Human Rights Data Analysis Group hrdag.org

Using science and data to advance human rights and justice

Deaths in custody during the armed conflict in Syria, 2011–2023

Maria Gargiulo* Tarak Shah† Megan Price‡

December 10, 2024

1 Executive summary

A key question of interest for the United Nations Commission of Inquiry on the Syrian Arab Republic is how many victims of the ongoing conflict were killed while in custody? Through our long collaboration with both the UN and multiple Syrian documentation groups, our team of data scientists at the Human Rights Data Analysis Group (HRDAG) have access to documented records of victims killed, under a variety of circumstances, in the Syrian Arab Republic between 2011 and 2023. This report is based on records collected by eight sources documenting deaths in the ongoing armed conflict in Syria. These sources are described in further detail in Section 2.

For the period 1 March 2011 through 31 December 2023, these sources documented a total of more than 21,300[1] unique, identifiable victims as being killed in custody. These documented killings, however, are an undercount of the total population of victims who were killed in custody during this period. The source of this undercounting is two-fold. First, not all conflict victims have been documented by the eight sources included in this analysis. Second, even among those documented, nearly 16% of records contained insufficient contextual information to determine whether the death occurred in custody. We use statistical and machine learning approaches to address these two sources of undercounting and to estimate the size of the total victim population, including those individuals who were not documented with sufficient contextual information or whose deaths were entirely undocumented.

*Statistician, Human Rights Data Analysis Group
†Data Scientist, Human Rights Data Analysis Group
‡Executive Director, Human Rights Data Analysis Group

[1]Following current recommendations, as described by Andrew Gelman (among others), we report results rounded to statistically meaningful precision.

From the more than 21,300 observed deaths in custody we estimate an overall total of **34,000** victims killed in custody within a 95% uncertainty interval between **(32,000, 37,000)**.[2] This estimate includes both documented and *undocumented* victims. As described in more detail in Section 4, these killings occurred most frequently early in the conflict, with an estimated peak in 2013, and a second notable spike in deaths in custody in 2018.

It is important to note that for this analysis we considered all groups who might be holding a victim in custody, not just State actors. We estimate that **80%**[3] of deaths in custody are attributed to State actors as the group accused of causing the death based on information recorded by the documentation groups. We relied on narrative descriptions of the circumstances of the victim's death or other notes recorded by the documentation groups to classify a death as occurring in custody or not. More details about this process are provided in Section 3.3.

The remainder of this report is organized as follows. Section 2 provides an overview of the eight data sources used in the analysis. Section 3 provides an overview of the data processing and the statistical methods used to link records across the eight input data sources, to statistically impute missing data within documented records, and to calculate estimates of the number of victims who have been killed while in custody, including those who were not documented by any data source. Section 4 contains the results of the analysis, including more information about temporal trends of killings in custody. Finally, Section 5 concludes.

2 Data sources and documentation

This work builds on several historical projects (Price, Klingner, and Ball 2013; Price, Gohdes, and Ball 2014, 2016; Gargiulo, Shah, and Price 2020; Shah et al. 2021). For the analyses presented here, we rely on records from eight sources:

- Center for Statistics and Research - Syria (CSR-SY)
- Damascus Center for Human Rights Studies (DCHRS)
- Office of the High Commissioner for Human Rights (OHCHR)
- Syrian Government (GoSY)
- Syrian Network for Human Rights (SNHR)
- Syrian Observatory for Human Rights (SOHR)
- Syria Shuhada website (SS)
- Violations Documentation Center (VDC)

We encourage all the groups to keep records in whatever way serves their internal processes and goals, and each of the data sources share records with us in different formats and organized in different ways. Most groups share excel files with us, but some share a combination

[2]This 95% uncertainty interval can be interpreted in the following way: "Given the observed data and assuming that the model is correct, there is a 95% chance that the true number of victims is between 32,000 and 37,000."

[3]With a 95% uncertainty interval of (76%, 84%).

of excel files and word documents. Groups may share a single excel file that includes all records of victims, others may share multiple excel files organized by date, geographic location, type of victim, or group accused of causing the death. Some excel files are single worksheets, others include records across multiple tabs. Again, these tabs may be organized by date, geographic region, or other information about the victim. Most records shared with us are predominantly in Arabic, but some contain a combination of English and Arabic. The groups also collect records covering different periods of the conflict, and not every group was actively documenting information about victims for the full period under analysis here.

This means that one of the first steps for our analysis is to standardize all the different data formats and organizational structures so that we can integrate the records of individual victims in a way that is internally consistent and coherent. This step is called data preprocessing and is described in Section 3.1.

A victim may be reported more than once, to the same documentation group, or to multiple documentation groups. We integrate all the information collected by all the groups, linking all the records that refer to the same victim. These multiple records are then combined into a single record with the most complete information available for each victim, generating a list of all uniquely identified victims. This set of records is referred to as an enumeration.

It is important to note that this enumeration, though informative, is *not* the complete number of deaths that occurred in custody during this time. As described in subsequent sections, though this enumeration is an important first step, we need to use statistical models to estimate how many victims are missing from this count: those whose deaths we do not know occurred in custody and those whose deaths have not been recorded at all.

Not all of the records shared by the sources are identifiable. For our purposes, a record is considered identifiable if it contains a full name, and month, year, and governorate of death. Some records are missing some of this content entirely, other records appear to have values for one or more of these columns, but further examination indicates they are not identifiable. For example, some records may contain name fields that describe that victim's relationship to someone else (such as "wife of" or "brother of") but without naming the victim. These records are also treated as unidentifiable. They represent victims who deserve to be acknowledged, however, they cannot be included when integrating information across groups because it is impossible to determine if the records with partial information refer to killings also described by other records. That is, anonymous or unidentifiable records cannot be matched or de-deduplicated. Records with partial information provide hints about the existence of killings which have not been fully documented and emphasize the need for the statistical estimation described in Section 3.5.

