Updated Statistical Analysis of Documentation of Killings in the Syrian Arab Republic

Megan Price, Anita Gohdes, and Patrick Ball*

Human Rights Data Analysis Group
everybody counts.

August 2014

Executive Summary

This report presents findings on documented conflict-related killings in the Syrian Arab Republic (Syria) between March 2011 and April 2014. This report is an update of work published in June 2013. The previous report presented an enumeration of reported killings between March 2011 and April 2013. The numbers presented in this report are based on the integration of information collected by multiple sources, including Syrian human rights groups and the Government of the Syrian Arab Republic. Although ongoing conflict conditions make the collection of accurate records of violent events difficult, different actors are continuing to gather information about killings through a variety of local sources and data collection methods.

The purpose of this report is to update previous quantitative assessments of the current documentation of violence in Syria. The previous report and this updated analysis both include three sources that cover the entire period under investigation: the Syrian Center for Statistics and Research, the Syrian Network for Human Rights, and the Violations Documentation Centre. The remaining two sources—the Syrian Government and the Syrian Observatory for Human Rights—cover a subset of that period. Records from the Syrian Government only cover the period from March 2011 through March 2012 and records from the Syrian Observatory for Human Rights only cover from March 2011 through April 2013. The records from the Syrian Center for Statistics and Research, the Syrian Network for Human Rights, and the Violations Documentation Centre reflect all further processing, verification, cross checking, and modification that these groups have conducted since the beginning of the conflict. Records from the Syrian Government and the Syrian Observatory for Human Rights included in this report are the same records provided in 2012 and 2013 and do not reflect any updates either of these groups may have made to their own databases.[1] The Office of the High Commissioner for Human Rights requested information from the Syrian Government in September 2013 and April 2014. No response was received at the time of publication. The Syrian Observatory for Human Rights declined to share an updated list in early 2014. Hereafter this collection of data will be referred to as the five sources on which this report is based.

The report considers documented killings that are fully identified by the name of the victim, as well as the date and location of death. Reported killings that are missing any of this information were excluded from this study. This report finds that when the fully identified records from four documentation groups plus the Syrian Government were combined and duplicates identified, these five databases identified 191 369 unique killings.[2] The integrated list of killings is called an enumeration.

We highlight four observations about patterns among the documented killings:

- *Sex of the Victims*: Of the 191 369 documented killings in this report, 85.1% are male, 9.3% are female, and 5.6% of records do not indicate the sex of the victim.

- *Age of the Victims*: Age is unknown for 83.8% of all records, which makes it impossible to draw conclusions about the overall distribution of the age of victims. However, the full enumeration does include 2165 records of victims 0-9 years old,

---

[1] See Section 1 and Appendix A for further details describing what is and is not included in the records analyzed in this report.

[2] All numbers in this report are formatted according to the International Standards Organization 31-0. The thousands separator is a small space and is used for numbers of more than four digits, and the decimal is marked by a point.

and 6638 records of victims 10-18 years old.

- *Location of Killing*: The three comprehensive non-governmental sources included in this report (the Syrian Center for Statistics and Research, the Syrian Network for Human Rights, and the Violations Documentation Centre) all record more killings in Rural Damascus than in other governorates.[3]

- *Combatant and Non-Combatant Status*: the status of the victims as combatants or non-combatants is unknown for all but a few records, and consequently, combatant status is not assessed in this report.

The enumeration is *not* the complete number of conflict-related killings in Syria. The enumeration may be a slight overcount of the number of the fully identified, reported killings, while at the same time the enumeration is likely undercounting the true total number of conflict-related killings that have occurred during this time period. This is because an unknown number of conflict-related killings are likely to have occurred without being documented by the sources used in this report. Therefore, the true total number of conflict-related killings must include both documented killings (those enumerated in this report) and *undocumented* killings, which must be estimated using statistical models.

The enumeration may be a slight overcount of identified reported killings due to two factors. First, the enumeration may include a small number of undetected duplicates among the unique killings, despite human efforts and computer modeling. Second, it may include records that are inaccurate in some sense, for example, records that describe deaths that were not conflict-related, or victims presumed dead who were later found to be alive.[4] Only a few examples of such records (fewer than 1000) have been shared with HRDAG so far. Both of these factors may slightly inflate the current enumeration.

However, based on experience in similar contexts, HRDAG believes that many more killings remain undocumented, and that statistical models could be used to estimate the probable number of undocumented killings.[5] The total 191 369 can be understood as a minimum bound of the number of killings between March 2011 and April 2014.

Report Organization

Section 1 provides a summary of documented killings in the Syrian Arab Republic between March 2011 and April 2014 and a brief description of the data sources. Sections 2 and 6 detail what is and is not included in these analyses and what can and cannot be concluded from them. Section 3 briefly describes how these five datasets were compared and integrated. A detailed analysis of how the datasets overlap with each other is presented in Section 4; the overlap analysis helps explain how the various data sources each capture distinct aspects of the total universe of killings. A comparative statistical analysis of all five datasets is presented in Section 5, including patterns of documented killings over time, as well as by geography, sex and age of the victims. Appendix A briefly describes each of the data sources and discusses concerns regarding potentially

---

[3]Records from the Syrian Government, which cover the period from March 2011 through March 2012, record the highest number of killings in Homs. The Syrian Observatory for Human Rights, with records from March 2011 through April 2013, records the highest number of killings in Rural Damascus.

[4]For more discussion of potentially inaccurate records, see Appendix A.3.

[5]HRDAG team members have analyzed documented killings and estimated undocumented killings in Guatemala, Kosovo, Perú, Timor-Leste, and Colombia. See the HRDAG publications page. For a discussion of the relevant methods, see HRDAG's page describing multiple systems estimation.

inaccurate records. Lastly, Appendices B and C provide technical and methodological detail.

1   Background

This report presents an analysis of killings that have been documented in the Syrian Arab Republic (Syria) between March 2011 and April 2014, based on five datasets. Based on a comparison of records from these five sources, HRDAG found a total of 191 369 unique records of documented killings. Importantly, this enumeration should not be inferred to include only civilian victims. The status of the victims as combatants or non-combatants is unknown for all but a few records, but both statuses are reported. Therefore, collectively the data sources include records of both combatants and non-combatants.

The analysis in this report updates two previous reports, released on 2 January 2013 and 13 June 2013. The most recent report, published on 13 June 2013, described 92 901 records of killings from March 2011 through April 2013.

It is important to note that the two previous reports integrated records from eight different sources. In the year since those publications, the conflict continues and these documentation efforts have evolved. Some organizations have stopped collecting data, while others have stopped sharing their data with this project.

This report presents results based on five sources (listed below). These sources were chosen as the basis for these analyses for two reasons. First, the four Syrian human rights groups (the Syrian Center for Statistics and Research, the Syrian Observatory for Human Rights, the Syrian Network for Human Rights, and the Violations Documentation Centre) have participated in a workshop on 27 and 28 February 2014 organized by OHCHR where OHCHR and HRDAG learned more about their data collection and verification methodology. Second, records from the Syrian Government are included since these are considered to be a valuable resource and to reflect a very different kind of knowledge about the conflict than that available to Syrian human rights groups.

For brevity, each list will be referred to by its acronym in the tables and figures throughout this report. It should be noted that each data collection organization determines their own methods for data gathering and verification. Further detail about each group is listed in Appendix A.1.

1. **GoSY**: the Syrian Government[6]

2. **CSR-SY**: the Syrian Center for Statistics and Research[7]

3. **SNHR**: the Syrian Network for Human Rights[8,9]

4. **SOHR**: the Syrian Observatory for Human Rights[10,11]

---

[6]Throughout the report, counts of records from GoSY are highlighted with an asterisk * as a reminder that GoSY records only cover the period March 2011 through March 2012.

