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

Commissioned by
the Office of the UN High Commissioner for Human Rights

Megan Price, Jeff Klingner, Anas Qtiesh, and Patrick Ball

Human Rights Data Analysis Group
everybody counts.

13 June 2013

Executive Summary

This report presents findings integrated from seven databases built by Syrian human rights monitors and one database collected by the Syrian government. The databases collect information about conflict-related violent deaths—killings—that have been reported in the Syrian Arab Republic between March 2011 and April 2013. Although conflict conditions make it difficult to identify an accurate record of events, governmental and non-governmental monitors are persevering in gathering information about killings through a variety of sources and data collection methods. The purpose of the report is to explore the state of documentation, the quantitative relationship of the sources to each other, and to highlight how understanding of the conflict may be affected due to variations in documentation practices.

This report examines only the 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. The status of the victims as combatants or non-combatants is unknown for all but a few records. This report finds that when the fully identified records were combined and duplicates identified, the eight databases collected here identified **92,901** unique killings. The listing of killings is called an enumeration.

The enumeration is *not* the complete number of conflict-related killings in the Syrian Arab Republic. The enumeration may be a slight overcount of the number of 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 anywhere. 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 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.[1] To

[1] For more discussion of potentially inaccurate records, see Appendix A.2.

date, the largest estimate of such records is 1,000. Both of these factors may slightly inflate the current enumeration.

However, based on experience in similar contexts, HRDAG believes that many killings remain undocumented. HRDAG has conducted analyses of documented killings in Guatemala, Kosovo, Perú, Timor-Leste, and Colombia. In each of those studies, HRDAG used the documented deaths in statistical models to estimate the probable number of deaths that remained undocumented.[2] A full estimate of the total killings, including documented killings plus undocumented killings, will need to wait for a future analysis in a scientific publication. Nonetheless, the enumeration here presents the most complete and precise listing of killings known by these eight sources as of April 2013. The total 92,901 can be understood as a minimum bound of the number of killings between March 2011 and April 2013.

This report is an update of work published in January 2013.[3] The January report presented an enumeration of reported killings in the Syrian Arab Republic between March 2011 and November 2012. This updated analysis finds that the eight sources integrated here have recorded 26,906 unique killings between December 2012 and April 2013. A combination of newly documented deaths that occurred between March 2011 and November 2012, and refinements to the matching model resulting from the new data add 2,956 killings to the total. Finally, a new source (the Syrian Center for Statistics and Research) added 3,391 records of previously undocumented killings between March 2011 and November 2012.

The rate of documented killings is slightly lower in this period than it was during the peak from July through October 2012. However, the level of documented killings has been sustained at greater than 5,000 observed killings each month from July 2012 through April 2013.[4] More killings are certainly unobserved. The rate of killing observed since July 2012 has been consistently and substantially greater than the period prior to July 2012.

[2]For these reports and various related publications, see the HRDAG publications page. For a discussion of the relevant methods, see HRDAG's page describing multiple systems estimation.

[3]The January 2013 report was published by Benetech. At the time of the January report, the Human Rights Data Analysis Group formed part of Benetech's Human Rights Program. HRDAG was spun-off into an independent non-profit organization in February 2013. The researchers working on this project are the same as those who contributed to the January report.

[4]Between July 2012–April 2013, the number of killings observed by one or more of the datasets integrated here has varied from a minimum of 5049 to a maximum of 7992.

Report Organization

Section 1 provides a summary of documented killings in the Syrian Arab Republic between March 2011 and April 2013. 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 eight 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 eight 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 eight data sources and discusses concerns regarding potentially inaccurate records. Lastly, Appendices B and C provide technical and methodological detail.

1 Documented Killings

This report presents an analysis of killings that have been reported in the Syrian Arab Republic between March 2011 and April 2013, based on eight datasets. Based on a comparison of records from these eight sources, HRDAG found a total of **92,901** unique records of documented killings. Importantly, this enumeration should not be inferred to include only civilian victims. The status of documented victims as combatants or non-combatants is reported in relatively few records in these datasets, but both status are reported. Therefore, collectively the data sources include records of both combatants and non-combatants.

