

Do WikiLeaks and Iraq Body Count tell the same story? No!

A Comparison of the Reports of Iraqi Civilian Deaths

Background:

On October 22nd, 2010, WikiLeaks released classified military records which included descriptions of approximately 66,000 Iraqi civilian deaths. Until this time there have been competing estimates of civilian deaths in Iraq. Since the beginning of the war, these estimates have tended to fall into two categories. The first category comprises passive monitoring systems which use sources such as Iraqi government records, press reports, or these sources plus a corrective factor. 1,2,3 Reports based on passive monitoring have included a range of estimates that generally cluster around 50,000 deaths by mid-2006 and greater than 100,000 deaths as of today. These estimates contrast with the second category of estimates, population-based surveys and polls. Population-based surveys and polls have generated violent death estimates up until some specific point in time that were roughly 3, 6, 10, and 20 times the passive estimates. 4,5,6,7

The New York Times was given exclusive US pre-release review rights to the so called WikiLeaks "Iraq War Logs" which were released on October 22nd, 2010. The Times was careful not to attribute these findings to supporting a particular civilian death total. But, many US papers ran an AP wire service article that stated: "The U.S. military has recorded just over 66,000 civilian deaths, according to the documents posted by WikiLeaks. Iraq Body Count (IBC), a private, British-based group that has tracked the number of Iraqi civilians killed since the war began, said in a press release that it had analyzed the information and found 15,000 previously unreported deaths, which would raise its total from as many as 107,369 civilians to more than 122,000 civilians. The Iraqi government has issued a tally claiming at least 85,694 deaths of civilians and security officials were killed between January 2004 and Oct. 31, 2008."8 This statement suggested that there was independent consistency among the reports, and perhaps even validation of estimates in the zone of 100,000 civilian deaths in Iraq. Many papers like the Washington Post included statements in their coverage such as, "There appear to be no major revelations in the latest logs."9 On October 26th the Post ran an editorial entitled, "WikiLeaks's leaks mostly confirm earlier Iraq reporting."10 The editorial concluded, "...claims such as those published by the British journal [The Lancet](#) that American forces slaughtered hundreds of thousands are the real 'attack on truth.'" Thus, the Washington Post explicitly, and other papers implicitly, sided with the narrative of Iraqi death tolls created by the passive surveillance sources like Iraq Body Count.

Over the past months, a group from the Columbia University School of Public Health has evaluated Iraq Body Count's (IBC) finding that the 66,000 civilian deaths reported by WikiLeaks only included about 15,000 new death reports, and that by implication 3/4th of the deaths in WikiLeaks were already reported and known through the IBC online database. Such analysis is crucial to evaluating the Washington Post's explicit claim and the general press interpretation that the war logs provided no new information about the civilian death toll in Iraq.

Methods

- a) **Iraq War Logs were downloaded** from the website of the UK newspaper, The Guardian, which had provided the data from the Iraq War Logs in excel format. Those files are available at <http://www.guardian.co.uk/news/datablog/2010/oct/23/wikileaks-iraq-data-journalism>.
- b) On October 25, 2010, the **Iraq Body Count data file of incidents was downloaded** to preserve a file uninfluenced by the WikiLeaks data.
- c) The WikiLeaks data file was **filtered to only contain civilian deaths**. These 33,809 entries contained reports of 65,679 civilian deaths. These deaths were recorded as 'killing events.' A killing event could include one or several deaths. Bombings, for example, generally included at least several deaths, while shooting events frequently entailed 1 death.
- d) Initially, each of 23 students in a graduate class on epidemiological methods for documenting human rights abuses was systematically (based on the date field) **assigned 100 killing event data entries to review** and determine if the killings described in those reports were included in Iraq Body Count's dataset, not included, or might be included.
- e) After the initial evaluation of 2,300 data files, the **participants developed a set of decision rules** to standardize the matching process. Participants were asked to code each WikiLeaks killing event record as: 0=no match, 1=unlikely match, 2=likely a match, 3= high probability it is a match. These included that to be coded above 0, two events must:
 - Not have contradictory locations or be contradictory types of event (in broad terms, shootings, torture, airstrikes, or explosive devices were different types of events).
 - IBC killing event reports could not be later than 2 days after the WikiLeaks event time. Moreover, IBC reports could not be any days before the WikiLeaks event time. (This is because the press reporting process might have delayed the arrival of reports in IBC. However, since IBC data are based on press reports, it would be impossible for them to have dates prior to those recorded in the WikiLeaks data). If there were multiple events in WikiLeaks that corresponded to multiple events in IBC, then participants were to match only the number that appeared in IBC. In other words, if, for example, there were 4 different shooting killing events in one neighborhood in Baghdad recorded in WikiLeaks and only 3 potential matches in IBC, then only 3 killing event matches could be recorded. If there was a disparate number of deaths in a killing event that otherwise appeared to be a likely match, then the WikiLeaks deaths could be considered matched or likely matched based on the reviewer's judgment.
- f) To aid with the matching process, the software Google Earth was utilized and a **Google Earth template was developed** to identify specific neighborhoods in Baghdad.
- g) Once the dataset had been examined by an individual, **a second individual** who was unaware of the original evaluator's identity or matching score (all 1,2,3 scores were converted to "M") was given the dataset and asked to **give his/her own match ranking (0,1,2,3) to any entries that were coded above 0** by the original evaluator. This was done to evaluate the consistency of the matching process.