[7]http://csr-sy.org/

[8]http://sn4hr.org/

[9]In previous reports we combined records from the Syrian Revolution General Council (SRGC) with SNHR. After speaking with the founder and head of SNHR, it is clear that this was incorrect and these two sources are no longer being combined. This report includes only records from SNHR and does not include any records from SRGC.

[10]www.syriahr.com, www.syriahr.net

[11]Throughout the report, counts of records from SOHR are highlighted with an asterisk * as a

5. **VDC**: the Violations Documentation Centre[12]

It should also be noted that three of these sources (CSR-SY, SNHR, and VDC) cover the entire period reported here (March 2011 through April 2014) whereas the remaining two sources cover a subset of that period. As summarized in Table 1, records from the Syrian Government only cover the period from March 2011 through March 2012 and SOHR records only cover from March 2011 through April 2013. Additionally, updated records from CSR-SY, SNHR, and VDC reflect all further processing, verification, cross-checking, and modification that these groups have conducted since the beginning of the conflict. Records from SOHR and GoSY included in this report are the same records provided in 2012 and 2013 and do not reflect any updates either of these groups may have made to their own databases.

Table 1: Time Period Covered by Each Source

Dataset	Period Covered
GoSY	March 2011–March 2012
CSR-SY	March 2011–April 2014
SNHR	March 2011–April 2014
SOHR	March 2011–April 2013
VDC	March 2011–April 2014

2 What These Analyses Do and Do Not Include

The first step in this analysis involves semi-automated examination of each individual record in each dataset in order to identify multiple records that refer to the same death. Sometimes these records occur within a single dataset (duplicate records) and other times they occur in multiple datasets (matched records). See Appendix C for a description of this process.

This comparison of records is only possible for records with sufficient identifying information, including the name of the victim, plus the date and location of death.[13] Each dataset considered in this study included a number of records which lacked this information. Table 2 lists the number of records from each dataset included in the analyses presented in this report (those with sufficient identifying information) and the number of records excluded from these analyses (those lacking sufficient identifying information). It is important to note that, as described above, records from two sources do not cover the entire period under study in this report. Therefore for GoSY and SOHR, the numbers presented in Table 2 do not represent the full set of records collected by these sources.

It is worth noting that none of the included counts in Table 2 match the total number of documented killings—191 369—because each dataset contains records that none of the other groups documented, duplicates within the dataset, as well as records that are common to two or more datasets. To be clear, neither anonymous nor duplicate record

---

reminder that SOHR provided records only through April 2013.

[12] http://www.vdc-sy.org/

[13] Ideally, records included an unambiguous governorate of death. In some cases location was inferred from other information included in the record. Complete details of this and other data processing can be found in Appendix B.

counts imply anything about the quality of any source's data collection efforts. These are natural and expected attributes of any data collection.

Table 2: Number of Records Included and Excluded in Analyses

Dataset	Identifiable	Unidentifiable
GoSY	2470	87
SOHR	45 402	4503
CSR-SY	75 149	13 840
SNHR	97 087	20 506
VDC	98 802	13 017
Total	318 910	51 953

Lastly, it is important to note that the records included from each source are specifically records of victims described as "martyrs." Some of these groups collect additional records of killings which describe other types of victims. Since these records were not available for the entire period under study, nor were they available from every group, for consistency, only records of victims described as "martyrs" are included from each of these groups in this analysis (see Appendix A for more details).

3 Methodology

This report begins with 318 910 records of reported killings of fully identified victims from five datasets. Many of these records are duplicates. An initial automated pass over all 318 910 records generates "blocks" of records for human review. These "blocks" are groups of records with identical names, dates, locations, or some combination of two or three of those attributes.

Next, several experts review these blocks, sometimes organized as pairs for comparison and other times organized as entire spreadsheets for review. These experts determine whether pairs or groups of records refer to the same death or not. Pairs or groups of records determined to refer to the same individual are assigned to the same "match group." All of the records contributing to a single "match group" are then combined into a single record. This new single record is then again examined as a pair or group with other records, in an iterative process. For example, two records with the same name, date, and location may be identified as referring to the same individual, and combined into a single record. In a second review process, it may be found that that record also matches the name and location, but not date, of a third record. The third record may list a date one week later than the two initial records, but still be determined to refer to the same individual. In this second pass, information from this third record will also be included in the single combined record. For further details on this process, including the incorporation of contradictory information across multiple records, see Appendix C

Three of the experts are native Arabic speakers; they review records with the original Arabic content. Two of the experts review records translated into English. These five experts review overlapping sets of records, meaning that some records are evaluated by two, three, four, or all five of the experts. This makes it possible to check the consistency of the reviewers, to ensure that they are each reaching comparable decisions regarding whether two (or more) records refer to the same individual or not. Formally, this process is referred to as inter-rater reliability, and is described further in Appendix C.1.

4 Documentation Patterns

Despite the enormous efforts by the data collecting groups, many killings in Syria are still undocumented. One way to imagine that is to consider that in any particular month, some killings are documented by four groups, other killings are documented by three groups, others by two groups, and some killings are reported by only one group. The question this observation raises is: how many killings are reported by zero groups?

The answer to this question requires statistical modeling and is beyond the scope of this report. Figure 1 provides a way to visualize the intensity of reporting by examining the killings documented by the four Syrian human rights groups (CSR-SY, SNHR, SOHR, and VDC).

Note the grey shading in Figure 1, indicating the period from May 2013 through April 2014, when the most number of sources available is three, since SOHR records at our disposal end in April 2013.

To interpret the graph, compare July 2013 to August 2013. Both months have roughly comparable numbers of killings recorded by two or three sources (the two darker shades at the bottom of each bar). But there are a notably larger number of killings in August documented by just a single source. Closer examination of these particular records may provide further context - were they all documented by the same source? In the same governorate? The key observation from Figure 1 is that in all months, at least some killings are reported by only one group as shown by the light pink part at the top of each bar.

Reframing the question posed above, how many new killings might be found if we included records from a fifth group? Then a sixth, seventh, eighth, etc., group? This is what modeling can answer, estimating the total number of deaths likely to have occurred, starting from those that have been documented and adding the number not yet documented.

It should also be noted that this documentation pattern, aggregated across the entire country, changes dramatically when broken down by geographic area. As one example, consider Figure 2 which shows the number of records documented by number of sources within the governorate of Tartus. This pattern differs substantially from the pattern in Figure 1. In particular, the vast majority of killings in Tartus appear to be documented by a single source (the majority of each bar in Figure 2 consists of the lightest pink portion, indicating one source). One exception to this pattern are killings during the month of May 2013. We will return to this observation in the following section.