The analysis in this report updates the previous report released on 2 January 2013. In the January report, 59,648 unique killings were identified for the period March 2011 through November 2012. This report adds 26,906 records from the period December 2012 through April 2013. A combination of newly documented deaths that occurred between March 2011 and November 2012, and refinements to the matching model resulting from the new data add 2,956 killings to the total. Finally, a new source (the Syrian Center for Statistics and Research) added 3,391 records of previously undocumented killings between March 2011 and November 2012.

The eight sources examined in this report are:

1. **15Mar**: the March 15 Group
2. **GoSY**: the Syrian government
3. **SCSR**: the Syrian Center for Statistics and Research[5]
4. **SNHR**: the Syrian Network for Human Rights[6]
5. **SOHR**: the Syrian Observatory for Human Rights[7]
6. **SRGC**: the Syrian Revolution General Council, which was combined with the SNHR[8]
7. **SS**: the Syria Shuhada Website [9]
8. **VDC**: the Violations Documentation Centre[10], the documentation arm of the Local Coordination Committees

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.

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.

Each dataset covered slightly different periods of time (see Section 5 for more detailed descriptions of each individual dataset) so this comparison of

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

[6] http://www.syrianhr.org/

[7] www.syriahr.com www.syriahr.net

[8] HRDAG learned that SNHR was a spin-off of SRGC, so the records of these two groups were integrated before comparing them with 15Mar, GoSY, SCSR, SOHR, SS, and VDC. From the time period covered by SRGC, 90.5% of killings recorded by SRGC were also recorded by SNHR. Considering the high level of overlap, the contextual knowledge that SNHR was originally a part of SRGC, and the fact that SNHR's dataset covers a longer period of time, HRDAG chose to combine the SNHR and SRGC datasets into a single dataset, referred to in the following sections as only SNHR.

[9] http://syrianshuhada.com/

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

records was conducted over several periods.[11]

Table 1: Time Period Covered by Each Source

Dataset	**Period Covered**
15Mar	March 2011–December 2011
SRGC	March 2011–January 2012
GoSY	March 2011–March 2012
SCSR	March 2011–April 2013
SNHR	March 2011–April 2013
SOHR	March 2011–April 2013
SS	March 2011–April 2013
VDC	March 2011–April 2013

The March 15 group stopped collecting data in December 2011, so records from this source were only included in the first ten months of analysis. Similarly, data from SRGC covered the period from March 2011 to January 2012 and the available government data extend until March 2012 only. Updated data from the Syrian government were not available. The five remaining organizations (SCSR, SNHR, SOHR, SS, and VDC) are still actively recording killings; this analysis only includes their records with dates of death through April 2013.

2 What These Analyses Do and Do Not Include

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.[12] Each dataset considered in this study included a number of records which lacked this information. Table 2 lists the number of records[13]

[11]The record linkage was conducted over four periods - an earlier report released in January compared records from March to December 2011, January to March 2012, and April to November 2012. Updated records examined in this report from December 2012 through April 2013 were treated as a fourth partition. Five sources cover the entire period from March 2011 to April 2013.

[12]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.

[13]Note that SOHR and SCSR coincidentally contributed the same number of identifiable records; these entries in the table are not a typo.

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 worth noting that none of the included counts in Table 2 match the total number of documented killings—92,901—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.

Table 2: Number of Records Included and Excluded in Analyses

Dataset	**Identifiable Records**	**Unidentifiable Records**
GoSY	2,469	80
15Mar	4,131	229
SRGC	6,151	424
SOHR	45,416	936
SCSR	45,416	2,643
SNHR	46,428	11,045
SS	50,658	14,750
VDC	62,386	7,881
Total	263,055	37,988