- h) A third **final arbiter was asked to review** those cases where match rankings given by the first two reviewers differed. The arbiter was not permitted to move the average rankings two units (e.g. a 2 & 3 ranking on the first two rounds could not be changed to a 0). If the arbiter felt that the score was off by two units or more, a second arbiter (4th reviewer) had to agree. At this third arbitration review, to receive a 3 ranking, the date of the event could not differ by more than 1 day in the two reports. Thus, the final data file involves data matches reviewed by at least two people, and if the first two did not exactly agree, the ranking is based on at least the judgment of three people. Reports that might only potentially match to the morgue tallies that were in IBC (IBC contained morgue tallies with no information about the individuals included therein) were given a score of 0 or 1 based on the reviewers' judgment. Previous evidence suggests that most violent deaths in Baghdad did not go to a morgue. 11
- i) An unintended natural experiment occurred when the War Logs file was imported into a US version of Excel. The software resorted the Iraq War Logs records by day so that all killing events on January 1st were listed first regardless of year. Because 1/3rd of the reviewers were initially told incorrectly the date data was in the format of month/day/year, some of these initial reviews attempted to match to the wrong date. These mismatches were used to gain insight into the rate of false positives but the data were not further explored.
- j) Some data fields in the War Logs were incomplete and did not allow proper matching. Others involved some question as to whether the victims were civilians or combatants, or whether the event was an accident or intentional violence. When in doubt, the data record was excluded.
- k) A systematic sample of ten locations within the data with a random start number was taken of the final file. At each interval, 20 negative (match=0) records were **examined to determine the fraction of false negatives** in the dataset.
- l) Match rates were defined as the number with a specific score divided by all fully evaluated reports. Match rates by the number of civilian casualties associated with an event was achieved by filtering the War Logs data set by casualty number and dividing the number with each score by the total number in that group. Total number of deaths matched by score was obtained by filtering the dataset by match number and dividing the total with a specific match number by the total number of civilian deaths.

Results

In total, 2409 files were examined by the three or four reviewer process. An additional 88 reports (3.5%) were unmatchable because they lacked information such as the date or because they were deemed to have been an accident and not conflict related.

By any measure of match (scores=1, 2 or 3), the majority of events described in the WikiLeaks War Logs were not included in the Iraq Body Count database. The figure below shows the fraction of 2409 War Log records that were matched to records in IBC. It is likely that 19.3% of records matched. 8.7% of records have a small chance of being listed in IBC. These were primarily single killings in Baghdad. Three fourths of records (72.0%) were judged as very likely not being in IBC.