Additionally, some of the records describe deaths that occurred outside of the Syrian Arab Republic. These analyses only consider identifiable victims recorded between 1 March 2011 and 31 December 2023 within one of the fourteen governorates of the Syrian Arab Republic.

The analyses presented here are also based on what is documented about these victims and the circumstances of their deaths. As described in other sections, in some cases, it is clear whether a death occurred in custody or not. But in other cases this information is missing entirely or is inconclusive. As described in more detail in Section 3.3, the results presented

here may be in some sense a narrow representation of deaths in custody. In addition to victims for whom we cannot determine if their death occurred in custody or not, we are also aware that many of these documentation groups maintain separate lists of victims who have been arrested or are missing, but whose deaths have not been confirmed.[4] These victims are not included in the current analysis.

Finally, of course, these records cannot include victims who have not been documented by any of these sources. Victims whose stories have not yet been told or who did not leave behind any witnesses to tell their story. Again, this emphasizes the need for the statistical estimation described in Section 3.5.

3 Methods

The statistical estimates presented here are the end result of a many-step process. This work can be thought of in four parts: accessing and preparing records for analysis (data processing or pre-processing), identifying multiple records that refer to the same victim (record linkage), estimating missing fields from observed records of victims (multiple imputation), and estimating *undocumented* victims (multiple systems estimation).

3.1 Data pre-processing

The first step in our work is to take all the records shared with us, in whatever format and organizational scheme they arrive, and develop standard column headings and information types (date formats, etc.), so that all the records from all the sources can be combined and reviewed in a coherent way. This involves a variety of data processing and cleaning steps; examples include converting all date formats to YYYY-MM-DD, standardizing recorded sex values as "M", "F", or "unknown", standardizing age group values as "child", "adult", or "unknown", and in the cases of categorical variables like sex or age group, identifying any extraneous values or information that may be contained in those columns.

During this step we also translate or transliterate the content of the records so they can be reviewed in English. For categorical variables with finite possible values (such as governorate) we have worked with Arabic speakers to develop a dictionary that translates these values. For other text that may take on a wider variety of values (such as names, or descriptions of the circumstances of a death) we use the Google Translate API to transliterate the original Arabic content. We are aware that there are limitations to this approach and that transliterations may not capture nuances in the language, but have found that in most cases key words (such as "bomb" or "detention") are adequately rendered for these analyses. More nuanced review of original Arabic content is a necessary next step for these analyses.

Importantly, each of these steps is carried out via code, or software, *not* manual manipulations to spreadsheets. This approach is part of our larger principled data processing

[4]See, for example, The Syrian Network for Human Rights (2020) on enforced disappearance.

philosophy,[5] which ensures that we can trace back any modifications we make from the original data content to the final analysis-ready data set.

3.2 Record linkage

Once all the records have been processed and standardized, they are combined across all sources into one set of what we refer to as *pooled records*. These records still contain many records that refer to the same victim. The next step is what is formally referred to as record linkage, but may also be considered de-duplication or data integration.

The specific process we use is called semi-supervised record linkage. Christen (2012) provides an overview of this method. This is a well-established method in computer science and involves an iterative combination of human review and computer modeling. We begin by grouping records that have some information in common: victims killed during the same month, or in the same governorate, or with phonetically similar names. These groups of records are referred to as "blocks" and randomly selected records from these blocks are manually reviewed to identify records that refer to the same individual. Machine learning models are then used to model the decisions the humans made in order to assign a likelihood (called a score) that any two records refer to the same individual. These scores are then used to cluster records into groups that represent the information available about a single individual across all the data sources. This approach also makes it possible to identify and link records that are likely to refer to the same person, even if the content of the records is not completely identical. For example, if a date of death varies by a small number of days or if a victim's name is reported slightly differently across sources. More technical detail about our record linkage approach is included in Appendix A.2

3.3 Classifying deaths in custody

Descriptions of the cause or circumstances of a victim's death were used to determine if that death occurred while in custody. For this report we considered all groups who might be holding a victim in custody, not just State actors. Records that described being held, field executed, tortured, arrested, detained, or kidnapped were all classified as a death that occurred in custody.

Certain keywords were considered to indicate a killing that did not occur in custody. These included causes of death that mentioned: bombing, bombardment, explosion, shell, cluster, airstrikes, artillery, clashes, Russian aviation, Russian airstrike, or warplane(s). Records containing these words were automatically labeled as not in custody. Other descriptions of the cause or circumstances of death were reviewed manually and labeled as in custody, not custody, or unknown (insufficient information to conclude one way or the other). Examples of content for each label are below.

[5] See, for example, Patrick Ball's blog post The Task is a Quantum of Workflow.

Examples of descriptions coded as not killed in custody:

- "Due to shelling"
- "Long-range missile"
- "A result of air raids on the city"

Examples of descriptions coded as killed in custody:

- "Under torture in detention system"
- "Executed in the military security branch in Deir al-Zour"
- "Executed on the ground in Sednaya Prison"
- "Martyred under torture in Branch 215"

Examples of descriptions coded as unknown:

- "Army shot dead"
- "Killed by Army"
- "At the hands of the Regime"

These labels were applied to each individual record, and sometimes these labels had to be reconciled during record linkage when multiple records were determined to refer to the same individual (during the clustering step described previously). Specifically, if records contained contradictory labels (one or more records labeled in custody and one or more labeled not in custody) then that victim's custody status was categorized as unknown.