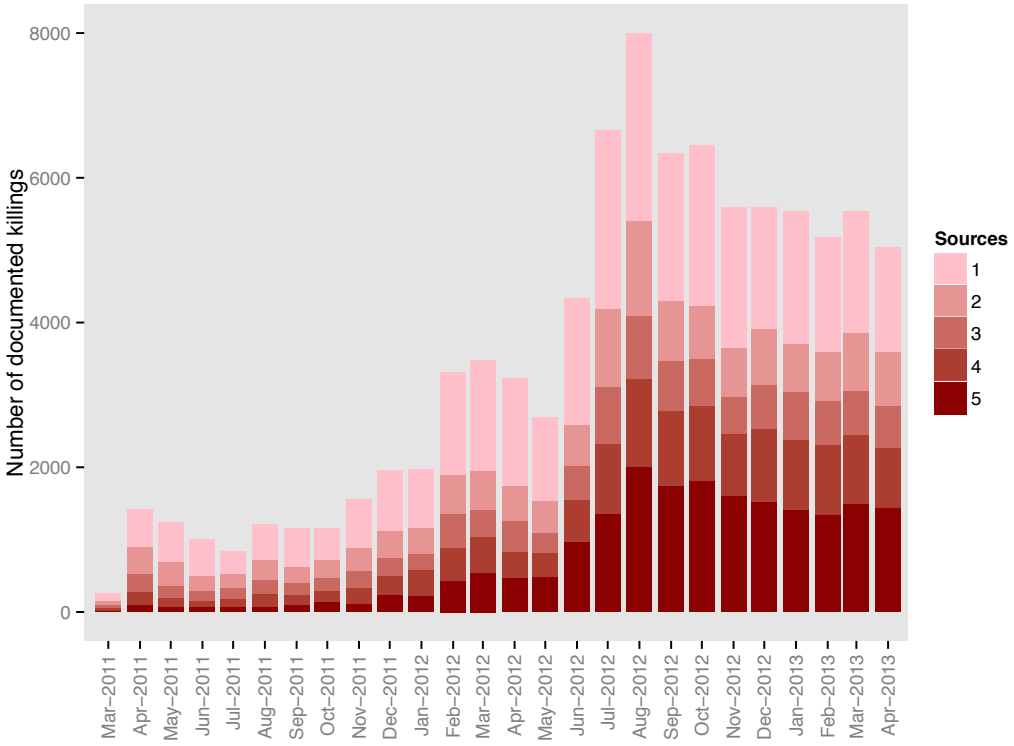
3 Methodology

This report begins with 263,055 records of reported killings of fully identified victims from eight datasets. Many of these records are duplicates. An expert whose native language is Syrian Arabic and who is fluent in English reviewed 14,160 pairs of reported deaths. He classified the reports as either referring to the same victim or to different victims. HRDAG used the expert's classifications with a computer algorithm called an Alternating Decision Tree to build a model to classify the remaining records as either matches or non-matches. The resulting records were merged into a combined dataset which, with duplicates removed, includes 92,901 records of documented killings. More detail on data processing is available in Appendix B, and on matching in Appendices C.1 and C.2.

4 Documentation Patterns over Time

This report began with a warning that despite the enormous efforts by the data collecting groups, many killings in the Syrian Arab Republic are still undocumented. One way to imagine that is to consider that in any particular month, some killings are documented by five groups, other killings are documented by four groups, others 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?

Figure 1: Documented Killings by Month and by Number of Sources per Killing



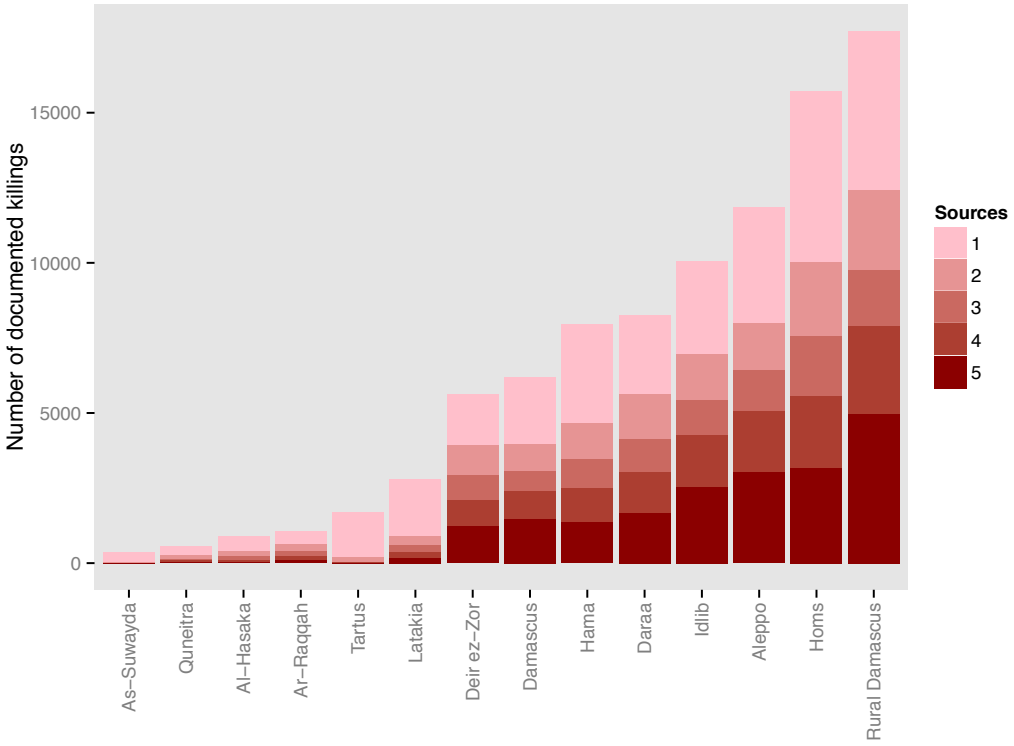
The answer to this question requires statistical modeling and is beyond the scope of this report. Figure 1 provides a way to visualize the intensity of reporting by examining the killings documented by the five datasets that cover the entire period (SCSR, SNHR, SOHR, SS, and VDC). To interpret the graph, compare July 2012 to April 2013. In July, the largest of the sub-bars is light pink, indicating that the largest fraction of documented killings was described by only one of the five sources. By comparison, in April, the number of killings documented by all five groups (in dark red) is approximately equal to the number documented by only one group. 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 there were a sixth group that documented the entire period? Then a seventh, eighth, ninth, 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.