Figure 1: Documented Killings by Month and by Number of Sources per Killing; Note that from May 2013 through April 2014 only three sources are available.
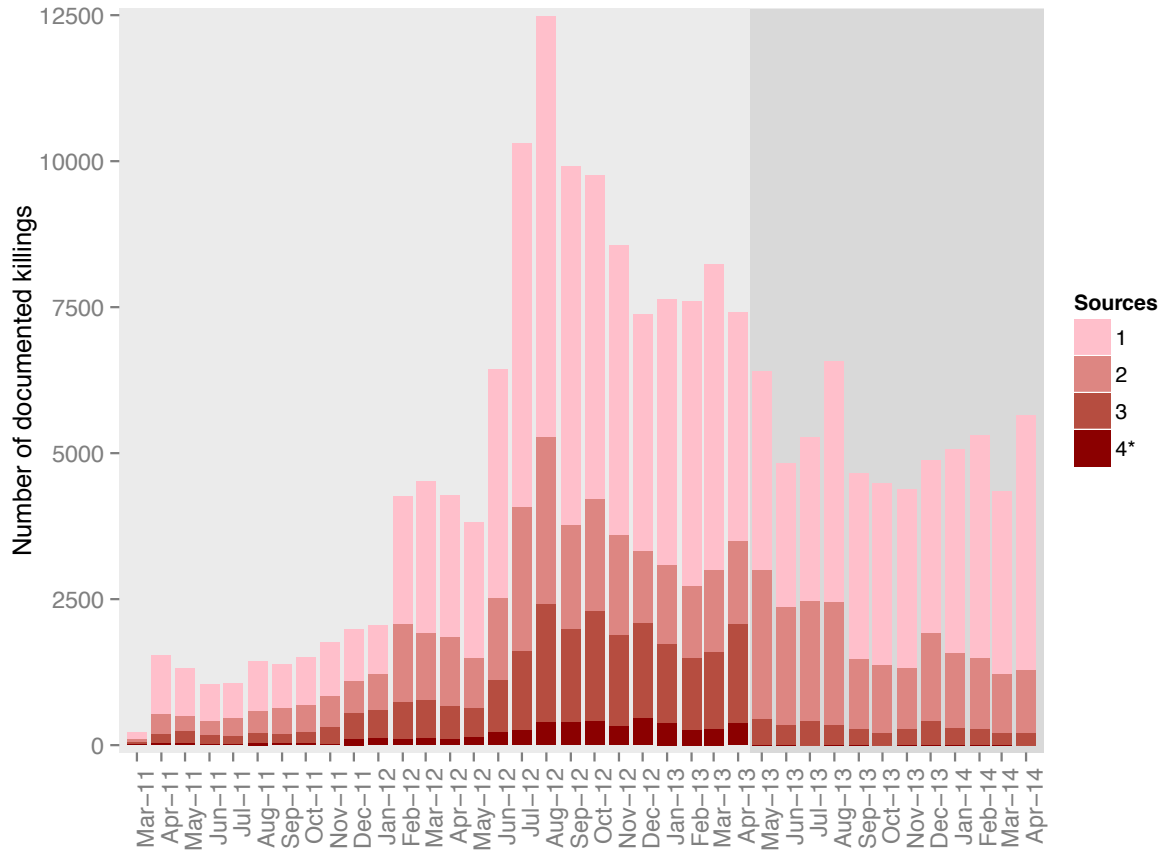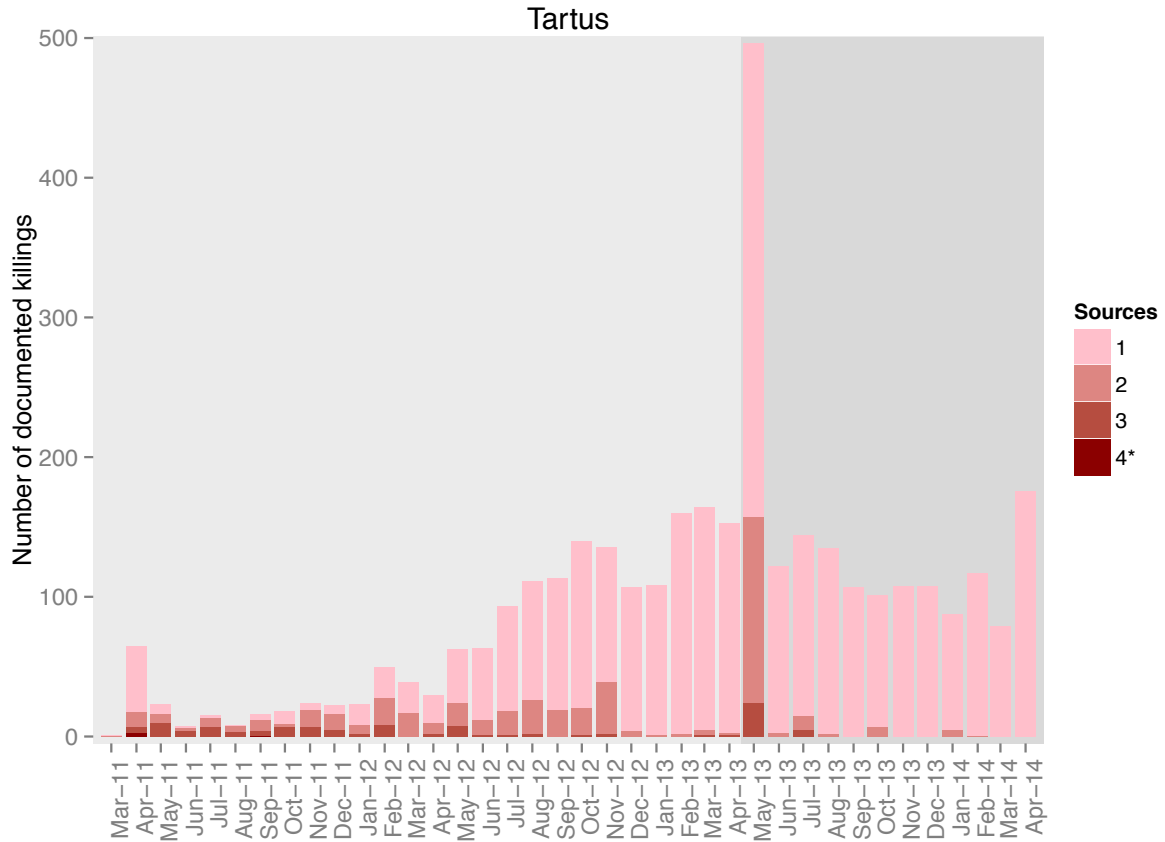
Figure 2: Documented Killings by Month and by Number of Sources per Killing in Tartus; Note that from May 2013 through April 2014 only three sources are available.

A similar comparison can be drawn by examining the number of sources documenting each killing reported in each governorate (Figure 3).[14] For example, very similar numbers of killings have been reported in Aleppo and Homs by two, three, and four sources (the darkest shading at the far left of each bar). However, somewhat more killings have been reported in Aleppo than in Homs by a single source. The implication of the greater number of singly-reported killings is that there may be many more killings in Aleppo than in Homs, not just the slightly greater number that have been reported. It is only possible to speculate about such potential patterns based on the observed data; statistical modeling is necessary to address questions about the total magnitude and true pattern of all killings.

---

[14]Graphs of categorical variables will be horizontally oriented throughout this report.

