Technical Memo for Amnesty International Report on Deaths in Detention

Megan Price, Anita Gohdes, & Patrick Ball*

August 18, 2016

*The authors thank: Jule Krüger and Kristian Lum for their constructive review of an earlier version of this report; James Johndrow and Daniel Manrique-Vallier for their methodological advice; the teams at CSR-SY, DCHRS, SNHR, and VDC for their brave and important work collecting information about victims in Syria; our project partner, the Human Rights, Big Data and Technology Project, funded by the Economic and Social Research Council; Ashraf Kheir, Mazin Abdelrahman Mohamed, Tayseer Mohammed Osman, and Michelle and Chris Dukich for their tireless work reviewing and classifying these records.

1 Executive Summary

This report is based on records collected by four sources investigating deaths in the ongoing conflict in Syria (described below and in Appendix A). For the period 15 March 2011 through 31 December 2015, we find a total of 12,270 documented, identifiable people killed while the victim was in detention.[1] Using a statistical method called multiple systems estimation (MSE), we calculate that an estimated total of 17,723[2] victims, both documented and *undocumented* were killed in detention during this time period. This implies that overall approximately 25% of killings that occurred in detention during this time were not documented by any of these four sources.

It is important to note that our estimate of 17,723 victims is likely to be an under-estimate. This is due to decisions we made regarding what constituted a killing that occurred in detention; this is described in more detail in Section 7.

There have been other investigations into deaths in detention; a particularly well-known example is the 2014 Syrian detainee report, also known as the Caesar Report. In Section 8, we describe why we believe that the statistical findings of the Caesar Report overestimate deaths in detention.

The sections of this report represent the chronological steps in our analysis: receipt of data from our partner organizations, classifying killings that occurred in detention, linking multiple records that refer to the same victim, imputing missing information about detention, and calculating MSE estimates. We then close with a few sections reflecting on this and other published estimates of deaths in detention. Several appendices provide further technical details.

2 Data Sources

The data used for this analysis are similar, but not identical, to the data we have used in previous descriptive reports published with the United Nations Office of the High Commissioner for Human Rights (OHCHR)(Price et al., 2013a,b, 2014). Four sources were used for this analysis. For more information about each source, see Appendix A.1:

- Syrian Center for Statistics and Research (CSR-SY)
- Damascus Center for Human Rights Studies (DCHRS)
- Syrian Network for Human Rights (SNHR)
- Violations Documentation Center (VDC).

[1] A small fraction of the total number of uniquely identified victims were missing sufficient information to determine whether or not the killing occurred in detention. These were taken into account in the estimate presented here through statistical imputation, described in Section 5.

[2] 95% credible interval (13,409, 18,713)

For brevity, each source is referred to by its acronym throughout this report. To integrate the four sources, we standardized the structure and content of the different sources; for full details see Appendix A.

3 Classifying Deaths in Detention

Each of the four organizations that collected data on victims in Syria include, whenever possible, information about the circumstances of that victim's death. This information is recorded in Arabic; we relied on Google's translation application programming interface to translate this information into English.[3] Two reviewers then read all of the incident details to classify the killings as deaths that occurred while in detention. The reviewers developed a set of rules to classify killings that occurred in detention, including if a victim is described as being:

- held by any of the regime's forces and executed
- held by any of the regime's forces and tortured
- arrested, detained, or kidnapped by any of the regime's forces, even if the body was later found outside of a detention center
- arrested, detained, tortured, or killed at one of the branches or prisons specifically named in Human Rights Watch's 2012 report "Syria: Torture Centers Revealed"

This is a strict classification of deaths that occur in detention. In particular, a number of "field executions" were described in the records, often of soldiers for refusing to fire. These were not considered to be deaths that occurred in detention unless additional details specified that the victim was arrested, imprisoned, or tortured.

In a small number of cases, details seemed to imply a killing that occurred in detention, but the perpetrator(s) was (were) not identified. This report (and the corresponding Amnesty International report "'It breaks the human': Torture, disease and death in Syria's prisons") is primarily focused on killings that occurred in state-run detention centers. Therefore, in these cases detention status was labeled as missing and was later imputed using a statistical model (for full details, see Section 5).

Additionally, some records lacked any information about the circumstances of a victim's death. These were also treated as missing, and whether the killing was in detention was imputed using a statistical model (for full details, see Section 5).

[3]We feel confident about using this translation since our analyses over the past three years have shown that reviewers examining the data in the original Arabic and in Google translated English make the same decisions regarding records that refer to the same individual. This is described in more detail in Appendix B.