A similar comparison can be drawn by examining the number of sources documenting each killing reported in each governorate (Figure 2). For example, slightly fewer killings are reported in Hama than Daraa. But a larger proportion of killings in Hama are reported by only one group (the light pink section at the top of each bar). Again, the question that is raised is: how many more killings have occurred than have been documented? It is possible that more killings are *occurring* in Hama than Daraa, but more killings are being *documented* in Daraa than Hama. 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, including those that have not been documented.

Lastly, another governorate that stands out in this figure is Tartus. Note that almost all of the documented killings in Tartus are reported by a single group. As Figure 4 in Section 5 will show, the majority of these records are reported by VDC.

Figure 2: Documented Killings by Governorate and by Number of Sources per Killing



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 ethnic group or region, 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.[14] 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 the Syrian Arab Republic 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

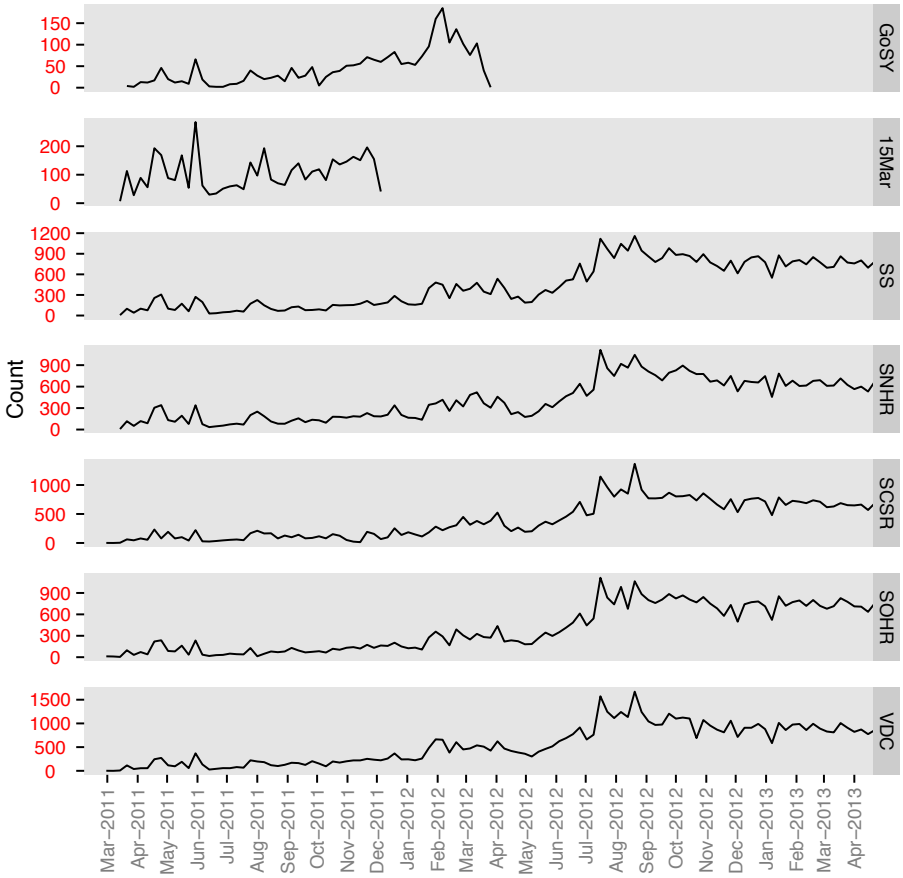
Figure 3 shows the frequency of reported killings by week for each dataset. Five datasets (SCSR, SNHR, SOHR, SS, and VDC) indicate roughly comparable patterns of violence over time. Note, however, that the y-axes are very different. VDC reports the highest number of killings, followed by SS, SCSR, and SNHR and SOHR with similarly-sized peaks of recorded violence.