Figure 3: Documented Killings by Governorate and by Number of Sources per Killing
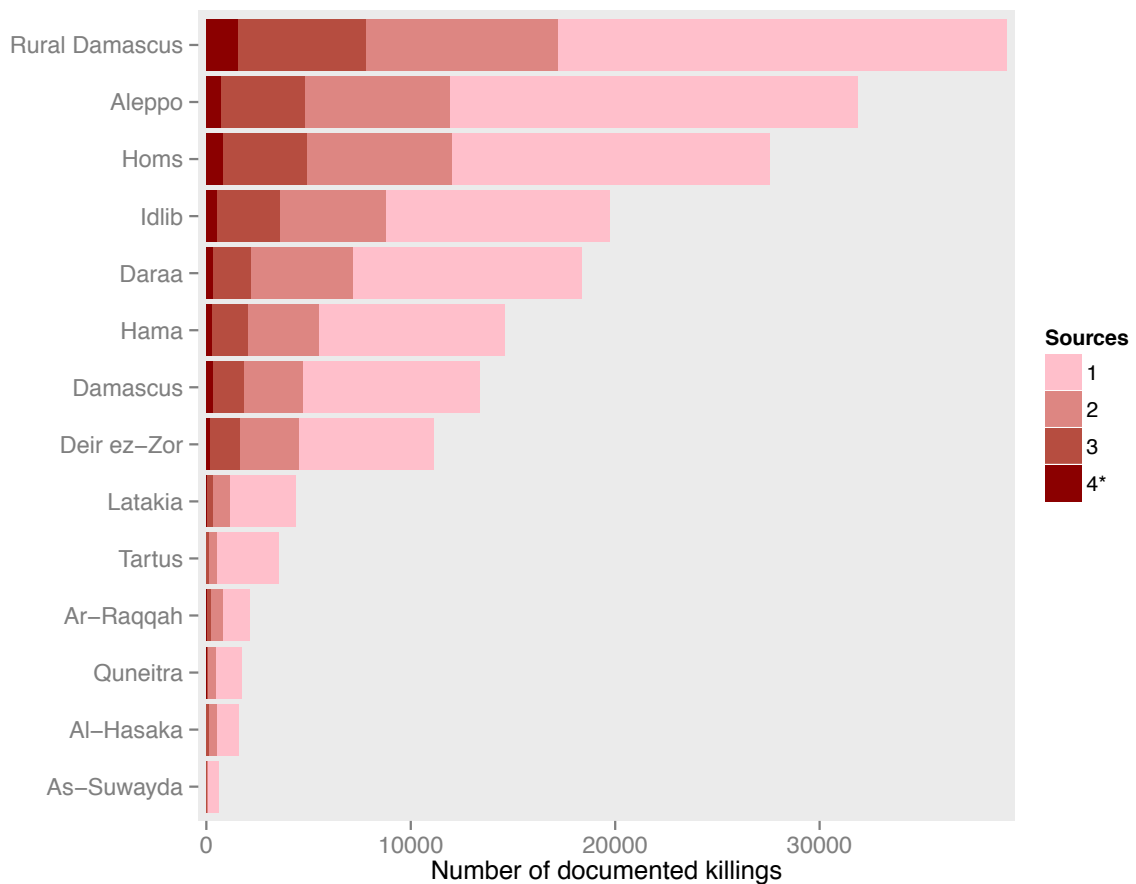
4.1 New Data Since Previous Report

Three of the sources (CSR-SY, SNHR, and VDC) have continued to share records with this project. The new records include killings documented between May 2013 and April 2014, the period of study since the previous report. These data groups have also documented additional killings that occurred between March 2011 and April 2013, but which were learned about, identified, or verified since the last report. In other words, records from these three groups reflect the ongoing work each of these groups conducts updating and verifying their existing records in addition to collecting new records.

Looking at the period of study from the previous report, this updated analysis finds that these three sources recorded 116 046 unique killings between March 2011 and April 2013. Notably, this is a higher count than the total enumeration previously reported (92 901). This is due to a combination of newly-documented deaths that occurred between March 2011 and April 2013, additional information discovered about previously-reported deaths, and refinements to the matching model.

From May 2013 through April 2014, CSR-SY, SNHR, and VDC have documented 61 816 unique killings. For this period, these three groups together have documented between 4344 and 6565 killings each month.[15]

---

[15]For this period, a monthly average of 5151 killings were documented by these three sources.

5 Descriptive Statistics

This section presents summary statistics that describe the datasets that were integrated for this enumeration. The analyses describe only identifiable victims reported by each individual dataset; unobserved and unidentifiable killings are not considered. Therefore, the analysis is affected by selection bias. That is, each killing has a different likelihood of being reported, due to individual characteristics of the victim and to field practices of each reporting group. For example, one data collection group may have better contacts within a certain religious or ethnic group, whereas another may have access to government personnel records. Another group may have excellent sources one week but be unable to contact these sources at other times. And of course, some violent events are not reported to any source, either because only the perpetrators survived the event, or because surviving witnesses were unable or chose not to report the incident. Raw data, including individual datasets and integrated enumerations such as the one presented in this report, are not suitable for drawing conclusions about statistical patterns. To draw rigorous conclusions, estimates that correct for selection bias must be made.

Nevertheless, analysis of the individual datasets explores what has been seen. This analysis is called "descriptive" because it describes the data. Although this may not provide insight into the unobserved true patterns, descriptive analysis shows what the datasets have in common, and how they differ.

These descriptive statistics only include records of identifiable victims. Records of identifiable victims include the victim's name, plus date and location of death.[16] The full identifying information is essential for the record comparisons required to match records across different datasets. Records lacking the complete information are considered "anonymous" and were excluded from the integration and analysis (see Table 2). The anonymous records describe victims of violence in Syria who deserve to be acknowledged. However, they cannot be included in this 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; a full accounting of killings—documented and undocumented—will require additional data analysis.

5.1 Documentation Over Time

Figure 4 shows the frequency of reported killings by week for each dataset. Three datasets (CSR-SY, SNHR, and VDC) indicate roughly comparable patterns of violence over time. Note, however, that the y-axes are very different. VDC and SNHR report the highest number of killings, followed by CSR-SY. Here it is important to note again that findings concerning GoSY and SOHR included in Figure 4 only reflect a portion of the conflict because the records available from these sources cover a shorter time period (GoSY records cover March 2011–March 2012 and SOHR records cover March 2011–April 2013).

It is important not to be misled by the apparent consistency across multiple sources when aggregated over the entire country of Syria. As hinted at in the previous section, expansion of these comparisons to specific locations of interest rapidly reveal differing patterns in the observed data. It is important to examine these patterns because it is

---

[16]A discussion of location and other data processing questions can be found in Appendix B.

precisely these differences where it is possible to learn the most, both about documentation processes themselves and about the true underlying patterns of violence.

Figure 5 shows the monthly count of documented killings in Tartus by each source. As indicated in Figure 2, the majority of observed killings in Tartus were documented by a single source (in this case, VDC). An exception to this pattern is the large peak visible in three of the graphs in May 2013. This corresponds to an alleged massacre in Banias.[17] It appears that CSR-SY, SNHR, and VDC all documented some portion of this event.

When interpreting figures such as Figure 5, the aim should not be to identify a single "correct" source. All of these sources are documenting different snapshots of the violence, and all of them are contributing substantial numbers of unique records of victims undocumented by the other sources. VDC is capturing some events that are not captured by the other sources, but there is no way to tell how many events are not being captured by VDC. It is worth noting the Oxford Research Group's finding that many of these groups report difficulty accessing information in areas considered to be loyal to the government, in particular, such as Tartus and Latakia (Salama and Dardagan, 2013).

Figure 4: Distribution of Total Reported Deaths by Week; Note that each y-axis is different.