This approach could be considered a narrow definition, since some of the descriptions currently categorized as unknown may in fact describe a victim who was killed while in custody. These unknown categories are one kind of missing data which needs to be addressed, as described in Section 3.4.

Importantly, this is also a narrower approach to potentially contradictory records than previous analyses we have conducted of deaths in custody. In a previous similar analysis (Price, Gohdes, and Ball 2016) we used a more expansive definition of custody, and considered any cluster of records with at least some labeled as custody to indicate the victim was killed in custody *unless* the majority of the contributing records were labeled as not custody.

Another important difference in this updated analysis is our inclusion of additional information from what the documentation groups typically call the "notes" field. This field contains any additional information that recorders chose to document, and includes phrases or sentences that may describe attributes of the victim (for example, their family structure or profession), but in some cases includes clear indications that a death occurred in custody (for example, "...among the leaked photos of Caesar"). Because this unstructured text field contains such a diverse set of information, we did not have the capacity to include it in our previous analysis, but we were able to transliterate and incorporate these additional details in the results presented here.

After deduplicating the database and identifying custody deaths using the procedure described above, the database contains information on more than 21,300 deaths documented as occurring in custody. Of these 21,300 records, slightly more than 12,900 records have state actors documented as the group accused of causing the death and slightly fewer than 1,900 records document another non-state actor as the group accused of causing the death. However, nearly 6,500 records of documented deaths in custody are missing information about the group accused of causing the death. These counts include only those records that were *documented* with sufficient information to determine that they occurred in custody. As will be discussed further in Section 3.4, there are many records missing information about whether the death occurred in custody, the group accused of causing the death, or both fields. Taking these limitations into consideration, and considering only records where the death was known to have occurred in custody *and* where the alleged perpetrator was known, 87% of records document state actors as the group accused of causing the death and 13% of records document another non-state actor as the group accused of causing the death. It is important to note that these proportions are not identical to the *estimated* proportions reported in Sections 1 and 4 since those are based on estimates that account for multiple types of missing data. The proportions reported here only consider documented records with a clear description of a death in custody and a known group accused of causing the death.

3.4 Imputation of missing fields

As described in preceding sections, there are two kinds of missing data that need to be considered and accounted for in these analyses. The first is specific pieces of information that are missing about a documented, identifiable victim, even after integrating all available information about that individual. For example, we may know the name, date and location of a victim's death, but not their age group or custody status. This kind of missing data can be addressed with a method called imputation.

The second kind of missing data are victims who are not recorded by any of these sources at all. Victims whose names we do not know, whose stories have not yet been told. It is possible to estimate the number of victims who have not been documented by any of these eight sources using multiple systems estimation (MSE), which is discussed in greater detail in Section 3.5.

Imputation is a statistical modeling approach used to fill in missing data from observed records using the non-missing information in the record and information from similar records that are not missing a particular field. Rather than filling in the missing values one single time, we create 25 distinct versions of the data—called replicates—with all of the missing fields imputed probabilistically, meaning that each replicate is slightly different from the others. For example, if a particular record is missing age group information, it may be imputed as adult in one replicate and minor in another. This difference in the imputed values represents the uncertainty inherent in the imputation model: since data about this field was not recorded for this victim, we cannot be certain what the correct answer should be. Rather than hiding this uncertainty by imputing the value a single time, we repeat this process 25 times and then construct population estimates using all 25 replicates so that the

imputation error can be propagated into the final uncertainty intervals. When the model has high confidence that a particular missing field value should be imputed in a particular way, we will observe more consistency in the imputed value across the 25 replicates. Conversely, if the model is highly uncertain about how a particular value should be imputed, there will be more variability in the imputed value across the replicates. This process of imputing missing fields multiple times and pooling the results is called multiple imputation (MI). We create the 25 replicate data files using multivariate imputation by chained equations with the predictive mean matching algorithm via the `mice` package for `R` (van Buuren and Groothuis-Oudshoorn 2011).[6][7]

To improve the predictive power of the imputation model, we also make use of information contained in the unstructured text fields present in many of the data files we received. These are the notes and details fields, which may include phrases or sentences describing the circumstances of the death. This often includes valuable information about the victim or the group accused of causing the death, but it takes time to process systematically. We do this by generating "support vectors" using natural language processing techniques that allow us to create summary measures of the unstructured text data that are correlated with each of the fields we are trying to impute. These variables aid in identifying similar records that are not missing information on a particular field that may be useful for imputing missing values in that field.[8] Amado et al. (2022) use the same approach as part of the imputation procedure in the context of the armed conflict in Colombia.

For this analysis, we used MI to fill in missing values for sex, age group, cause of death, civilian status, custody status, and the group accused of causing the death. We also used MI to fill in records with contradictory information in the civilian status and custody status fields.

Table 1: Percentage of records with missing information

Variable	Percent missing
Group accused	24
Civilian status	22
Custody status	16
Age category	15
Cause of death	14
Sex	8

[6]The classic advice for MI is to use 3–5 replicates. However, to have very little additional variation due to the replicate sampling, we spent extra computational resources to create and analyze 25 replicates. This exceeds the 20 replicates recommended in more recent literature. Increasing the number of replicates has diminishing returns due to the inverse square root relationship with variation. See Section 2.8 of van Buuren (2018) for more information.

[7]See Chapter 3.4 of van Buuren (2018) for an overview of the predictive mean matching algorithm.