4 Record Linkage

In order to know how many deaths have been documented, we need to identify the multiple records that refer to the same individuals. These records may be within the same list (e.g., the same source records the same victim multiple times, perhaps because he or she was reported by multiple community members) or across sources (e.g., different sources record information about the same victim). The records may contain impartial and imperfect information, and the records contain no unique identifying number such as a national id number. The challenge of deduplicating databases with imperfect information arises in a variety of contexts and has been studied across disciplines for decades. It is called "record linkage" when there are two databases, and "database deduplication" when there are three or more. Dunn (1946) and Newcombe et al. (1959) first formalized the problem and Fellegi and Sunter (1969) developed some of the first theory. Winkler (2006), Herzog et al. (2007), and Christen (2012) provide overviews of the problem and various methods.

In our case, we begin with records of identifiable victims. An identifiable victim is a record that includes at least two words from the victim's name, plus the date and location of death. The full identifying information is essential for the comparisons required to match records to each other. Records lacking the complete information are considered "anonymous" and are excluded from the analysis. The anonymous records describe victims of violence in the Syrian Arab Republic who deserve to be acknowledged. However, they cannot be included in this step of the analysis because it is impossible to determine if the records with partial information refer to killings also described by other records. That is, anonymous records cannot be matched or de-deduplicated. Records with partial information provide hints about the existence of killings which have not been fully documented; estimating the number of *undocumented* killings is the goal of the next step in our analysis (see Section 6).

To identify multiple records that refer to the same individual, we employ a combination of human review and computer modeling. Humans review subsets of records—some in the original Arabic, others in translated English[4]—and combine groups of records that they believe refer to the same individual. Computer algorithms are then used to model the decisions the humans made, to assign a probability that any two records refer to the same individual. These probabilities are then used to cluster records into groups that represent the information available about a single individual across all the data sources. This is called "semi-supervised" modeling.

[4]We have found that these reviewers make highly comparable decisions, regardless of the language in which they review the records. See Appendix B for details.

4.1 Hand-Labeled Data

Although this report only considers killings that occurred while the victim was in detention, for the record linkage part of this analysis we began with all victims recorded by one or more of the documentation groups, regardless of the circumstances of their death. Later, we filtered records based on the classification described in Section 3. As a result, we begin with 413,954 total records of victims, both those killed while in detention and those killed under other circumstances. It is important to note that this number refers to the total number of records prior to matching and de-duplication—it should *not* be inferred to indicate the total number of victims killed in the ongoing conflict. This count includes many examples of duplicated information.

From this set of total records, we selected small groups of records with similar names, locations, and dates of death, which a human reviewer sorted into even smaller groups of records, called clusters, in which all the records in each cluster refer to the same person. From these clusters we can identify pairs of records that refer to the same person ("positive pairs" or "matches", of which there were 128,741 pairs) and pairs of records that do *not* refer to the same person ("negative pairs" or "non-matches", of which there were 427,638 pairs). It is useful to organize records as either clusters or sets of pairs for various different steps in the record linkage process. These human-labeled pairs and clusters are used to evaluate decisions in the next step (blocking) and to train the pairwise classification model described in the classification step.

4.2 Blocking

In order to link records that refer to the same person, we create a model that estimates the probability that any *pair* of records refer to the same person. Rather than estimate this model on the full combination of all possible pairs (which would be approximately 85 billion pairs), we limit the consideration to only the pairs that have a reasonable chance of being matched. This process of limiting the analysis to a subset of pairs is called "blocking" or "indexing."

In brief, we consider the total number of positive pairs identified in the hand-labeled data, looking for combinations of common field values that define subgroups ("blocks") within which all the positive pairs are found.[5] The rules for creating these blocks are described in Figure 1 in Appendix D.

This approach covered all but 0.4% of the hand-labeled positive pairs and generated a total of 44,448,855 pairs. This is the set of pairs that must be considered in the remaining steps.

[5]For full technical details, please see this blog post.

4.3 Feature Definition

The primary bases for comparing records are the name of the victim, and the date and location of the death. In order to represent the similarities and differences among records, there are many possible comparisons among these fields, including whether the values in a field in two records are exactly equal, or much more complicated comparisons. According to the pairwise classification models that we have considered, the most useful comparisons for determining matches/non-matches are:

- Given a list of all the names (in English and in Arabic), sort them alphabetically, then calculate the number of deletions and insertions required to transform one into the other (this is called the Jaro-Winkler distance). This is an edit distance normalized to the length of the string.
- Calculate the number of first, middle, and last names common to both records divided by the total number of names in either record (this is called the Jaccard index). Calculated for both English and Arabic versions of the names.
- The Jaccard index for the words in the location descriptions (in Arabic).
- Whether the names share substrings (measured by locality sensitive hashing, see Rajaraman and Ullman (2011), especially chapter 3), the first five characters, or the last five characters (in Arabic).
- The number of days between the two dates of death.
- Whether the year, month, and governorate are exactly the same.