---
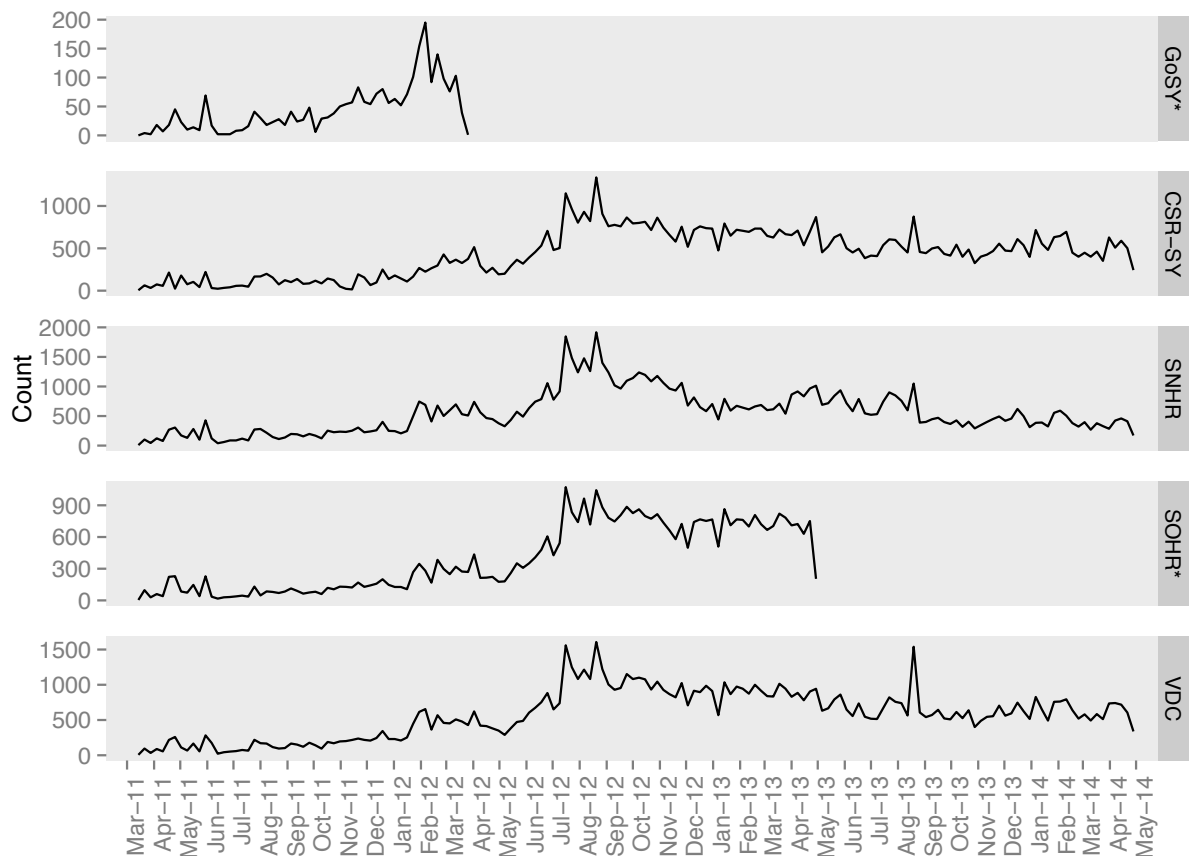
[17]See report in the LA Times, BBC, and The Independent, among others.

Figure 5: Distribution of Total Reported Deaths by Month in Tartus
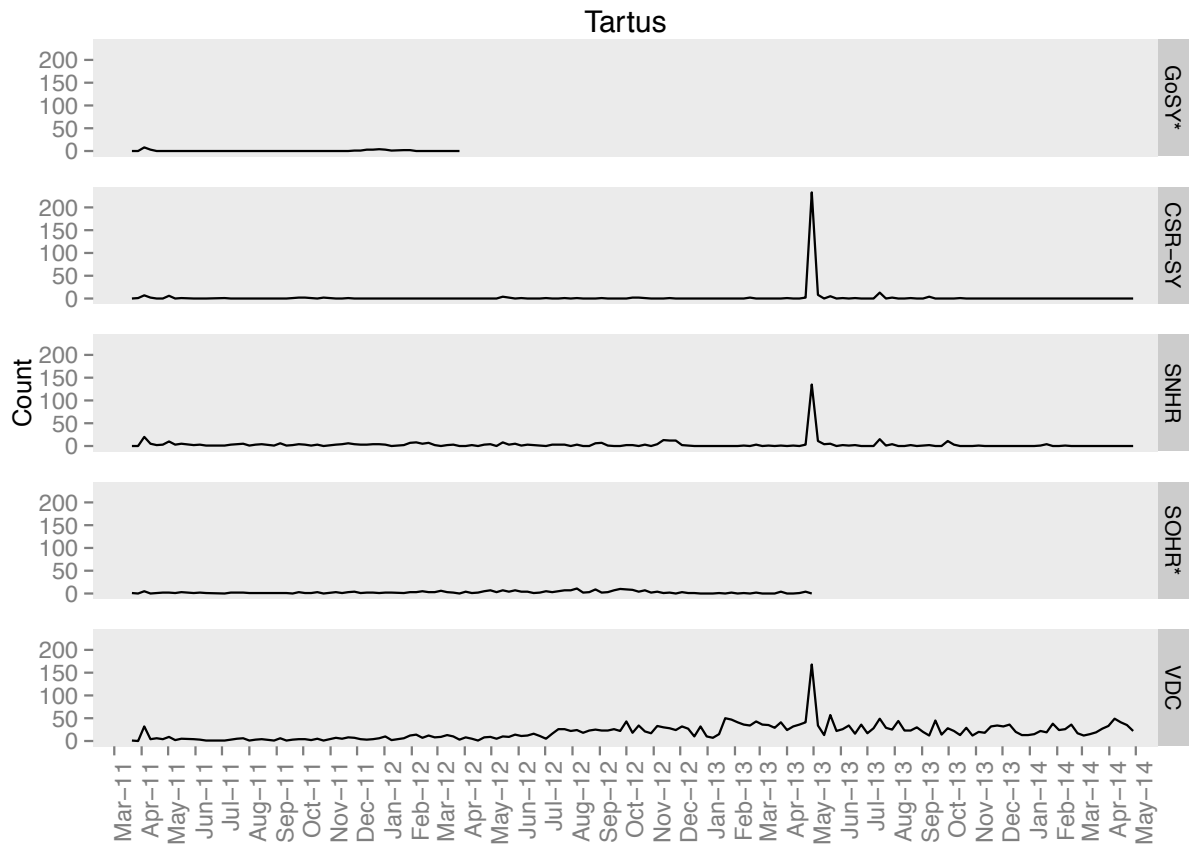
5.2 Documentation Over Geographic Area

Figure 6 indicates patterns of reported violence over geographic area for each of the five datasets. Four of these datasets (CSR-SY, SNHR, SOHR, and VDC) record the highest number of reported killings in Rural Damascus. GoSY records the highest number of killings in Homs. Homs is the second-most frequently reported governorate in SNHR and SOHR, third-most frequent in VDC and CSR-SY. Rural Damascus, Idlib, Daraa, and Hama are all reported comparably frequently in GoSY records, and all are a distant second to the number of killings recorded in Homs by this source.

When drawing such comparisons it is important to note that GoSY and SOHR records cover a shorter time period than records from CSR-SY, SNHR, and VDC. Specifically, records available from GoSY cover March 2011–March 2012 and records available from SOHR cover March 2011–April 2013, whereas the other three sources cover March 2011–April 2014. It is possible that updated records from GoSY and SOHR would indicate different patterns across governorates than those indicated in Figure 6.

It should also be noted that it is possible the geographic pattern in some data is being misinterpreted. It was not always possible to determine the governorate of death precisely from the available data. In some cases it was necessary to assume that the governorate of death was the same as the governorate of birth. See Appendix B for further details.