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

The patterns of violence recorded by the remaining two datasets, 15 Mar and GoSY, are smaller and look different. The pattern shown by 15 Mar approximately tracks SCSR, SNHR, SOHR, SS, and VDC, but the similarity is difficult to see in these graphs because 15 Mar documents so many fewer cases. The variation in 2011 in SCSR, SNHR, SOHR, SS, and VDC is much smaller than the variation in 2012. Because 15 Mar stopped documenting killings in December 2011, its pattern seems different. Data from the Syrian government includes very few records after March 2012 and shows a February 2012 peak that is not found in the other datasets.

Although five of the datasets (SCSR, SNHR, SOHR, SS, and VDC) indicate a substantial increase in documented killings over time, it is important to note that these are recorded killings and this increase may reflect an overall increase in violence or an increase in documentation efforts and therefore in *records* of violence. Alternatively, it may be that documentation has weakened over time, which would mean that violence has increased even more than shown in Figure 3. Because this report includes only the fully-identified reported deaths, it is impossible to rigorously distinguish between these alternatives.

Figure 3: Distribution of Total Reported Deaths by Week; Note that each y-axis is different and SRGC records are included with SNHR.



5.2 Documentation Over Geographic Area

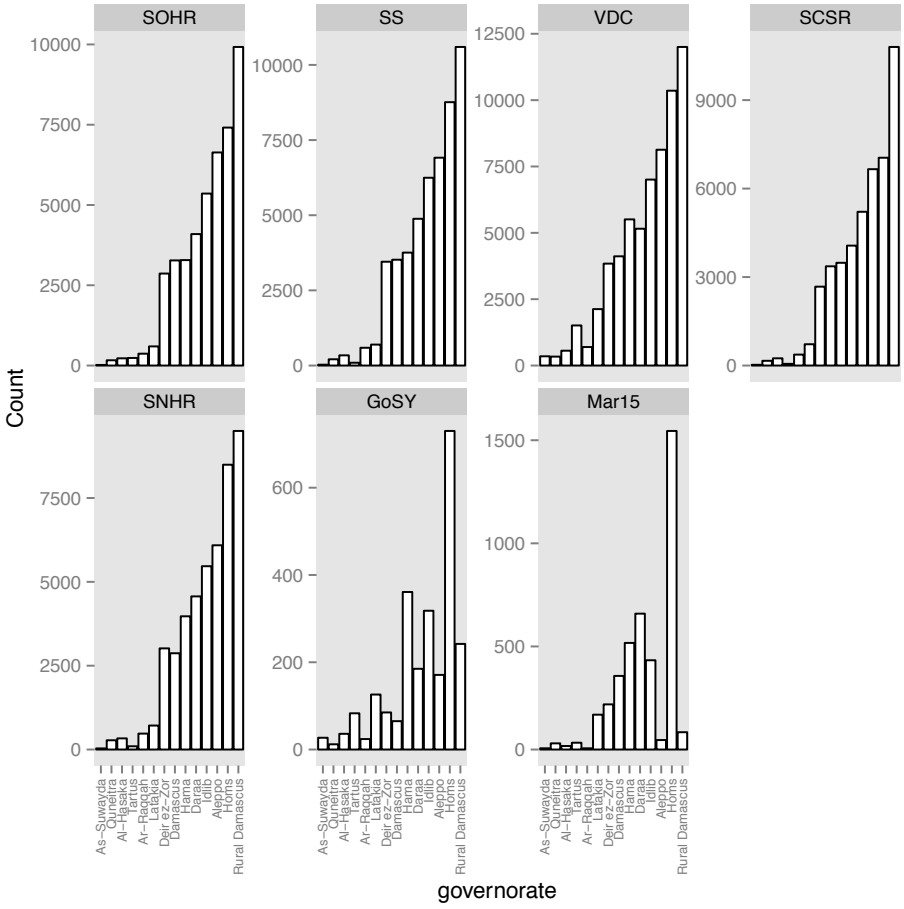
Figure 4 compares patterns of reported violence over geographic area across all seven datasets. Five of these datasets (SOHR, SS, VDC, SCSR, and SNHR) record the highest number of reported killings in Rural Damascus. The other two (GoSY and 15 Mar) record the highest number of killings in Homs. Homs is the second-most frequently reported governorate in the