We found a total of 32 comparisons to contribute substantially to the probability of matching. These comparisons are the "features," or predictor variables, used in the pairwise classification of whether a given pair is or is not likely to be a match.

4.4 Pairwise Classification

The hand-labeled data, both positive and negative pairs, was randomly divided into two groups, one for training with 440,464 "training" pairs, and a second group of 115,915 pairs used for testing. The training records were given to a "gradient boosted trees" algorithm[6], and the resulting model was used to classify the testing data.

Note that the pairs in the training and testing sets are "hard." That is, we did not include pairs that are obviously non-matches, and many of the

[6]The implementation described here.

matches are pairs with slightly different dates or names. The classification was nonetheless quite accurate. The results of classifying the testing pairs are shown in Table 1.

Table 1: Confusion Matrix for Held-Out Testing Data

		model	
		match	non-match
human	match	27,510	4,521
	non-match	3,089	80,795

Table 1 shows that 27,510 pairs are classified by both the human reviewer and the model as matches; these are called "true positives." In the next cell, 4,521 pairs were classified as matches by the human reviewer but as non-matches by the model, these are "false negatives."

Combined, the confusion matrix creates a mean positive-negative F_1 score of 0.917. Another way to evaluate a pairwise classification model is through the calculation of a Brier score. In our case this metric suggests that on average, the classification scores are approximately 0.23 away from the hand-labeled values of zero (for non-match) or one (for a match).

The model was applied to the 44,448,855 pairs generated by blocking; each pair was assigned a probability of being a match.

4.5 Clustering

Once the records are classified, we need to decide which groups of records refer to the same person; together, the records that refer to a single person are called a *cluster*. There may be one, two, or more records in a cluster.[7]

Our approach first partitions the records into groups via transitive closure, linking all the pairs which have even a small probability of being matches (a classification score > 0.4) into a super-cluster, called a "connected component."

We next separate each connected component into smaller clusters which maximize the similarity of the records to each other using a method called "hierarchical agglomerative clustering" (HAC). [8]

4.6 Merged Records

For clusters with more than one record, the information across all the records must be merged to form a single, unified record. For a few groups of records, this means that contradictory information from different records that refer

[7]See this recent blog post for additional technical aspects of clustering.

[8]Specifically, we use an average weighting method and a threshold-based cluster flattening, with $t = 0.4$. This is a distance measure rather than similarity measure, so it is slightly more strict than the threshold used in transitive closure.

to the same person must be resolved. When there are multiple records in a cluster with differing values, the most common (nonmissing) value for each field is chosen; ties are broken by random selection.

5 Imputing Missing Data

After record linkage was complete, a total of 3,892 records were missing information about detention classification (as described in Section 3) because the original raw records did not include any details about the circumstances of death. An additional 4,766 records were also missing detention status because the original raw records did include some details about the circumstances of death, but those details did not identify a perpetrator. Since this report is concerned only with killings that occurred in state-run detention centers, these records also must be treated as missing since we cannot conclude based on the information in the original record whether the perpetrator was a representative of the state.

These records account for a small proportion of the total number of uniquely observed victims (both those killed in detention and those killed under other circumstances). This relatively small amount of missing data is unlikely to have much impact on substantive conclusions, however the statistically appropriate way to handle missing data is through imputation. Imputation means using a statistical model to predict what the missing values might have been, based on the values that were observed.

In this particular case we developed two different models, one for records missing details about circumstances of death and one for records with details but missing perpetrator information. Both were logistic regression models, using information about the documentation pattern (whether a record was documented by each source), date, and location of death as the predictors.

The second model, used for records missing perpetrator information, used an additional 26 predictors describing the presence or absence of the most frequently reported "tokens" in the phrases or sentences describing the circumstances of death (a "token" is a single word from a string of text, for example "arrest" or "prison"; see Appendix C for details). Both models predicted a probability of custody status for each record missing this information. Custody status was then drawn from a binomial distribution with that probability.

Some of the fully observed records were held out as a test set to evaluate the performance of these two models. Specifically, we examined how each predicted known detention classifications. We can think of these predictions much like the confusion matrix in Section 4, where some predictions will be true positives, others false negatives, etc. A common measure to summarize this information is the Area Under the Curve, or AUC. For the first model, using only documentation pattern, date, and location of death, the AUC was 0.7. For the second model, including information about "tokens" the AUC

was 0.99. Values closer to 1 indicate a better performing model. Perhaps not surprisingly, when we have information about the circumstances of death, even if perpetrator is uknown, our predictions are better, but even lacking this information our model performs adequately.

Finally, these models were used to calculate twenty five imputed datasets, meaning that 25 different possible versions of the set of merged records, with complete detention classification, were calculated. MSE estimates were then calculated for each of these 25 datasets and results were averaged across the datasets. In this way the final MSE estimates presented in this report account for any additional uncertainty introduced by imputation.