Figure 6: Distribution of Recorded Deaths by Governate
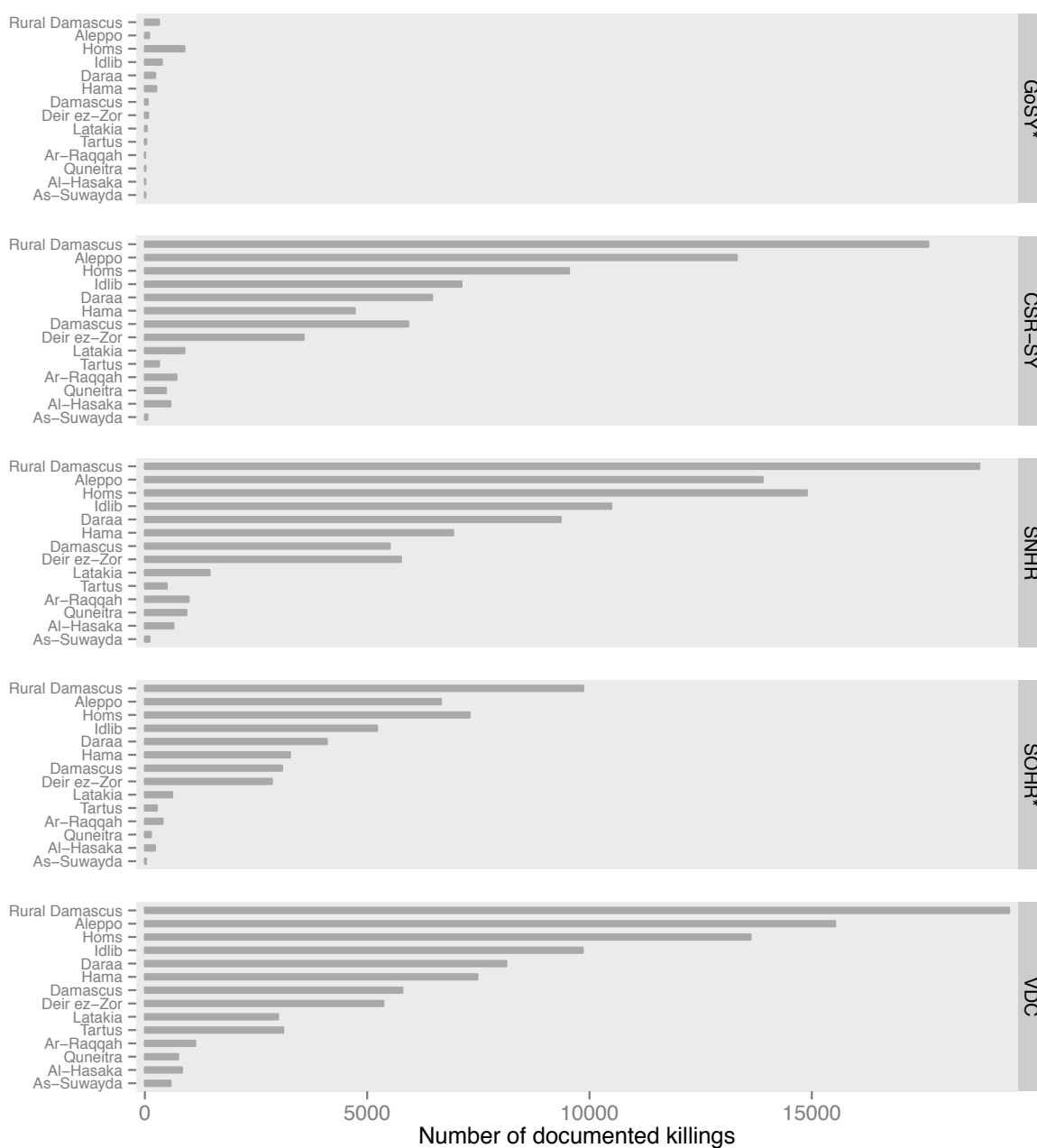
5.3 Documented Victim Demographic Characteristics

All five datasets include information about the sex (Table 3) and age of victims (Figures 7 and 8). There seems to be general agreement across the datasets that the majority of victims are male. Considering the integrated data, of the 191 369 unique records of documented killings in this report, 85.1% are male victims, 9.3% are female victims, and 5.6% of records do not indicate the sex of the victim. This is likely a very slight undercount of female victims, as the expert reviewers have discovered a small number of records with a clearly female name incorrectly recorded as male. In future research, it might be possible to infer and correct the recorded sex of these victims based on their

names. This was beyond the scope of the current project.

Table 3: Total Documented Killings by Sex

Dataset	Female	Male	Unknown
GoSY	0	2443	20
SOHR	3993	29 137	11 001
SNHR	8413	79 839	2398
CSR-SY	7533	63 604	378
VDC	8937	84 047	1806

The majority of records (83.8%) lack information about the age of victims, which makes it impossible to draw conclusions about the distribution of violence over age categories. To be clear, this is not a criticism of any of these documentation efforts, but rather an indicator of just how difficult it can be to record accurate age information. Of the records that do include age information, 2165 indicate victims 0-9 years old, and 6638 victims 10-18 years old.

Consider the histograms in Figure 7. With the exception of GoSY, the remaining datasets are all missing information on age for nearly 80% or more of records. The records without ages could have substantially different ages than the records with reported ages. For example, the age of very young people and very old people is often relevant to their identity. "He was only four years old" or "he was over seventy years old" are common phrases, but there is no comparable salience for an adult's age. It may be that most or all of the records with missing age data are in fact adults, which would make most distributions look more like the GoSY patterns. The high proportion of missing age data makes it impossible to draw conclusions about the true distribution of the age of victims reported to each group.