other five sources, but Rural Damascus is rarely reported in GoSY and 15 Mar.

It is important to note the different limits of each y-axis in Figure 4. VDC reports the highest number of records, and although sharing a general pattern with SOHR, SS, SCSR, and SNHR, the proportion of deaths reported in Daraa and Hama is different in the VDC data. VDC also reports a small peak in Tartus that is not reflected in any of the other datasets, with the exception of GoSY.

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 4: Distribution of Recorded Deaths by Governate; Note that each y-axis is different and SRGC records are included with SNHR.



5.3 Documented Victim Demographic Characteristics

All seven datasets include information about the sex (Table 3) and age of victims (Figure 5). There seems to be general agreement across the datasets that the majority of victims are male. Considering the integrated data, of the 92,901 unique records of documented killings in this report, 82.6% are male victims, 7.6% are female victims, and 9.8% of records do not indicate

the sex of the victim. In future research, it might be possible for an expert familiar with Syrian names to examine the names and infer the sex of these victims.

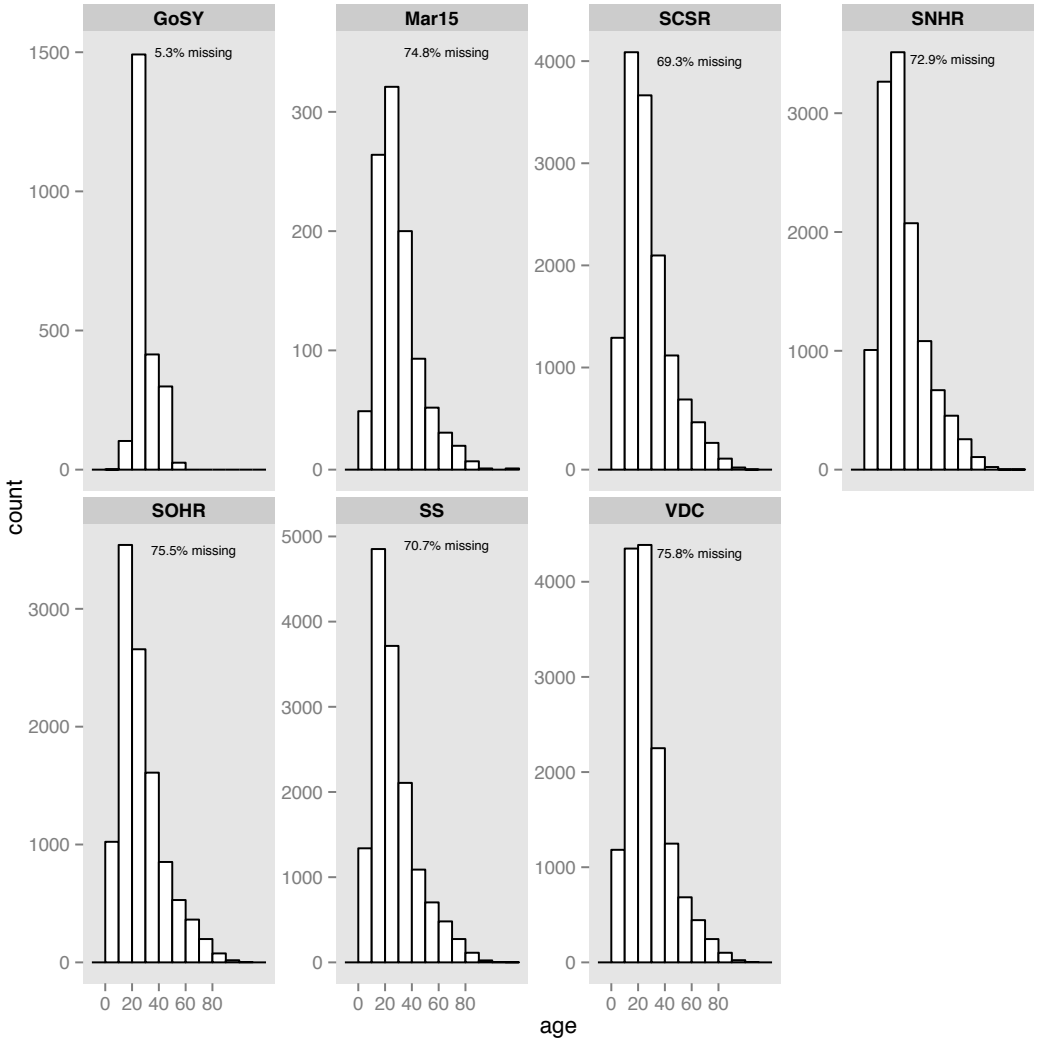
Table 3: Total Documented Killings by Sex (SRGC records are included with SNHR)