Figure 1: Fraction of War Logs in IBC

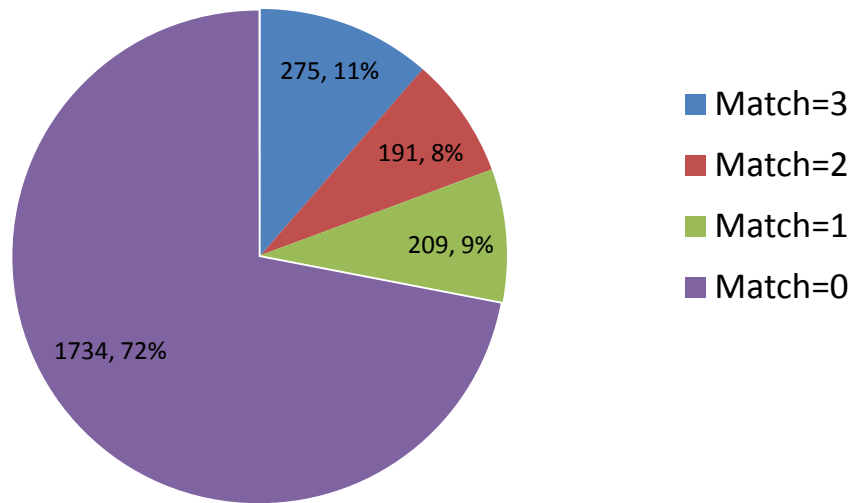
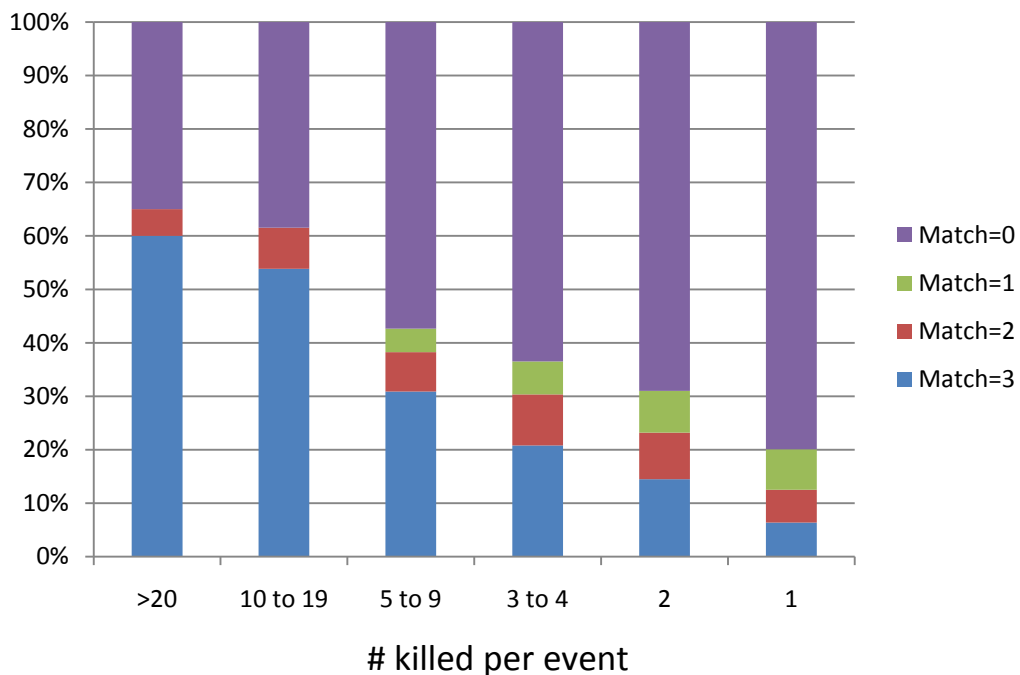
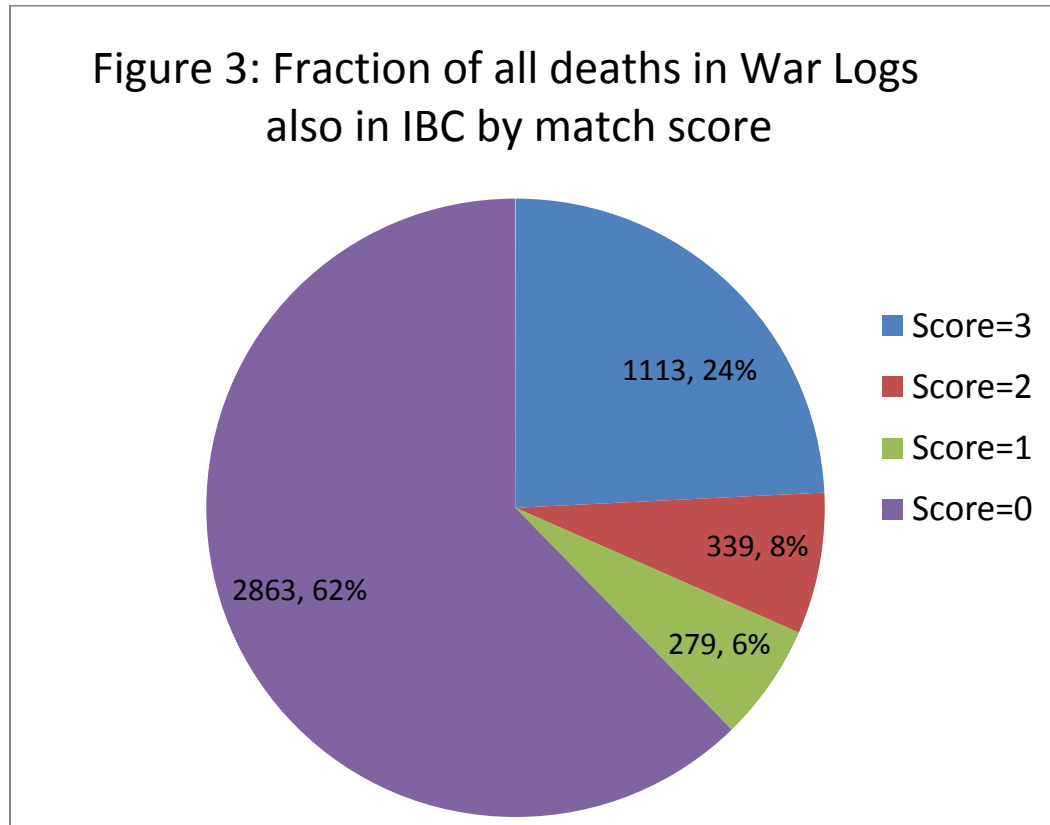