6 Multiple Systems Estimation (MSE)

By merging the multiple records that refer to the same individual, we create a single list with one row for each uniquely identifiable victim. The row contains information about which source(s) recorded information about that victim. The number of victims documented by a single source, by each possible combination of two sources, three sources, and all four sources, provides insight into the size of the total victim population. In other words, by examining the documentation patterns across the four sources, we can learn about the only partially observed underlying population that generated those documentation patterns. We organize this information, referred to as "overlap patterns", as shown in Table 2.

As noted above, missing values were imputed, creating 25 datasets on which MSE estimates were calculated. Table 2 shows just one possible imputation example. Therefore the total estimate implied by this table does not match precisely the total estimate reported in Section 1 because it is only one of 25 possible (but very similar) estimates; the total reported in Section 1 was averaged over all 25 imputed datasets.

A '1' in a column indicates records documented by that source. For example, the first row of Table 2 contains a '1' under each source (CSR-SY, DCHRS, SNHR, and VDC) and indicates that 1,493 records of victims killed in detention (according to this particular imputation) were documented by all four sources. The second row has a '1' only under three of the sources (CSR-SY, DCHRS, and VDC) and indicates that only 248 records of victims were common across these three sources.

The goal of this step in the analysis is to estimate the last row, the number of victims that have not yet been documented by any of these sources. Our total estimate (for this imputation) of 17,848 implies an estimate of 4,518 as yet undocumented victims.

For each imputed dataset, we calculate this estimate using a class of methods called multiple systems estimation (MSE). MSE has been used over the past century by ecologists (Peterson, 1894; Lincoln, 1930; Otis

Table 2: Distribution of Records into Four Sources for One Imputation

CSR-SY	DCHRS	SNHR	VDC	Number of Records
Documented				
All Four Sources				
1	1	1	1	1,493
Three Out of Four Sources				
1	1	0	1	248
1	0	1	1	2,599
0	1	1	1	457
1	1	1	0	301
Two Out of Four Sources				
1	0	0	1	1,302
0	1	0	1	171
1	1	0	0	163
0	1	1	0	350
0	0	1	1	1,148
1	0	1	0	602
Unique to a Single Source				
1	0	0	0	1,260
0	1	0	0	473
0	0	1	0	1,177
0	0	0	1	1,586
Estimated				
Undocumented by All Sources				
0	0	0	0	4,518

et al., 1978), demographers (Sekar and Deming, 1949), statisticians (Fienberg, 1972; Darroch et al., 1993; Agresti, 1994; Fienberg et al., 1999; Fienberg and Manrique-Vallier, 2009), public health researchers (International Working Group for Disease Monitoring and Forecasting, 1995a,b), and human rights researchers (Ball et al., 2002, 2003; Brunborg et al., 2003; Silva and Ball, 2008; Zwierzchowski and Tabeau, 2010; Lum et al., 2010; Manrique-Vallier et al., 2013; Lum et al., 2013; Mitchell et al., 2013) to estimate difficult to observe populations of animals and humans. MSE includes a broad set of mathematical models, all of which are designed to use data structured like Table 2 and estimate the last row of unobserved information.

For this specific analysis we used the model developed in Madigan and York (1997) as implemented in the R software package `dga`. In this model the list overlap counts (as described in Table 2) are specified to follow a multinomial distribution with a hyper-Dirichlet distribution as a prior. In brief, this approach uses Bayesian model averaging (Hoeting et al., 1999) to incorporate uncertainty about potential relationships between the lists.

7 Why This is Likely an Under-Estimate

As described in Section 3, we implemented a strict definition of killings that occurred while a victim was in detention. As a result, a relatively small number of observed records are used to estimate the total number of both observed and unobserved records. Although 13,340 (the average imputed observed total) may not seem small, Table 2 shows that most of the overlap patterns contain relatively few records. These patterns are what we model to estimate the unobserved number of records. The particular method that we use is conservative in the sense that if there are too few records to produce a meaningful estimate, the model defaults to an estimate that is closer to the observed number of records than estimates produced by some other methods.

Additionally, although MSE methods are designed to estimate missing data, these methods can only estimate cases where there is a non-zero probability of the death being reported. In other words, we can estimate the number of unobserved victims killed in detention by assuming that those victims have something in common with the victims we were able to observe. For example, we can estimate the unobserved victims who were killed around the same time or in the same detention center as victims who were in fact observed. We cannot estimate the number of undocumented victims that are truly invisible to all the documentation groups. For example, victims whose families do not know they were arrested, or victims for whom there are no community members left in Syria and who thus may have no chance of being reported to one of our data collection partners. As a result, there are likely more total victims than we are able to estimate.