Dataset	Female	Male	Unknown
GoSY	0	2,465	0
15 Mar	119	2,446	1,556
SOHR	4,117	33,993	6,340
SNHR	4,336	39,981	1,583
SCSR	4,526	40,366	0
SS	5,020	45,069	0
VDC	5,348	56,344	0

As indicated in Figure 5, these seven datasets indicate a similar *reported* age distribution pattern. While the 15 Mar data has relatively few children less than ten years old, the SCSR, SNHR, SOHR, SS, and VDC datasets show substantial numbers of young children. It could be that more children have been affected in 2012, after the March 15 group stopped their documentation efforts. However, many records are missing indication of age. 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.

Consider the histograms in Figure 5. With the exception of GoSY, the remaining datasets are all missing information on age for nearly 70% 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 or 15 Mar 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 5: Distribution of Recorded Deaths by Age; Note that each y-axis is different and SRGC records are included in SNHR



6 What these Numbers Can and Cannot Tell Us

This report shows that since July 2012, more than 5,000 unique killings have been documented each month, a considerable increase over the sixteen prior months. From the enumeration, several questions remain: does the increase in documented killings reflect an increase in true violence, or does it show an improvement in the documentation groups' capacity to report? Is there really more violence, are groups reporting a greater proportion of the total violence, or both: it could be that violence is increasing while at the same time groups are learning how to do more complete documentation. In either case, the level of violence is extraordinarily high.

The enumeration provided in this report—92,901—is the most accurate accounting available based on identifiable victims reported by these eight 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.[15] Bias means that patterns in the raw data may be misleading regarding the pattern and magnitude of violence occurring in the Syrian Arab Republic. It may be that more violence is occurring in Rural Damascus than Tartus, but it may just be that violence in Tartus 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 July 2012 level. In 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.

[15] For an analysis of event size bias in similar documentation projects in Iraq, see Carpenter et al. (2013).

A Data

A.1 Sources

HRDAG obtained data from the eight sources listed in Section 1 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.

- March 15 Group: This list was provided to HRDAG by OHCHR in February 2012. The group was recommended to OHCHR by the Local Coordination Committees, among others.
- Syrian Government: This list was provided to HRDAG by OHCHR in September 2012.
- Syrian Center for Statistics and Research: HRDAG scraped[16] SCSR's website[17] in May 2013 to obtain a copy of their published data. Individuals can fill out a form on the SCSR website to add victim information. HRDAG established direct contact with SCSR in late May 2013 and in the future will be able to access data directly from SCSR.
- 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. These updates were not shared in time to be included in the data used in this analysis. SNHR maintains a website[18] where they describe that they 'adopt the highest approved documentation principles by the international bodies.'
- Syrian Observatory for Human Rights: This list was provided to HRDAG by OHCHR in December 2012 and again in May 2013. 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

[16] Using a computer program to extract information from websites.
[17] http://csr-sy.org/
[18] http://www.syrianhr.org/

fighters. SOHR maintains a website[19] on which they describe themselves as '. . . a group of people who believe in Human Rights, from inside and outside the country, observing the Human Rights situation in Syria, documenting and criticizing all Human Rights violations, filing reports and spreading it across a broad Human Rights and Media range.' The website also specifies that SOHR '. . . is not associated or linked to any political body.'