[8]In particular, we use a long short-term memory neural network on the entire contents of the records as we received them from the documentation groups (prior to data cleaning), but use standard pre-processing procedures, such as lemmatizing text and removing stop words, before fitting the models and calculating the support vectors.

Table 1 shows the percentage of records for each field in the deduplicated database that is missing information. Here we observe that almost a quarter of records are missing information about the group accused of causing the death, whereas a much smaller percentage of records are missing information about the sex of the victim. Imputation will be used to address all of the varying amounts of missing data across all of these fields.

3.5 Multiple systems estimation (MSE)

Multiple systems estimation (MSE) is a broad class of statistical tools designed specifically to estimate the size of a hard to reach population based on multiple overlapping, although not necessarily representative, samples. Originally proposed in 1783 to study the size of the population of France (Amorós 2014), this class of tools has been developed and expanded since then and across a wide variety of fields, including ecology (Otis et al. 1978), demography (Seber 1965), public health (Wittes and Sidel 1968; Yip et al. 1995), and human rights research (Ball and Price 2019). This family of methods, applied to conflict casualty research, allows us to estimate the *total* number of victims and construct an uncertainty interval around the estimate. Detailed examples of the use of this method to study conflict casualties, and particularly the verification of this approach, are provided in Appendix A.1.

Once we have completed record linkage, and identified multiple records that refer to the same victim, we can summarize how many victims were only reported by one source, by two sources, by every possible combination of two sources, and so on.

A way to visualize these overlap patterns is shown in Figure 1 which shows the overlap patterns across the groups reporting deaths in custody in 2014. A black colored-in circle indicates records documented by that source; black lines connecting circles indicate records covered by *all* of the black colored-in and connected sources. For example, the leftmost bar in Figure 1 is above filled in and connected dots for six sources: VDC, DCHRS, CSR-SY, SOHR, SNHR, and SS. This means the most frequent documentation pattern in 2014 was when six sources all recorded victims in common. However, Figure 1 indicates a number of other documentation patterns, including substantial unique records contributed by only one source (the single black circle with no connecting lines corresponding to CSR-SY under the bar third from the left and the single black circle with no connecting lines corresponding to SNHR under the bar fourth from the left). Note that in 2014, two sources—GoSY and OHCHR—did not document any deaths, hence why all of the circles are grey for all observed documentation patterns. This is because not all sources cover all years for the period of analysis.

Each of these patterns of documentation is important because statistical models are fit to these observed overlap patterns to estimate the "unobserved" pattern: how many victims are not reported to any source?

The intuition behind MSE is based on these patterns. Consider the following analogy: say you have been asked to determine which of two dark rooms is larger.[9] To do this, you were given rubber balls with special properties. The balls do not make any noise when they hit

[9]This analogy is used in many of HRDAG's reports using MSE.

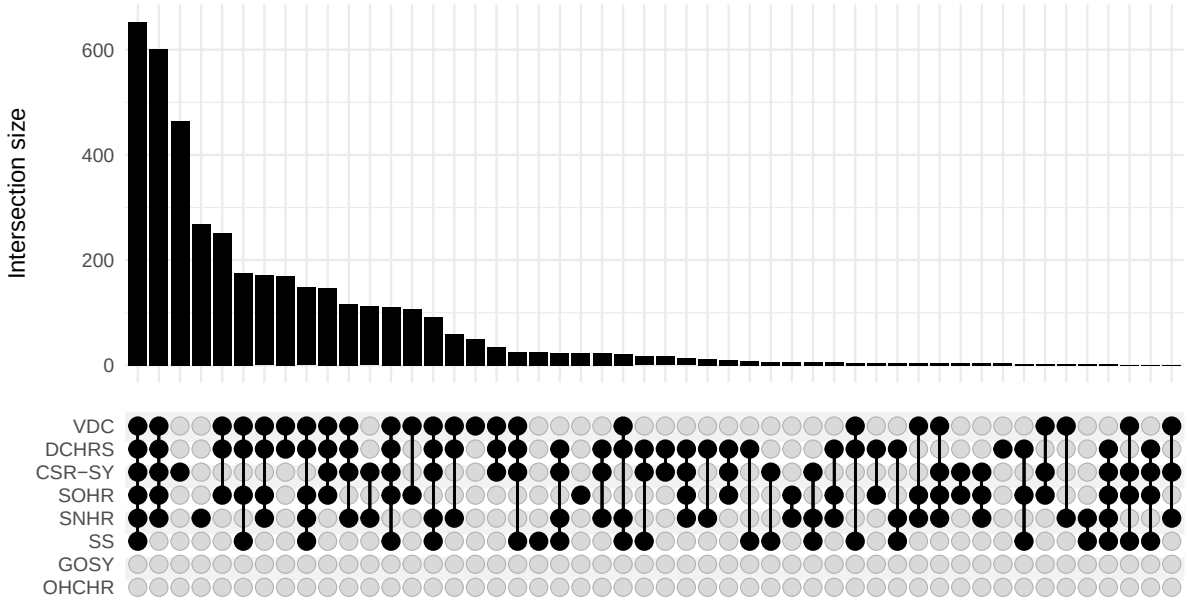

Figure 1: Documentation patterns across eight sources for 2014

the walls, floor, or ceiling of the room, but when two or more balls collide, they make a clicking noise. You pick up the balls and throw them in the first of the dark rooms and hear several clicks—*click*, *click*, *click*. You pick up the balls, go to the second room, and toss the balls again with the same force. This time you hear only one—*click*. Which room is larger? The second room must be larger; the balls had more space to spread out, so they collided with each other less frequently than they did in the first room and therefore made fewer clicks.