8 The Caesar Report

In 2014 a report on Syrian detainees, commonly referred to as the Caesar Report, was published. This report was prepared by an investigative team "...mandated to determine the credibility of a defector from Syria." In addition to interviewing the defector, the team's main objective was to assess the forensic credibility of approximately 55,000 photographs smuggled out of Syria, images which are described to show "signs of starvation, brutal beatings, strangulation, and other forms of torture and killing."

Although it was not their main objective, the investigative team also drew conclusions about the total number of victims based on the number of photographs. The report describes that the team examined 5,500 photographs, but it does not describe how these photographs were selected from the full collection of 55,000 photos. From these 5,500 photographs, the team estimated that 1,300 distinct, individual corpses were depicted, but it does not describe how these individuals were identified, or how photos were matched to specific victims; this is a problem analogous to our work matching multiple records to the same unique victim. From these comparisons,

the team concludes that "... most deceased persons had between four or five images taken of them," and on this basis, they extrapolate that the 55,000, photographs probably include approximately 11,000 individual victims.

It is possible that this many victims are depicted in this collection of photographs, but it is impossible to assess the statistical accuracy of this claim based on the information available. There are several statistical concerns with the estimate of 11,000 victims. For the estimate to be correct, we must assume that when images were selected for examination, *all* the images relevant to each identified victim were sampled. That is, if an individual had 6 images in the full set of 55,000 images, all 6 images must have been chosen among the set of 5,500 images that were examined. If this is not correct, if for example only 4 or 5 of the images of this victim were selected for examination, then the images per victim would be underestimated.

It seems to us unlikely that each victim's images were selected as a complete set. It was likely difficult to identify victims by the images, and some of the images relevant to each of the identified victims were likely missed when the sample was drawn for examination. Consequently, the estimated number of photos per victim is probably too low, and thus, the true number of victims in the images is probably fewer than the estimated 11,000.

Until the individuals in the photographs are all positively identified, we will not know how many victims this collection represents. The authors of the report mention that each detainee was given two identifying numbers, and that only the intelligence service knew the identity of the victims. If more information about these lists of numbers were available, we might be able to draw stronger conclusions about the likely number of victims included in the collection.

We emphasize that we are not criticizing the substantive analysis of the photographs, which seems to us extremely important and informative. Our concerns are narrowly directed to the extrapolation of the total number of victims, given the design of the examination.

9 Conclusion

Based on records collected by four organizations we identified 12,270 records with complete information sufficient to determine that they described killings that occurred while the victim was in detention. Using statistical imputation to infer information about records with missing or incomplete incident details, we concluded that an average total of 13,340[9] records describe victims who were killed while in detention. Applying MSE methods to these data we estimate that a total of 17,723[10] victims have been killed while in detention, including both those whose deaths have been reported to one or more of these sources and those whose stories have yet to be told.

[9]The full range of observed records across the 25 imputation datasets was from a minimum of 13,312 to a maximum of 13,381

[10]95% credible interval (13,409, 18,713)

A Data

A.1 Sources

- Syrian Center for Statistics and Research: This list was initially provided to HRDAG in November and December 2013. Subsequent updates to their files were shared with HRDAG in June 2014, October 2014, May 2015, and March 2016. As described on its website, "The center includes a local network of reporters and a team of researchers and academics inside and outside of Syria."

- Damascus Center for Human Rights Studies: This list was provided to HRDAG in April 2015 and updated records were shared with HRDAG in January and February 2016. The Damascus Center for Human Rights Studies maintains several documentation projects in addition to lobbying and advocating for Syrian human rights and working to draw attention to the situation in Syria.

- Syrian Network for Human Rights: This list was initially provided to HRDAG by OHCHR in August 2012. Beginning in February 2013, HRDAG established a direct relationship with SNHR. SNHR conducts monthly reviews of their records and subsequently updates their dataset with newly discovered or verified victims. SNHR shared their list and subsequent updates with HRDAG in February 2014, June 2014, October 2014, March 2015, and June 2016. SNHR maintains a website where they describe that they "adopt the highest approved documentation principles by the international bodies." Also available on their website is a description of their three phase documentation process and the six categories of victims they document.

- Violation Documentation Centre: This list was initially provided to HRDAG by OHCHR in February 2012. Subsequently HRDAG scraped[11] the website several times between 2012 and 2016 to obtain updated data. This process captures two of the lists maintained by VDC, "Martyrs" and "Regime fatalities." The "About" page of their website describes the data classification methods and three-stage data verification process implemented by the VDC.