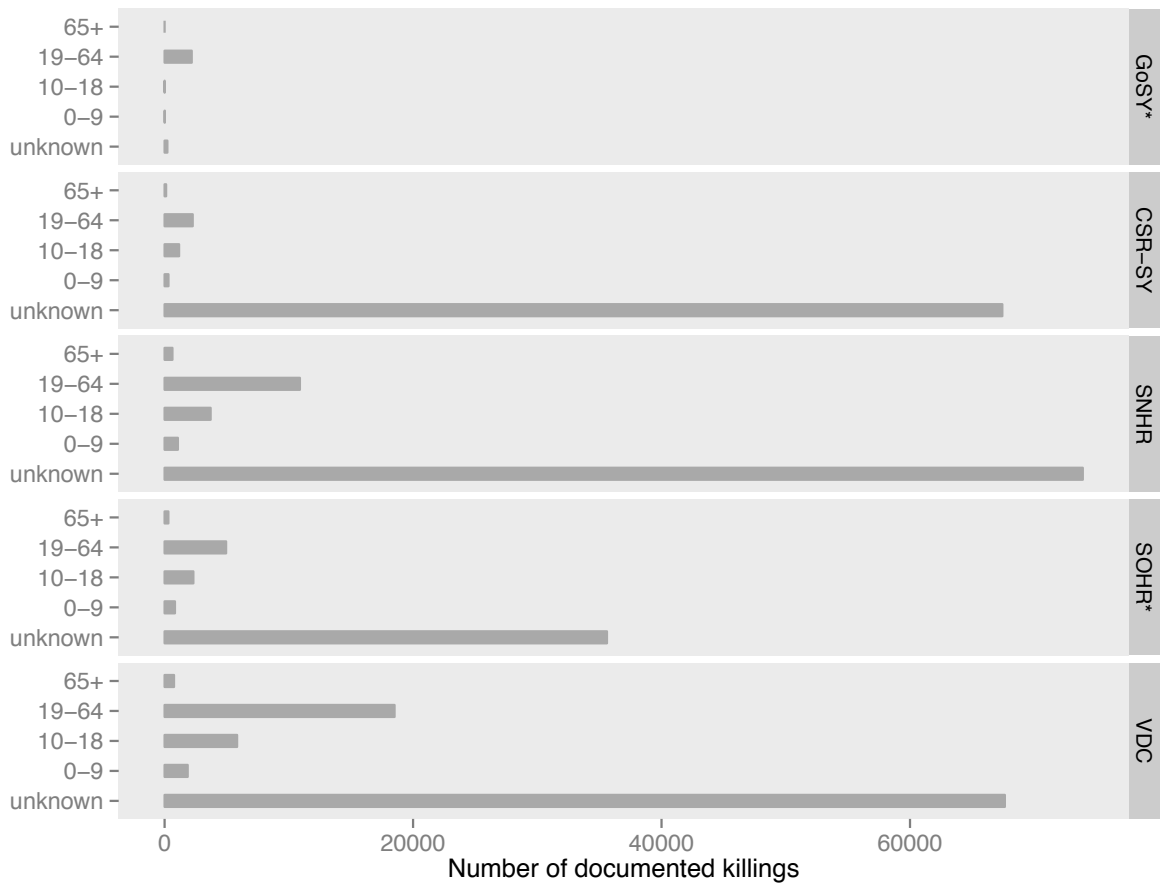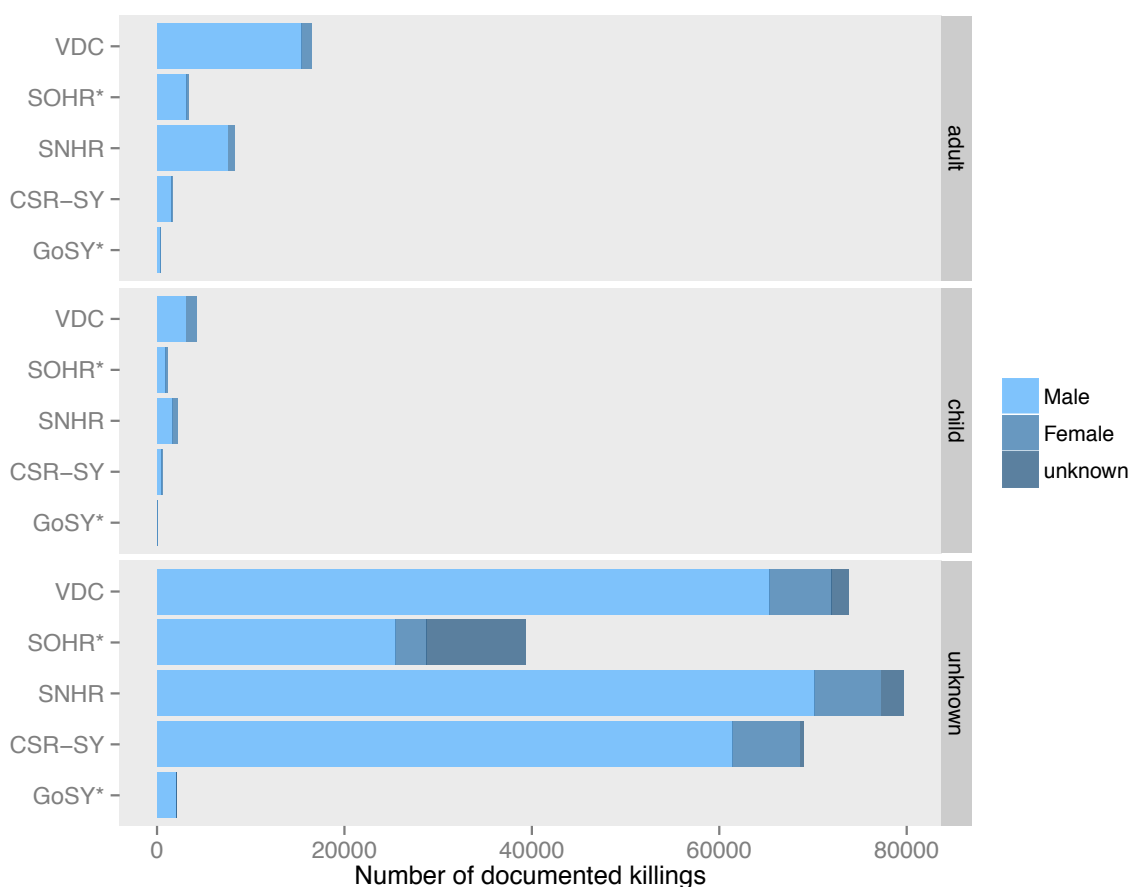Figure 7: Distribution of Recorded Deaths by Age

Number of documented killings

Figure 8: Distribution of Recorded Deaths by Age Group

6 What these Numbers Can and Cannot Tell Us

The enumeration provided in this report—191 369—is the most accurate accounting available based on identifiable victims reported by these five groups. However, many victims are not yet included in these databases, and the excluded victims may be systematically different from the victims who are recorded. Well-known individuals who are victims of very public acts of violence, and victims who are killed in large groups tend to attract public attention, and they are therefore likely to be reported to one or more of these sources. By contrast, single individuals killed quietly in a remote corner of the country tend to be overlooked by media and documentation projects.

Different proportions of killings are reported depending on when, where, and how the killing happens, and who the victims are. These differences are generally called "selection bias," and there are many variations.[18] Bias means that patterns in the raw data may be misleading regarding the pattern and magnitude of violence occurring in Syria. It may be that more violence is occurring in Idlib than Daraa, but it may just be that more violence in Daraa is not being documented. It may be that violence peaked in the summer of 2012, or it is possible that documentation efforts have suffered so the apparent violence is declining but true violence continues at the August 2012 level. In

---

[18]For an analysis of event size bias using examples from Syria and similar documentation projects in Iraq, see Carpenter et al. (2013) and Price and Ball (2014).

order to understand the true underlying patterns of violence, statistical estimates will be needed to identify and correct biases.

Examining reported killings is an important step in understanding violence in Syria. But it is only the first step. Further analysis is necessary to answer substantive questions about patterns of violence during this conflict.

A Data

A.1 Sources

HRDAG obtained data from various sources via different mechanisms and at different times. Below is a brief description of each source, how and when HRDAG obtained data from each source, and any additional information available about each source's mission and data collection and verification methods.

- Syrian Government: This list was provided to HRDAG by OHCHR in September 2012. OHCHR requested updated information from the Syrian Government in September 2013 and April 2014; no response was received at the time of publication.

- Syrian Center for Statistics and Research: HRDAG scraped[19] CSR-SY's website[20] in May 2013 to obtain a copy of their published data. Individuals can fill out a form on the CSR-SY website to add victim information. HRDAG established direct contact with CSR-SY in late May 2013 and all data used in this report were provided directly from CSR-SY.

- Syrian Network for Human Rights: This list was 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 victims. All data included in this report were provided directly from SNHR. SNHR maintains a website[21] 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.

- Syrian Observatory for Human Rights: This list was provided to HRDAG by OHCHR in December 2012 and again in May 2013, as agreed to by SOHR. This list includes only "[c]ivilians and opposition fighters who are not defectors" as categorized by SOHR. SOHR also collects data on defectors, pro-government militia (Sabihha), military and police personnel, unidentified persons, unidentified and foreign fighters, and Hezbollah fighters. SOHR declined to share an updated list in early 2014.

- Violation Documentation Centre: This list was provided to OHCHR in February 2012. Subsequently HRDAG scraped the website[22] several times in 2012, 2013, and 2014 to obtain updated data. HRDAG only includes records from the list labeled "Martyrs." The "About" page of their website describes the data classification methods and three-stage data verification process implemented by the VDC.

---

[19]Using a computer program to extract information from websites.
[20]http://csr-sy.org/
[21]http://sn4hr.org/
[22]http://www.vdc-sy.info

A.2 Overlap Among Sources

As noted above, a total of 191 369 unique killings were documented by five sources (GoSY, CSR-SY, SNHR, SOHR, and VDC). Each of these sources contribute some records that were only documented by that source, as well as records that were reported by that source and by one or more other sources.

The remainder of this discussion focuses only on the human rights groups; these four sources (CSR-SY, SNHR, SOHR, and VDC) contributed a total of 189 739 records. The remaining 1630 records were contributed only by the Syrian Government and were not included on any of the human rights groups' lists. Of the 189 739 unique killings, only 4877 were reported by all four Syrian human rights groups.

A more detailed way to look at this distribution of records is presented in Table 4. The first row of Table 4 indicates the 4877 records reported by all four human rights groups. Note that in the four leftmost columns, there is a '1' in each, indicating that the row includes records from this dataset.