Translating this analogy back to the language of MSE, the rubber balls represent our data sources and the dark rooms represent the unknown size of the victim population we are trying to estimate. In this language, a "collision" occurs when two or more sources document the same victim. The more often the same victims are reported across multiple sources, the closer the total number of victims will be to the observed total, whereas fewer repeat reports implies a larger number of unobserved victims.

The results presented in this report are based on a specific MSE modeling approach called Bayesian Non-Parametric Latent-Class Capture-Recapture (Manrique-Vallier 2016).[10] This approach is well-suited to handle the number of sources we have available and the varying time periods covered by each source. We apply LCMCR to each of the 25 imputed replicate data sets and combine the results using the laws of total expectation and variance, such that the final uncertainty intervals include both the uncertainty from imputation and the

[10]We use the software package LCMCR for R to construct the estimates.

uncertainty from MSE.[11] By combining these approaches, the results presented here should account for the different kinds of missing or partially observable information described in preceding sections. In other words, the results presented here account for documented victims as well as those who are partially documented, and those who were completely undocumented.

4 Results

As described in Section 1 for the period 1 March 2011 through 31 December 2023, these sources documented a total of more than 21,300 unique, identifiable victims as being killed in custody. Applying the methods described in the preceding sections to these observed records, we estimate 34,000 victims killed in custody within a 95% uncertainty interval between (32,000, 37,000). This estimate includes all groups who might be holding a victim in custody, not just State actors. We estimate that 80%[12] of deaths in custody are attributed to State actors as the group accused of causing the death. In total, we estimate that 28,000 deaths in custody with a 95% uncertainty interval between (26,000, 31,000) as being attributed to State actors and 7,000 with a 95% uncertainty interval between (6,000, 9,000) as being attributed to other groups.

It is important to note that due to the way that the estimates are constructed, it is not correct to simply add the results together. The same applies for the lower and upper bounds of the uncertainty intervals. As described in Section 3.5, the laws of total expectation and variance must be applied to the posterior distribution of the sum to correctly sum estimates across categories and calculate the appropriate uncertainty interval.

Figure 2 shows the total number of observed and estimated victims by year. The dotted line at the bottom of the graph is the total number of uniquely identified victims killed in custody, the solid line is the estimated total, and the shading around the solid line indicates the 95% uncertainty interval. There are a few notable takeaways from Figure 2. First, it appears that deaths in custody occurred more frequently in the early part of the conflict, peaking in 2013. There is another notable spike, in both documented and estimated total numbers of victims, in 2018. The distance between the dotted line and the bottom of the shaded region also indicates that there is a statistically significant number of undocumented victims: those who have been killed in custody but whose deaths have not been recorded by any of these eight sources. For the period 2012–2014 and the spike in 2018, the number of undocumented victims is substantial (as indicated by the distance between the dotted line and the bottom of the shaded region.)

[11] See Section 18.2 of Gelman et al. (2014) for more information about combining estimation results constructed using data imputed using MI. These rules generally assume that the outcome parameter, in this case, the estimated number of victims, is distributed according to a normal distribution. However, the posterior distribution of the probable number of victims using MSE models is often right skewed, approximating a log-normal distribution. There is usually more uncertainty in the upper limit than there is in the lower limit because the lower limit is bounded by the number of documented victims. As a result, we log transform our estimates to ensure that the uncertainty interval is properly calculated and then exponentiate the results to have them back in the original scale.

[12] With a 95% uncertainty interval of (76%, 84%).