A.2 Data Cleaning

In this step, invalid data values are filtered from the data. For example, in many datasets the "age" variable includes a combination of ages in years as well as specific birth years; for example, ages recorded as "1970" are clearly a birth year rather than an age in years. These values are subtracted from the year of death, and the difference in years is recorded as the approximate age

[11] Using a computer program to extract information from websites.

of the victim. Another data cleaning task is simply removing obvious typos from data values. For example, strings of unstructured text in otherwise numeric or categorical variables (such as age or sex) can usually be trimmed from those variable values.

A.3 Data Translation

In this step, key analysis variables, such as sex and governorate, are translated from Arabic to English. HRDAG's Syrian expert (one of the native Arabic speakers who review records) confirms the translation of these values. Other Arabic content, such as names and locations (a finer geographic description than governorate) are reviewed in their original form by the native Arabic speaking reviewers.

For other reviewers, HRDAG uses Google's translation application programming interface to translate names and locations, which they then review in English. Close comparison of these decisions to those made by the native Arabic speakers confirm a high level of consistency, regardless of whether review is conducted in English or Arabic. For full details, see the section on inter-rater reliability in Appendix B.

A.4 Data Canonicalization

In this step, analysis variables are transformed to have a common structure across all of the data sources. For example, the different datasets collect a variety of information about the location of death. These locations may be recorded across numerous variables and in varying levels of precision (e.g., neighborhood, area, governorate). HRDAG matches records based on governorate and compares results for different governorates, so the location variable must be standardized across data sources. In some cases, this is straightforward, in some cases HRDAG uses other location information (such as city) to map to governorate.

B Inter-Rater Reliability (IRR)

When two or more individuals review and code data, such as the reviewers employed by HRDAG to determined whether multiple records refer to the same individual, it is common to need to assess the consistency of the decisions made by those individuals. Formally, this assessment is referred to as inter-rater reliability (IRR) and is generally described using the overall percent agreement among coders and a kappa coefficient. There are a variety of other statistical measures to evaluate IRR, but kappa is commonly used for categorical measures, such as assigning match/non-match to groups of records.

First, the overall agreement rate is the proportion of times that multiple coders make the same decision. For example, for this project coders A and B each reviewed the same 63,249 pairs of records (coder A in English, coder B in Arabic) with overall agreement 96%. Coders B and C each reviewed the same 63,951 pairs of records (both working in Arabic) with overall agreement 95.4%. Finally, coders A (English) and C (Arabic) each reviewed the same 86,371 pairs of records, with overall agreement 94.7%.

Second, kappa is calculated as this agreement, adjusted to consider the amount of agreement that might be expected by chance. Specifically:

$$\kappa = \frac{p_a - p_c}{1 - p_c}$$

where p_a is the overall agreement and p_c is the amount of agreement expected by chance. p_c is calculated from the total number of matches and non-matches assigned by each coder. For the same combinations of coders described above, coders A and B have a kappa value of 0.813, B and C a value of 0.817, and A and C a value of 0.796.

In general, a kappa above 0.8 is considered very good, 0.6-0.8 is good, and 0.4-0.6 is considered moderate. For more about kappa, see Gwet (2012) or Hallgren (2012).

Perhaps even more important than the raw percent agreement and kappa values is the consistency of those values regardless of whether the coders being compared are both working in Arabic or one is working in English and the other Arabic. These results imply that the decisions made by each of the reviewers are highly consistent, regardless of whether they were reviewing the records with the original Arabic content or translated into English.

C Incident Detail Tokens

For the imputation described in Section 5 we needed to summarize the content contained in the incident details field, and particularly the relationship between that content and whether or not a record was likely to be categorized as describing a killing that occurred in detention. The incident details field is an unstructured text field, meaning that it contains descriptions in phrases, sentences, or, in a few cases, paragraphs. We split this text into separate words, then counted the frequency with which each word appeared among records that were coded during the hand-review as a killing that occurred in detention ("Y") or not ("N").

Perhaps not surprisingly, among those records that described a killing that occured in detention, some of the most frequently appearing words included "torture", "prison(s)", and "arrest". Among those that were not categorized as killings that occurred in detention, some of the most frequently appearing words included "clashes", "shelling", and "bombardment". After removing uninformative common words, such as "a", "an", "the," we identi-

fied 16 words (tokens) strongly associated with records coded as "N" and 10 words strongly associated with records coded as "Y". Indicators for whether or not each of these 26 tokens were present in the details field were used as predictors in the second logistic regression model described in Section 5.