The second row of the table has the pattern 1101, which indicates that it counts killings recorded by SOHR, CSR-SY, and SNHR but not by VDC; there are 1107 killings with this reporting pattern.

The sixth row of the table has the pattern 1001, which indicates that it counts killings recorded by SOHR and SNHR but not by CSR-SY or VDC; there are 4734 killings with this reporting pattern.

The final row of the table has the pattern 0001, and it shows the killings reported only by SNHR, not by SOHR, CSR-SY, or VDC. There are 19 917 killings with this reporting pattern.

Table 4 makes clear that VDC and SNHR share the most overlap, while CSR-SY has the most uniquely reported killings. SOHR is relatively smaller than the other sources, but recall that data from SOHR covers only the period March 2011–April 2013.

Table 4: Distribution of Records into Four Sources				
SOHR	**CSR-SY**	**VDC**	**SNHR**	**Number of Records**
All Four Datasets				
1	1	1	1	4877
Three Out of Four Datasets				
1	1	0	1	1107
1	0	1	1	17 579
0	1	1	1	5851
1	1	1	0	494
Two Out of Four Datasets				
1	0	0	1	4734
0	1	0	1	1693
1	1	0	0	526
0	1	1	0	1458
0	0	1	1	34 447
1	0	1	0	2817
Unique to a Single Dataset				
1	0	0	0	11 877
0	1	0	0	55 289
0	0	1	0	27 073
0	0	0	1	19 917

A.3 Potentially Inaccurate Records

Numerous sources have mentioned to HRDAG the possibility of inaccurate records.[23] HRDAG is very interested in learning more about potentially inaccurate records. To this end, HRDAG has asked several of the documentation groups if they would provide further information about and examples of inaccurate records so that these may be excluded from the count. It would be even more interesting to use examples of potentially inaccurate records to model the impact of inaccuracy in the statistical analyses. Only a few examples (fewer than 1000) have been shared so far.

There are a variety of ways in which a record may be potentially inaccurate. For example, some records may describe people who died of non-conflict-related causes; in the context of a database of killings, these records are potentially inaccurate. For example, victims of accidents or illness mistakenly included in lists of conflict-related killings would be one kind of inaccurate record. Victims who were believed to be dead but are later discovered to be alive would also be inaccurate. Individuals who were missing following a violent event, or who disappeared for some time may have been mistakenly recorded on a list of conflict-related killings. There is also the possibility that some records are fabricated, that is, records of victims who do not in fact exist at all. HRDAG is only aware of a small number of specific examples of inaccurate records (examples provided by VDC), but it is possible that some additional inaccurate records are included in these data.

Although some inaccurate records may be found, there may be others which cannot be identified, which is precisely why the characteristics of inaccurate records need to be modeled. Statistical modeling is the tool scientists use when some information is known

---
[23]For one example, see Stephen Starr of the Toronto *Globe and Mail* (Starr).

but other information needs to be estimated. With a large number of known inaccurate and known accurate records, it might be possible to compare these two groups of records and identify key characteristics that differ between them. With this information, a classification model could be built to identify sets of potentially inaccurate records (i.e., "scenarios" of kinds of inaccuracy). The modeling would use records previously identified by human reviewers to suggest records not yet identified as potentially inaccurate. With the modeled information, HRDAG could examine how these records are distributed with respect to geography, time, and the characteristics of the victims, and thus determine how inaccuracy might be affecting the substantive conclusions of future analysis.

B   Data Processing

The data processing step standardizes the structure and content of the different sources prior to matching and de-duplication. Processing includes three important steps: cleaning, translating, and what HRDAG refers to as canonicalizing.

B.1   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. 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 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.

B.2   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 three native Arabic speakers who review records) (see Section 3 and Appendix C) 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 three native Arabic speaking reviewers.

For the fourth and fifth 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 three 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 C.1.

B.3   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, and in some cases HRDAG assumes that the governorate of birth matches the governorate of death.

C   Matching

As mentioned in Section 1, to use the records described in this report, they must be linked together, identifying groups of two or more records which refer to the same person. This is challenging because each data source may record slightly different information (as indicated by Section 5). Furthermore, each data source is working to overcome the difficulties inherent in collecting complete, accurate information in the midst of a conflict.

Linking records within a single data source is called de-duplication, and identifying the same death in records found in different sources is called record linkage. Both of these tasks were performed together, by looking for duplicates within a single list of all records from all data sources with sufficient information, including name plus date and location of death. Other variables, such as age (or date of birth), sex, and location of birth, were also used for matching.

From the full set of pooled records, groups of records that might be matches were first identified using very broad rules, dividing the pool of records according to governorate and year. In other words, all killings recorded in Homs in 2011 were examined as a group, looking for records with similar names and dates. Reviewers then labeled groups of records as referring to the same individual or not.

The matched groups of records which all refer to the same person were then combined; this is called clustering. For example if record A matches record B, and record B matches record C, then the group (A, B, C) might be formed. Each of these groups contains all the records from all the databases that reviewers believe refer to the same death. Finally, the records in each group are merged into a single record containing the most precise information available from each of the individual records. If some records contain contradictory information (for example, if records A and B record the victim as age 19 and record C records age 20) the most frequently reported information is used (in this case, age 19). If the same number of records report each piece of contradictory information, a value from the contradictory set is randomly selected.

After an initial round of clustering, subsets of these combined records were then re-examined to identify previously missed groups of records that refer to the same individual, particularly across years (e.g., records with dates of death 2011/12/31 and 2012/01/01 might refer to the same individual) and governorates (e.g., records with neighboring locations of death might refer to the same individual).

C.1   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 4084 pairs of records (coder A in English, coder B in Arabic) and made the decisions summarized in Table 5. Their overall agreement is 99.8%. This is calculated based on the 2593 pairs both coder A and B labeled as matches plus the 1481 they both labeled as non-matches. Thus they agreed on 4074 pairs out of 4084, for an overall agreement of 99.8%.

Table 5: IRR Example

| | Coder A | |
Coder B	Match	Non-Match
Match	2593	3
Non-Match	7	1481

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 (99.8% in the example above) 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 example above:

$$p_c = \frac{1484}{4084} \times \frac{1488}{4084} + \frac{2600}{4084} \times \frac{2596}{4084}$$

In this case $p_c = 0.54$, so Kappa $= \frac{0.998-0.54}{1-0.54} = 0.996$, which is very good. 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).

For all of the analyses presented in this report, kappa statistics between reviewers were over 0.9 and overall agreement was over 97%. Therefore, HRDAG considered the decisions made by each of the reviewers to be highly consistent, regardless of whether they were reviewing the records with the original Arabic content or translated into English.

References

Carpenter, D., Fuller, T., and Roberts, L. (2013). WikiLeaks and Iraq Body Count: the Sum of the Parts May Not Add Up to the Whole—a Comparison of Two Tallies of Iraqi Civilian Deaths. *Prehospital and Disaster Medicine*, 28(3).

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. *utor Quant Methods Psychol*, 8:23–34.

Price, M. and Ball, P. (2014). Big Data, Selection Bias, and the Statistical Patterns of Mortality in Conflict. *SAIS Review of International Affairs*, 34:9–20.

Salama, H. and Dardagan, H. (2013). Stolen Futures: The Hidden Toll of Child Casualties in Syria. Oxford Research Group.

Starr, S. Casualty count proves inexact science, rights group finds. *The Globe and Mail* 27 March 2013 Available: http://www.theglobeandmail.com/news/world/casualty-count-proves-inexact-science-rights-group-finds/article10467644/.

About HRDAG

The Human Rights Data Analysis Group is a non-profit, non-partisan organization[24] 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) and Facebook page.

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. The content of this analysis does not necessary reflect the opinion of OHCHR.

---

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