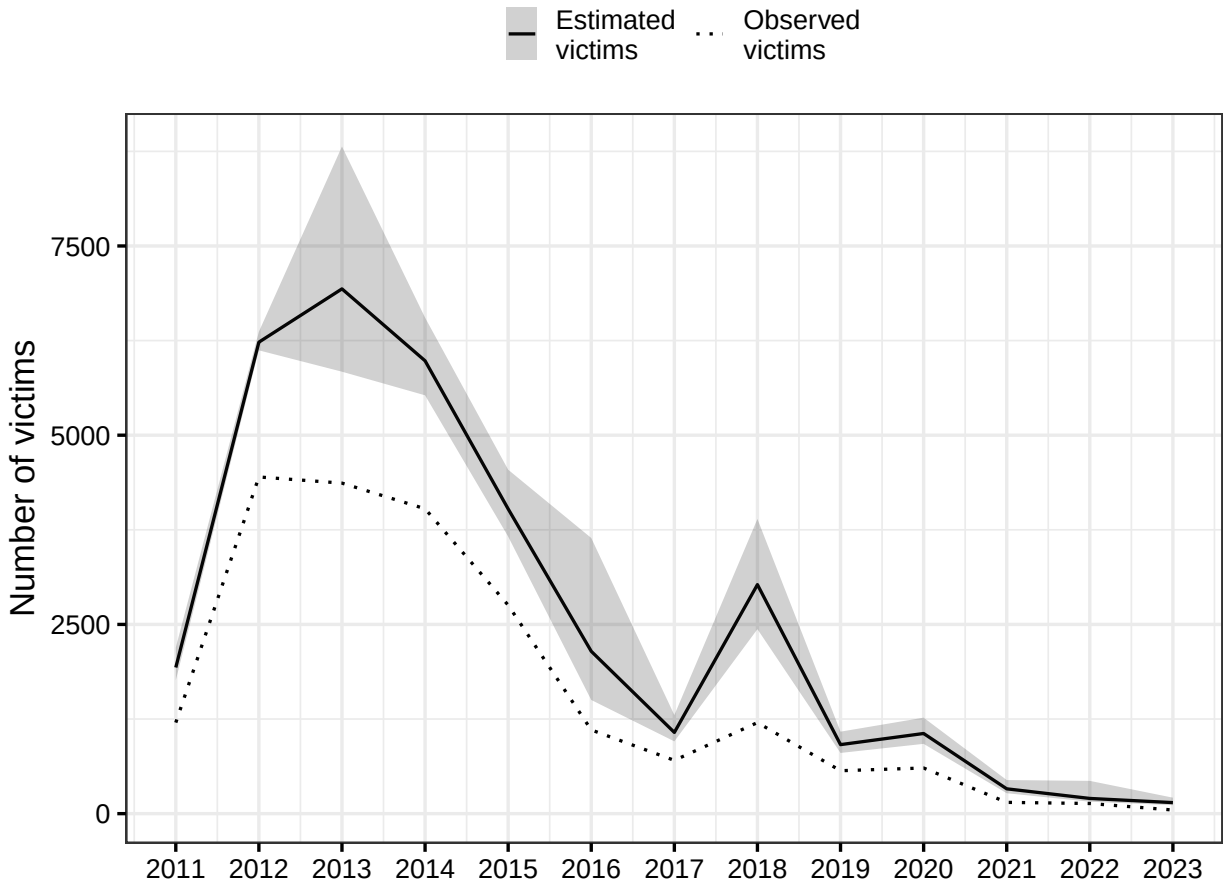

Figure 2: Observed and estimated deaths in custody over time

Further investigation is needed to understand why deaths in custody follow these patterns over time, but this analysis should provide a starting point.

5 Conclusion

This analysis is an update and expansion of previous analyses, but in many ways is still just the beginning of investigative questions that could be raised about deaths in custody during the ongoing conflict in Syria. The results presented here consider patterns of death in custody over time (by year) and one comparison of groups accused of being responsible for the killing. It is possible that for some early years in the conflict, and potentially 2018, a more granular time trend, by month, could reveal other informative patterns. Qualitative research into the circumstances of the conflict during those years might help us to further interpret these estimated trends over time.

Additional data sources may also deepen our understanding of this and other aspects of the conflict. The current analysis is based on records of confirmed deaths, but we know the groups who contributed information to this project, as well as other documentation groups, collect additional information about victims who are reported as missing or disappeared. Matching those lists of identifiable victims to existing lists of confirmed deaths has the potential to at least provide closure to communities who may still be waiting to know what happened to their loved ones.

A Appendix

A.1 How do we know MSE works?

In addition to over a century of methodological development across a wide variety of fields, there are multiple examples of MSE used specifically in the analysis of conflict casualties. These include a direct confirmation of MSE results from both alternative estimation methods and a complete census of victims. There are also multiple examples of MSE results being considered as evidence in courtrooms and truth commissions.

In Kosovo, two epidemiological surveys estimated total mortality consistent with previously calculated MSE estimates (Spiegel and Salama 2000; Iacopino 1999). More importantly, after more than a decade's work, a Serbian and Kosovar NGO, the Humanitarian Law Centre, conducted a complete enumeration of the victims of the conflict, and their count is consistent with the MSE estimate, at an aggregate level and stratified by time and space (Krüger and Ball 2014).

Historically, scientists at HRDAG first used multiple sources to estimate total deaths in conflict in Guatemala. Following Guatemala's internal armed conflict, the Commission for Historical Clarification (CEH) sought to estimate both the overall number of individuals who had been killed during the conflict and the relative number of victims who were members of the indigenous Mayan population. Three sources collected lists of victims: the CEH, the International Center for Human Rights Research in Guatemala, and the Catholic Church's Recovery of Historical Memory project. The MSE results provided the basis for the CEH's conclusion "...that agents of the State of Guatemala, within the framework of counterinsurgency operations carried out between 1981 and 1983, committed acts of genocide against groups of Mayan people which lived in the four regions analysed" (Commission 1999).

Nearly fifteen years later, HRDAG scientists updated those analyses with new methods and a fourth data source (from the National Reparation Program). The updated approach confirmed the earlier findings: indigenous people in the Ixil region in the early 1980s were approximately eight times more likely to be killed by the Army than non-indigenous people in the same time and region. These analyses were presented in 2013 in the trial of former *de facto* President General José Efraín Ríos Montt for acts of genocide, and contributed to the judges' determination of a guilty verdict (Ball and Price 2018).

MSE methods have also been used in a few different investigations related to conflicts among member states of the former Yugoslavia. Zwierzchowski and Tabeau (2010) used MSE to estimate the total number of war-related deaths in Bosnia and Herzegovina from April 1992 to December 1995. They relied on official death notifications, military records, lists of missing persons, and exhumation records. Brunborg, Lyngstad, and Urdal (2003) used data collected by International Committee for Red Cross and Physicians for Human Rights about missing persons to conduct two systems MSE to estimate the total number of people killed in Srebrenica. Brunborg, Lyngstad, and Urdal (2003) concluded that more than a third of all Muslim men living in Srebrenica before the war were killed. This statistical pattern is consistent with acts of genocide, and in the ICTY trial against General Radislav Krstic, the court agreed.