D Blocking Rules

The full set of rules mentioned in Section 4.2 are:

$(name\_last5 \wedge name\_meta\_first \wedge yearmo) \cup$
$(governorate \wedge name\_en\_meta\_first \wedge yearmo) \cup$
$(sortedname\_en\_first5 \wedge name\_en\_meta\_last \wedge sortedname\_en\_meta\_last) \cup$
$(date\_of\_death \wedge name\_first5 \wedge name\_meta\_last) \cup$
$(name\_en\_last4 \wedge yearqtr \wedge reg) \cup$
$(date\_of\_death \wedge sortedname\_meta\_first \wedge sortedname\_en\_meta\_last) \cup$
$(name\_last5 \wedge name\_en\_meta\_first \wedge reg) \cup$
$(sex \wedge date\_of\_death \wedge sortedname\_en\_first5) \cup$
$(sortedname\_en\_last5 \wedge name\_en\_no\_mo\_lsh \wedge yearmo) \cup$
$(date\_of\_death \wedge sortedname\_last5 \wedge reg) \cup$
$(sex \wedge date\_of\_death \wedge location) \cup$
$(location \wedge sortedname\_en\_meta\_last \wedge year) \cup$
$(sortedname\_last5 \wedge sortedname\_en\_first5 \wedge yearmo) \cup$
$(name\_first5 \wedge name\_en\_meta\_last \wedge yearsem) \cup$
$(date\_of\_death \wedge name\_en\_meta\_first \wedge reg) \cup$
$(name\_en\_last4 \wedge name\_en\_meta\_first \wedge sortedname\_en\_meta\_last) \cup$
$(governorate \wedge name\_last5 \wedge sortedname\_last5) \cup$
$(sortedname\_first5 \wedge sortedname\_en\_meta\_last \wedge yearmo) \cup$
$(date\_of\_death \wedge governorate \wedge sortedname\_lsh) \cup$
$(location \wedge sortedname\_en\_first5 \wedge yearqtr) \cup$
$(age \wedge governorate \wedge name\_first5) \cup$
$(governorate \wedge sortedname\_en\_last5 \wedge yearqtr) \cup$
$(location \wedge sortedname\_last5 \wedge yearmo) \cup$
$(sortedname\_en\_last5 \wedge name\_en\_meta\_first \wedge yearmo) \cup$
$(age \wedge sortedname\_first5 \wedge yearsem) \cup$
$(location \wedge sortedname\_first5 \wedge name\_en\_first4) \cup$
$(name\_en\_first4 \wedge name\_en\_meta\_last \wedge yearmo) \cup$
$(sex \wedge sortedname \wedge name\_first5) \cup$
$(date\_of\_death \wedge name\_en\_meta\_last \wedge sortedname\_lsh) \cup$
$(date\_of\_death \wedge name\_en\_first4 \wedge name\_lsh) \cup$
$(location \wedge name\_en\_first4 \wedge name\_en\_last4) \cup$
$(location \wedge sortedname\_first5 \wedge yearmo) \cup$
$(age\_group \wedge name\_en\_first4 \wedge name\_en\_last4)$

Figure 1: Blocking Rules

References

Agresti, A. (1994). Simple Capture-Recapture Models Permitting Unequal Catchability and Variable Sampling Effort. *Biometrics*, 50(2):494–500.

Ball, P., Asher, J., Sulmont, D., and Manrique, D. (2003). *How many Peruvians have died?* American Association for the Advancement of Science, Washington, DC.

Ball, P., Betts, W., Scheuren, F., Dudukovich, J., and Asher, J. (2002). *Killings and Refugee Flow in Kosovo, March–June, 1999*. American Association for the Advancement of Science and American Bar Association's Central and Eastern European Law Initiative, Washington, D.C.

Brunborg, H., Lyngstad, T. H., and Urdal, H. (2003). Accounting for Genocide: How Many Were Killed in Srebrenica? *European Journal of Population*, 19(3):229—-248.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, New York.

Darroch, J., Fienberg, S., Glonek, G., and Junker, B. (1993). A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability. *Journal of the American Statistical Association*, 88(423):1137–1148.

Dunn, H. L. (1946). Record linkage. *American Journal of Public Health and the Nations Health*, 36(12):1412–1416.

Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

Fienberg, S. (1972). The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables. *Biometrika*, 59:591–603.

Fienberg, S., Johnson, M., and Junker, B. (1999). Classical Multilevel and Bayesian Approaches to Population Size Estimation Using Multiple Lists. *Journal of the American Statistical Association*, 162(3):383–405.

Fienberg, S. E. and Manrique-Vallier, D. (2009). Integrated Methodology for Multiple Systems Estimation and Record Linkage Using a Missing Data Formulation. *Advances in Statistical Analysis*, 93(1):49–60.

Gwet, K. L. (2012). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters*. Advanced Analytics, LLC.

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quant Methods Psychol*, 8:23–34.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*. Springer.

Hoeting, J. A., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14(4):382–417.

International Working Group for Disease Monitoring and Forecasting (1995a). Capture-Recapture and Multiple-Record Systems Estimation I: History and Theoretical Development. *Am. J. Epidemiol.*, 142(10):1047–1058.

