in size.2002.mig than in size.2002. But, there are no NA values in immigration or emigration, so it is not clear where these additional NAs come from.

mid.2002 The average of size and size.2002. There are 13 NA observations, consistent with the NAs in size.2002.

mid.2002.mig The average of size and size.2002.mig. There are 71 NA observations, consistent with the NAs in size.2002.mig.

has.death A binary variable indicating if there were any deaths in the household. There are 526 households with at least one death.

deaths.violent The number of violent deaths.

```
   0    1    2    3
1562  269   14    4
```

deaths.nonviolent The number of non-violent deaths.

```
   0    1    2    3    7
1583  221   40    4    1
```

Note that the number of violent/non-violent deaths is inconsistent between the **deaths** and **houses** data frames. In **deaths**, there are 302 violent deaths and 327 non-violent deaths, just as Burnham et al. (2006) reports. In the **houses** data frame, there are 309 violent deaths and 320 non-violent ones. It is not clear what the reason is for this discrepancy.

Comments

This section provides a guided tour of some of the more interesting features of the data frames.

Clusters

The clusters in the data set are labeled from 1 to 51, with missing numbers at 17, 19, 29 and 50. In general, cluster numbers are grouped within governorates. For example, the three Thi-Qar clusters are 11, 12 and 13. The five in Ninewa are 34 – 38. The two in Saleh Al-Din are 47 and 48. A prominent exception to this pattern is Baghdad, which includes clusters 14 – 24 (with 17 and 19 missing), 33, 39 and 40.

Burnham et al. (2006) state:

> Only 47 of the sought 50 clusters were included in this analysis. On two occasions, miscommunication resulted in clusters not being visited in Muthanna and Dahuk, and instead being included in other Governorates. In Wassit, insecurity caused the team to choose the next nearest population area, in accordance with the study protocol. Later it was discovered that this second site was actually across the boundary in Baghdad Governorate. These three misattributed clusters were therefore excluded, leaving a final sample of 1849 households in 47 randomly selected clusters.

It is not clear why there is a cluster numbered 51 (located in Anbar) if the original plan called for only 50 clusters. The other two Anbar clusters are 30 and 31.

Burnham (2007) mentions that they sampled three clusters in Falluja (which is in the Anbar province) even though the plan called for only one cluster. They did this because the Falluja data from Roberts et al. (2004) was such an outlier that they wanted a better estimate for this violent city. Having interviewed in three clusters, the authors then selected one of the three randomly. The selected cluster was the least violent of the three. It seems plausible that this "extra" cluster is 51.

Interview Procedure

The paper states that the interviewers went to houses in each cluster until they completed 40 interviews.

> Empty houses or those that refused to participate were passed over until 40 households had been interviewed in all locations.

However, the data shows that only 29 of 47 clusters featured exactly 40 interviews. The following table shows the number of clusters for each total number of houses interviewed:

33	36	38	39	40	41
1	2	5	8	29	2

The clusters with 41 households were 23 (Baghdad) and 37 (Ninewa). Those with fewer than 38 were 3 (Najaf), 9 (Babylon) and 26 (Sulaymaniyah). This variation is troubling. If clusters in violent regions had more interviews than those in less violent regions, the calculated mortality rate might be too high. This problem may have occurred. For example, Sulaymaniyah (in the Kurdish north) should have featured 120 interviews, 40 for each of the three clusters. Instead, only 113 households were interviewed. Since Sulaymaniyah featured *no* violent deaths, interviewing fewer households there inflates the estimate of post-invasion violent mortality in Iraq.[2]

Yet the claim that interviews were conducted until 40 households agreed to participate is contradicted elsewhere in the article.

> In 16 (0.9%) dwellings, residents were absent; 15 (0.8%) households refused to participate. In the few apartment houses visited, the team progressed to the nearest households within the building. One team could typically complete a cluster of 40 households in 1 day.

If 40 households had been interviewed in each of the 47 retained clusters, there would be 1,880 rows in **houses**. The claim that 31 households were either absent or refused to participate is consistent with the 1849 observations actually present. It is unclear how one can reconcile this description (absences/refusals were not replaced) with the previous claim that "40 households had been interviewed in all locations."

[2]This would depend on the precise method used to calculate excess mortality. If the calculation were based on estimating mortality within each cluster and then aggregating these cluster estimates, it would not bias the results if less violent clusters had more interviews.