Other publications from HRDAG using MSE methods include estimates for the Peruvian Truth Commission (Ball et al. 2003), of the total number of killings and disappearances in Casanare (Lum et al. 2010; Guberek et al. 2010) and of social movement leaders killed in Colombia (Rozo Ángel and Ball 2019), of the total number of Syrians killed in custody (Price, Gohdes, and Ball 2016), of the total number of people disappeared after surrendering during the last three days of the Sri Lankan civil war (Ball and Harrison 2018), of the total number of civilians killed and disappeared during the civil war in El Salvador (Hoover Green and Ball 2019), of drug-related killings in The Philippines (Ball et al. 2019), and estimates of homicides, enforced disappearances, kidnappings, and the recruitment of child soldiers for the Colombian Truth Commission (Amado et al. 2022).

A.2 Record Linkage - Technical details

In the first step of the record linkage process, a human reviewer starts with "blocks" of records, and then sorts those 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") and pairs of records that do not refer to the same person ("negative pairs" or "non-matches"). 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 train a pairwise classification model. In total, over 77 million pairs were generated from human review.

Rather than estimate the classification model on the full combination of all possible pairs (which would be over 500 billion pairs), we limit the consideration to only the pairs that have a reasonable chance of referring to the same individual victim. We start by considering 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 we maximize coverage of positive pairs. This process is referred to as "blocking". We learn the optimal set of blocking rules using a variation of the method proposed by Michelson and Knoblock (2006).[13] This approach generated a total of more than 77 million pairs, called "candidate pairs" (it is a coincidence that the number of pairs generated from human review and the number of candidate pairs are similar). This is the set of pairs that is considered in the remaining steps of the record linkage process.

Next, we generate a series of features that can be used to compare pairs of records. Features are any numerical summary that can describe the similarity, or difference, between a pair of records. Examples include the number of days between recorded dates of death, the phonetic similarity of their names (in English or Arabic), and the number of characters that would need to change to make names (in English or Arabic) identical. In total, we generate more than 70 features to compare pairs of records. After developing these features, we use the hand-labeled data, both positive and negative pairs, to train and validate a binary classifier using the gradient boosted trees algorithm (Chen and Guestrin 2016), which we implemented using

[13]For even more technical detail about blocking, see Patrick Ball's blog post on Adaptive Blocking. Tarak Shah's blog post on Greedy Blocking introduces the local search approach to blocking we used for this analysis.

the xgboost package for R. We used 70% of the hand-labeled data to train the classifier, and the remaining 30% of the hand-labeled data to evaluate model performance. The precision of the final model was 0.992 and the recall was 0.988. This yields an overall F-score of 0.99, indicating good model performance.

We then applied this trained model to the more than 77 million candidate pairs generated in the blocking step. Pairs with classification scores closer to 1 are likely to refer to the same victim, whereas pairs with classification scores closer to 0 are likely to refer to different victims. With the classification scores for all candidate pairs, we now need to determine which groups of records, called clusters, refer to the same individual victim. A cluster may consist of one, two, or more records.

Our clustering approach first partitions records into groups using transitive closure: we link all pairs that have even a small probability of being a match (here defined as a classification score greater than 0.4) into super-clusters called "connected components". Next, we need to separate each connected component into smaller clusters, each representing an individual victim, which maximize the similarity of the records contained within. We separate the connected components using a method called Hierarchical Agglomerative Clustering (HAC). 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 stricter than the threshold used in transitive closure.[14]

After applying transitive closure and HAC, we can identify which records refer to which unique victims. For clusters with more than one record, the information across all the records must be merged to form a single, unified record. After identifying multiple records that refer to the same victim, we sometimes find that those records contain contradictory information across some fields, including the field indicating whether the victim was killed in custody. In other cases, even after gathering all available information, that particular field value may be unknown. This also occurs with the custody field in some records. For records with either a contradictory set of values for custody status or missing custody status, we used a statistical method called multiple imputation to probabilistically assign a custody status. This method is described in detail in Section 3.4.

Acknowledgements

The authors are deeply indebted to Michelle Dukich for her meticulous data review and processing. We thank the documentation groups for sharing their records with us. And we thank Sir Bernard Silverman, Dr. Patrick Ball, Dr. Shira Mitchell, and Professor Nils Lid Hjort for reviewing an earlier draft of this report. Any errors or omissions are those of the authors.

[14] For even more technical detail about clustering, see Patrick Ball's blog post on Clustering.

About HRDAG

The Human Rights Data Analysis Group is a non-profit, non-partisan organization[15] 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 is available on HRDAG's website (https://hrdag.org).

HRDAG is composed of 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. The human rights movement is sometimes described as "speaking truth to power": HRDAG believes that statistics about violence need to be as true as possible, with the best possible data and science.

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

[15] Formally, HRDAG is a fiscally sponsored project of Community Partners, see https://communitypartners.org/.

References

Amado, Paula, William Acero, Camilo Argoty, Giovanny Babativa, Luz Karime Bernal, Alejandro Castro, María Juliana Durán, et al. 2022. "Informe Metodológico Del Proyecto Conjunto JEP-CEV-HRDAG de Integración de Datos y Estimación Estadística."

Amorós, Jaume. 2014. "Recapturing Laplace." *Significance*. https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2014.00754.x.

Ball, Patrick, Jana Asher, David Sulmont, and Daniel Manrique. 2003. "How Many Peruvians Have Died." *Washington, DC: American Association for the Advancement of Science*.

Ball, Patrick, Sheila Coronel, Mariel Padilla, and David Mora. 2019. "Drug-Related Killings in the Philippines." Human Rights Data Analysis Group.

Ball, Patrick, and Frances Harrison. 2018. "How Many People Disappeared on 17–19 May 2009 in Sri Lanka?" Human Rights Data Analysis Group.