International Working Group for Disease Monitoring and Forecasting (1995b). Capture-Recapture and Multiple-Record Systems Estimation II: Applications in Human Diseases. *American Journal of Epidemiology*, 142(10):1059–1068.

Lincoln, F. (1930). Calculating Waterfowl Abundance on the Basis of Banding Returns.

Lum, K., Price, M., Guberek, T., and Ball, P. (2010). Measuring Elusive Populations with Bayesian Model Averaging for Multiple Systems Estimation: A Case Study on Lethal Violations in Casanare, 1998-2007. *Statistics, Politics, and Policy*, 1:1–26.

Lum, K., Price, M. E., and Banks, D. (2013). Applications of Multiple Systems Estimation in Human Rights Research. *The American Statistician*, 67:191–200.

Madigan, D. and York, J. C. (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika*, 84:19–31.

Manrique-Vallier, D., Price, M. E., and Gohdes, A. (2013). Multiple Systems Estimation Techniques for Estimating Casualties in Armed Conflict. In Seybolt, T. B., Aronson, J. D., and Fischhoff, B., editors, *Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict*. Oxford University Press.

Mitchell, S., Ozonoff, A., Zaslavsky, A. M., Hedt-Gauthier, B., Lum, K., and Coull, B. A. (2013). A Comparison of Marginal and Conditional Models for Capture–Recapture Data with Application to Human Rights Violations Data. *Biometrics*, 69:1022–1032.

Newcombe, H. B., Kennedy, J. M., Axford, S., and James, A. P. (1959). Automatic linkage of vital records. *Science*, 130(3381):954–959.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). *Statistical Inference from Capture Data on Closed Animal Populations*. The Wildlife Society, Wildlife Monographs No. 62, Washington, D.C.

Peterson, C. (1894). The Yearly Immigration of Young Plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station*, 6:1–48.

Price, M., Gohdes, A., and Ball, P. (2014). Updated Statistical Analysis of Documentation of Killings in the Syrian Arab Republic.

Price, M., Klingner, J., and Ball, P. (2013a). Preliminary Statistical Analysis of Documentation of Killings in the Syrian Arab Republic.

Price, M., Klingner, J., Qtiesh, A., and Ball, P. (2013b). Full Updated Statistical Analysis of Documentation of Killings in the Syrian Arab Republic.

Rajaraman, A. and Ullman, J. D. (2011). *Mining of Massive Datasets*. Cambridge UP, London.

Sekar, C. C. and Deming, E. W. (1949). On a Method of Estimating Birth and Death Rates and the Extent of Registration. *Journal of the American Statistical Association*, 44(245):101–115.

Silva, R. and Ball, P. (2008). The Demography of Conflict-Related Mortality in Timor-Leste (1974–1999): Empirical Quantitative Measurement of Civilian Killings, Disappearances & Famine-Related Deaths. In Asher, J., Banks, D., and Scheuren, F., editors, *Statistical Methods for Human Rights*, chapter 6, page 117. Springer, New York.

Winkler, W. E. (2006). Overview of Record Linkage and Current Research Directions. Technical Report RRS2006/02, Statistical Research Division, U.S. Census Bureau.

Zwierzchowski, J. and Tabeau, E. (2010). The 1992-95 War in Bosnia and Herzegovina: Census-Based Multiple System Estimation of Casualties' Undercount. In *Conference Paper for the International Research Workshop on 'The Global Costs of Conflict' The Households in Conflict Network (HiCN) and The German Institute for Economic Research (DIW Berlin).*

About HRDAG

The Human Rights Data Analysis Group is a non-profit, non-partisan organization[12] that applies scientific methods to the analysis of human rights violations around the world. This work began in 1991 when Patrick Ball began developing databases for human rights groups in El Salvador. HRDAG grew at the American Association for the Advancement of Science from 1994–2003, and at the Benetech Initiative from 2003–2013. In February 2013, HRDAG became an independent organization based in San Francisco, California; contact details and more information are available on HRDAG's website (`https://hrdag.org`) and Facebook page.

HRDAG is staffed by applied and mathematical statisticians, computer scientists, demographers, and social scientists. HRDAG supports the protections established in the Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights, and other international human rights treaties and instruments. HRDAG scientists provide unbiased, scientific results to human rights advocates to clarify human rights violence.

This work was supported by funding from the Economic and Social Research Council, the John D. and Catherine T. MacArthur Foundation, the Oak Foundation, Open Society Foundations, the Sigrid Rausing Trust, an anonymous U.S.-based private foundation, and individual donors.

The materials contained herein represent the opinions of the authors and editors and should not be construed to be the view of any of HRDAG's constituent projects, the HRDAG Board of Advisers, or the donors to HRDAG.

[12] Formally, HRDAG is a fiscally sponsored project of Community Partners.