Death Certificates

The authors claim that missing death certificates, whether caused by the failure of the interviewer to ask or by the inability of the interviewee to produce one, are not a problem because there is no correlation between this missingness and other variables.

> Survey teams asked for death certificates in 545 (87%) reported deaths and these were present in 501 cases. The pattern of deaths in households without death certificates was no different from those with certificates.

There were 501 deaths with death certificates, 45 without and 83 observations with NA values. The two locations with the largest number of NA values were both in Baghdad: cluster 33 (24 deaths) and cluster 24 (10 deaths).

There were 35 deaths for which a certificate was provided, but no date of death is listed. This is interesting as one would expect that death certificates provide the date of death. There were 14 deaths for which the death certificate was not asked for and the date of death is NA. It would seem especially important to ask for a death certificate when the interviewee can not recall the date of death.

The asking rate for death certificates is correlated with the year of the death — the later the year, the higher the likelihood of the interviewers not asking for a death certificate. It is unclear why interviewers would be more likely to "forget" to ask for a certificate if the death occurred in 2006 rather than 2002.

	Year				
Asked For Death Certificate	2002	2003	2004	2005	2006
FALSE	2	9	13	23	36
TRUE	62	71	128	171	114

Of the 64 deaths in 2002, the interviewers forgot to ask for a death certificate 3% of the time. This percentage increases steadily until it hits 24% in 2006. There is correlation with the nature of the death as well.

	Nature of Death	
Asked For Death Certificate	violent	non-violent
FALSE	70	13
TRUE	232	314

It is unclear why the nature of a death would make an interviewer more likely to forget to ask for a death certificate. In the case of non-violent deaths, the interviewers forget to ask 4% of the time. For violent deaths, the forget-to-ask rate is 23%.

Cluster 33

Cluster 33 has a total of 25 deaths. All 24 of the deaths with NA death certificates had the same cause ("exploded vehicle") and occurred in the same month (July 2006). Since data collection ended on July 10th, these deaths must have happened prior to that date. It is not clear why death certificates were not asked for. Since the survey started on May 20th, the interviewers, by this point, had 6 weeks to learn to remember to ask for death certificates. It is also not clear how they remembered to ask for a death certificate in the one other death in this cluster, a non-violent heart attack.

Besides the issue of missing death certificates, there are two other problems with this cluster. First, the authors report that "deaths from car explosions have increased since late 2005." This is true but misleading. Consider total mortality from car bombs in 6 month intervals.

pre-2004	Jan-Jun 2004	Jul-Dec 2004	Jan-Jun 2005
0	2	4	3
Jul-Dec 2005	Jan-Jun 2006	Jul 2006	
3	2	24	

There were no deaths from car bombs prior to 2004, but mortality was largely constant for the 2.5 years from January 2004 through June 2006. There were 6 deaths in 2004, 6 in 2005 and 2 in the first six months of 2006. The entire rise in car bombs deaths come from the data for one month in one cluster.

The second problem with this cluster is the implications of having so many deaths from car bombs (or, more likely, a single car bomb) centered on a single neighborhood. Johnson et al. (2006) argue that the methodology of Burnham et al. (2006) generated a "main street bias" because interviewers were more likely to select households near main streets where violence is more common. This cluster would seem to provide evidence that main street bias might be a concern.

Conclusion

I hope that this brief description of the data is useful to those who are interested in the Lancet articles but do not have access to the data and that the tools provided here are also useful to those who have such access. I hope to extend this article in the near future to include more analysis. Please contact me with suggestions.

References

G. Burnham. Presentation at MIT, February 2007. URL `http://web.mit.edu/webcast/tac/2007/mit-tac-wgbh-e51345-27feb2007-220k.ram`. [Online; accessed 14-May-2007].

G. Burnham, R. Lafta, S. Doocy, and L. Roberts. Mortality after the 2003 invasion of Iraq: a cross-sectional cluster sample survey. *The Lancet*, 368: 1421–1428, October 2006.

N. F. Johnson, M. Spagat, S. Gourley, J.-P. Onnela, and G. Reinert. Bias in epidemiological studies of conflict mortality, 2006.

L. Roberts, R. Lafta, R. Garfield, J. Khudhairi, and G. Burnham. Mortality before and after the 2003 invasion of Iraq: cluster sample survey. *The Lancet*, 364:1857–1864, October 2004.