Figure 2 shows the fraction of events that match as a function of the number of civilians reportedly killed. Events that killed many people were far more likely to be reported in both datasets than were killing events that involved few people. The majority of War Logs reports (73.6%) were single killing events and only 15.6% of these were judged to be in IBC.

Figure 2: Matching scores by # civilians killed per event



Because the larger events were most likely to be matched to IBC, we assessed the fraction of total deaths (not reports of killing events) reported in WikiLeaks that were also in IBC. Figure 3 shows these results. Roughly two-thirds of deaths, probably 68%, were not included in IBC in spite of the fact that the larger killing events were likely to be recorded by both listings. This is because the vast majority of reports involved only one or two deaths.



In the initial review, among 822 records which were matched to the wrong date, 72 (8.8%) were matched (e.g. a score of 3) and 132 (16.1%) were matched with a score of “2”. In this same initial round, 135 or 515 records (26.2%) matched to the correct date were initially (before the decision rules and Google Earth tool were developed) given a score of 3 and 43 (8.3%) a score of 2. The process of using the Google Earth Map, more clarified matching rules, and the process of having three reviewers lowered the match rate from 34% to 19%. In a best case scenario, if we assume the more rigorous procedures would have caught almost all of the “2” attributed to the wrong date, it is likely that several percent, perhaps 6 or 8%, of the WikiLeaks reports which we judged to be in IBC are in fact not, but are false positive matches. Given the different language and geographic delineations used between the two lists, especially within the city of Baghdad, we found it exceedingly difficult to assure against false positive matches.

Because the triple review process was only applied to the initial positive matches, a systematic random sample was taken of 20 non-matches (match score=0) from ten locations within our finalized reviewed data file. Among these 200 “0” matches, one was deemed to be a “2” and 20 were deemed a “1.” Most

“1”s were single killing events in Baghdad. Given that 1 means that it is probably not a match, this suggests a false negativity rate of ~0.5%.

Limitations

Several serious limitations exist with this analysis. These include:

- i. We have no capacity to assess if the Iraq War Logs are truly U.S. Department of Defense reports or are in part, or in total, fabrications.
- ii. It is likely that the WikiLeaks Iraq War Logs and IBC have common sources and are correlated listings. This would artificially increase the overlap between the datasets and make the overall death toll seem lower than it really is (that is more than 3 to 5 times the individual sums of these listings).
- iii. While it is likely that the false positive rate is far greater than the false negative rate, these estimates were made by having multiple persons independently review the same datasets. If there is a correlation that results in two people being likely to mismatch the same records, these estimates of false positive and false negative reports could be inaccurate.
- iv. Any conclusions about the fraction of deaths missed by the IBC data set must assume that the “sensitivity” of IBC was consistent between 2003 and October 2010.
- v. Single killings in Baghdad were particularly difficult to match. This difficulty was complicated by the fact that the IBC database has an undetermined number of double counted deaths. Since they include press reports that generally provide detail about the event, and morgue tallies that provide no details about the persons included, it is assumed that there are an unknown number of deaths that appear in a press report and a morgue tally.
- vi. After the initial reviews were done but before the second or third reviews were completed, cyber-attacks on WikiLeaks’ website prevented reviewers from accessing the supportive descriptions of the War Logs reports. This may have increased discrepancies between reviewers for the small fraction of reports where the supportive description was important.

Discussion

Our analysis of the WikiLeaks Iraq War Logs suggests that hundreds of thousands of civilians were killed since the invasion of Iraq. The widespread dissemination of a contrary message is a profound testimony of the US media’s inability to digest detailed reports with detailed analysis. It took 10 minutes of initial review of the Iraq War Logs for the study team to realize most events outside of Baghdad did not even have a comparable event in the same Governorate that day in IBC. Thus, while this analysis represents hundreds of hours of work, a brief check of a handful of reports would have shown any reporter writing about the WikiLeaks October releases that most of these killings were previously un-reported.