Ball, Patrick, and Megan Price. 2018. "The Statistics of Genocide." *CHANCE* 31 (1): 38–45.

———. 2019. "Using Statistics to Assess Lethal Violence in Civil and Inter-State War." Journal Article. *Annual Review of Statistics and Its Application* 6 (Volume 6, 2019): 63–84. https://doi.org/https://doi.org/10.1146/annurev-statistics-030718-105222.

Brunborg, Helge, Torkild Hovde Lyngstad, and Henrik Urdal. 2003. "Accounting for Genocide: How Many Were Killed in Srebrenica?" *European Journal of Population/Revue Européenne de démographie* 19 (3): 229–48.

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–94.

Christen, Peter. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31164-2.

Commission, Historical Clarification. 1999. "Guatemala: Memory of Silence." *Guatemala City: Historical Clarification Commission*.

Gargiulo, Maria, Tarak Shah, and Megan Price. 2020. "Documented Identifiable Victims in the Syrian Conflict 2015-2017." Human Rights Data Analysis Group.

Gelman, Andrew, John B Carlin, Hal Steven Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2014. *Bayesian Data Analysis*. New York: CRC Press. http://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=1438153.

Guberek, Tamy, Daniel Guzmán, Megan Price, Kristian Lum, and Patrick Ball. 2010. "To Count the Uncounted: An Estimation of Lethal Violence in Casanare." Benetech.

Hoover Green, Amelia, and Patrick Ball. 2019. "Civilian killings and disappearances during civil war in El Salvador (1980–1992)." *Demographic Research* 41 (27): 781–814. https://doi.org/10.4054/DemRes.2019.41.27.

Iacopino, Vincent. 1999. *War Crimes in Kosovo: A Population-Based Assessment of Human Rights Violations Against Kosovar Albanians*. Physicians for Human Rights.

Krüger, Jule, and Patrick Ball. 2014. "Evaluation of the Database of the Kosovo Memory Book." *Human Rights Data Analysis Group*. https://hrdag.org/wp-content/uploads/2015/04/Evaluation_of_the_Database_KMB-2014.pdf.

Lum, Kristian, Megan Price, Tamy Guberek, and Patrick Ball. 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).

Manrique-Vallier, Daniel. 2016. "Bayesian Population Size Estimation Using Dirichlet Process Mixtures." *Biometrics* 72 (4): 1246–54.

Michelson, Matthew, and Craig A Knoblock. 2006. "Learning Blocking Schemes for Record Linkage." In *AAAI*, 6:440–45.

Otis, David L, Kenneth P Burnham, Gary C White, and David R Anderson. 1978. "Statistical Inference from Capture Data on Closed Animal Populations." *Wildlife Monographs*, no. 62: 3–135.

Price, Megan, Anita Gohdes, and Patrick Ball. 2014. "Updated Statistical Analysis of Documentation of Killings in the Syrian Arab Republic." Human Rights Data Analysis Group. https://hrdag.org/wp-content/uploads/2013/06/HRDAG-Updated-SY-report.pdf.

———. 2016. "Technical Memo for Amnesty International Report on Deaths in Detention." Human Rights Data Analysis Group. https://hrdag.org/wp-content/uploads/2016/07/HRDAG-AI-memo.pdf.

Price, Megan, Jeff Klingner, and Patrick Ball. 2013. "Full Updated Statistical Analysis of Documentation of Killings in the Syrian Arab Republic." Human Rights Data Analysis Group. https://hrdag.org/wp-content/uploads/2013/06/HRDAG-Updated-SY-report.pdf.

Rozo Ángel, Valentina, and Patrick Ball. 2019. "Killings of Social Movement Leaders in Colombia: An Estimation of the Total Population of Victims - Update 2018." Human Rights Data Analysis Group.

Seber, George AF. 1965. "A Note on the Multiple-Recapture Census." *Biometrika* 52 (1/2): 249–59.

Shah, Tarak, Maria Gargiulo, Anita Gohdes, and Megan Price. 2021. "Documented Identifiable Victims in the Syrian Conflict 2011-2020." Human Rights Data Analysis Group.

Spiegel, Paul B, and Peter Salama. 2000. "War and Mortality in Kosovo, 1998–99: An Epidemiological Testimony." *The Lancet* 355 (9222): 2204–9.

The Syrian Network for Human Rights. 2020. "The Ninth Annual Report on Enforced Disappearance in Syria on the International Day of the Victims of Enforced Disappearances; There Is No Political Solution Without the Disappeared." https://snhr.org/wp-content/pdf/english/The_Ninth_Annual_Report_on_Enforced_Disappearance_in_Syria_on_the_International_Day_of_the_Victims_of_Enforced_Disappearances_en.pdf.

van Buuren, Stef. 2018. *Flexible Imputation of Missing Data*. CRC press.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in r." *Journal of Statistical Software* 45: 1–67.

Wittes, Jane, and Victor W Sidel. 1968. "A Generalization of the Simple Capture-Recapture Model with Applications to Epidemiological Research." *Journal of Chronic Diseases* 21 (5): 287–301.

Yip, PSF, G Bruno, N Tajima, GAF Seber, ST Buckland, RM Cormack, N Unwin, et al. 1995. "Capture-Recapture and Multiple-Record Systems Estimation i: History and Theoretical Development." *American Journal of Epidemiology*.

Zwierzchowski, Jan, and Ewa Tabeau. 2010. "The 1992-95 War in Bosnia and Herzegovina:

Census-Based Multiple System Estimation of Casualties' Undercount." *Berlin: Households in Conflict Network and Institute for Economic Research*, 539.