There is a practice developed in wildlife biology for estimating population numbers from incomplete independent samples called capture-recapture. The logic goes: if you try to find every moose on an isolated island, and you find 10, and put a tag in their ear, and then you go and try again later and you find 10, but only 5 of them have been tagged, it must be there are 20 moose, twice as many as initially found. This assumes the population is not moving in and out, and that the samples of moose were

*On Nov. 30th 2010, after this project was underway, the Office of Career Services at the Columbia School of International and Public Affairs sent an e-mail to students on the anonymous advice of an alumni employed by the US Department of State that said, “He recommends that you DO NOT post links to these documents nor make comments on social media sites such as Facebook or through Twitter. Engaging in these activities would call into question your ability to deal with confidential information, which is part of most positions with the federal government.” Thus, all graduate students have been strongly advised against placing their name on this document.

References

1. While the Official Iraqi estimates have tended to be released sporadically through the press, the Congressional Research Service has compiled reports through 2008 at <http://www.fas.org/sgp/crs/mideast/RS22537.pdf>
2. Iraqbodycount.org
3. Brookings Institution Iraq Index. <http://www.brookings.edu/saban/iraq-index.aspx>
4. Iraq Family Health Survey Study Group. *New England Journal of Medicine*, Vol. 358, Issue 5, Pages 484-493.
5. Roberts L, R.Lafta, R.Garfield, J.Khudhairi, G.Burnham. (2004). Mortality before and after the 2003 invasion of Iraq: cluster sample survey. *Lancet*, Volume 364, Issue 9448, Pages 1857-1864
6. Burnham G, Lafta R, Doocy D, Roberts L. (2006). Mortality after the 2003 invasion of Iraq, cross-sectional cluster sample survey. *Lancet*; Issue 368, Pages 1421-28.
7. ORB’s website includes a press release of these findings. http://www.opinion.co.uk/Newsroom_details.aspx?NewsId=120
8. See for example, <http://www.washingtontimes.com/news/2010/oct/22/wikileaks-plans-major-announcement-europe/>
9. Greg Miller and Peter Finn. “Secret Iraq war files offer grim new details.” <http://www.washingtonpost.com/wp-dyn/content/article/2010/10/22/AR2010102201682.html?sid=ST2010102206224>
10. Washington Post. “WikiLeaks's leaks mostly confirm earlier Iraq reporting.” <http://www.washingtonpost.com/wp-dyn/content/article/2010/10/25/AR2010102504643.html>
11. Siegler A. et al. (2008). Media Coverage of Violent Deaths in Iraq: An Opportunistic Capture-Recapture Assessment. *Prehospital Disaster Medicine* Volume 23, Issue4, Pages 369–371.

Appendix 1: Capture-recapture type comparison ignoring correlation.

The capture-recapture process is now widely applied to a variety of public health settings. Annually, the logic is used to count homeless in New York City by planting decoys across the city on the nights when enumeration takes place and assessing the fraction of the decoys identified. Most commonly, it is used to count wildlife.

In our moose example above, where 10 moose were identified and tagged, in the next round 10 moose are again caught but only 5 are tagged, we would assume the true population was twice the 10 originally seen. The table below outlines the math associated with estimating the population from two lists. In our moose counting example, A equals the 5 moose caught twice. B is the 5 moose caught the first time but not the second. C equals the 5 moose caught the second time but not the first. D would equal 5 because there are 5 moose out in the wild assumed to be there even if they have never been seen.

Table 1: Simplest type of capture-recapture analysis.

	On list 2	Not on List 2
On list 1	A	B
Not on list 1	C	D – unseen but real

If $AD=BC$, $D=BC/A$; Total = $A+B+C+D$

In this logic, the more complete list 1 and list 2, the smaller is D, the unseen portion. In fact if either list was complete, D would go to zero.

On October 25th, 2010, IBC estimated they had tallied between 99,000 and 108,000. If we apply this capture-recapture logic to the data we have from the Iraq War Logs and IBC using the October 25th midpoint of the IBC range, we would get the following:

	In IBC	Not in IBC	
In WL	20,856	45,144	Total WL=66,000
Not in WL	81,644	D	
	Total IBC= 104,500	D= 176,723	Total = 324,367

In the most conservative analysis, the WikiLeaks War Logs data when compared to IBC suggest over 300,000 deaths. If one assumes a 0.5% false negative and 6% false positive rate (best guess), the fraction of overlap would be 26.1% and death toll would be over 400,000. On the extreme other end, if we assume all morgue tallies are double counts, the death toll would be reduced by 20%.