- Syrian Revolution General Council: This list was provided to HRDAG by OHCHR in February 2012 along with the description that '. . . a staff of 5 is diligently dedicated to documenting numbers of deaths using different means including visiting families of those killed and contacting mosques and also verifying medical records and in some cases inspection of the body by person when possible.'

- Syrian Shuhada: This list was provided to HRDAG by OHCHR in February 2012. Subsequently HRDAG scraped the website[20] several times in 2012 and 2013 to obtain updated data. It is worth noting that the SS website collects data from several sources, including the Syrian Network for Human Rights (one of the other sources for this analysis). As of 24 May 2013 the SS website reported 5,605 total records from SNHR, about 8% of SS's total database.

- Violation Documentation Centre: This list was provided to OHCHR in February 2012. Subsequently HRDAG scraped the website[21] several times in 2012 and 2013 to obtain updated data. The 'About' page of their website describes the data classification methods and three-stage data verification process implemented by the VDC.

A.2 Potentially Inaccurate Records

Numerous sources have mentioned to HRDAG the possibility of inaccurate records, for example, Starr. 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 and used to model the impact of inaccuracy in the statistical

[19] www.syriahr.com
[20] http://syrianshuhada.com
[21] http://www.vdc-sy.info

analyses. Only a few examples have been shared so far. HRDAG would welcome any examples of potentially inaccurate records so that the records can be examined and removed from the enumeration if verified as inaccurate. More importantly, known inaccurate records can be used to create computer models of potentially inaccurate records to adjust future analyses.

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. Another example includes victims who were believed to be dead but are later discovered to be alive. 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. Although HRDAG is only aware of a small number of specific examples of inaccurate records (those provided by VDC), 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 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

As mentioned at various points throughout this report, the matching and de-duplicating process requires records of identified victims. It also requires that each contributing data source has a similar structure prior to integrating

them into a single list of documented killings. Each data collection organization records slightly different information and organizes that information slightly differently. 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 2012, 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 who reviews training pairs to build the matching model (see Section 3 and Appendix C) confirms the translation of these values. Names are not translated: they remain in the original language of the data source. Names in all the data sources are in Arabic, except in the the 15 Mar list for which the names were transliterated into Latin characters.

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, since each data source records slightly different information (as indicated by Section 5), not to mention each data source is working to overcome the difficulties inherent in collecting complete, accurate information in the midst of a conflict.

C.1 Non-technical matching overview

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. We performed both of these tasks together, by looking for duplicates within a single list of all records from all data sources with sufficient information, including name, and date and location of death. We also used other variables, such as age (or date of birth), sex, and location of birth, for matching.

The records were divided in four groups, called partitions. The first includes data from eight sources (15 Mar, GoSY, SCSR, SNHR, SOHR, SRGC, SS, and VDC) during March to December 2011. The second partition includes seven sources (GoSY, SCSR, SNHR, SOHR, SRGC, SS, and VDC) for January to March 2012. The third partition includes five sources (SCSR, SNHR, SOHR, SS, and VDC) for April to November 2012. The fourth partition includes the same five sources (SCSR, SNHR, SOHR, SS, and VDC) for December 2012 to April 2013. Although these last two partitions include the same data sources, the updated data subsequent to the report published in January was treated as a fourth partition.

From the full set of pooled records, we first identified pairs of records that might be matches using very broad rules. We considered any pair of recorded killings that were reported in the same governorate within one day of one another, and any pair of recorded killings within one week of each other where the names of the two people were similar (with an Arabic edit distance of at most 2). We identified 38.8 million of these "candidate pairs" across all four partitions.

In the next step, we generated detailed comparisons of the two records in each candidate pair and all the training set pairs. These comparisons form a numerical summary of how 'similar' the two records are. We computed eighteen different numerical similarity scores, including:

- Whether the names are the same
- If the names are different, how many words differ between them
- If the names are different, how different the spellings are (edit distance)
- For comparisons of a name recorded in English to a name recorded in Arabic, a phonetic-domain edit distance
- How different the reported ages were
- Whether the records recorded the same sex
- How far apart in time were the two reported deaths
- Whether the two deaths were reported in the same governorate

Then we created a 'training set' consisting of 14,160 record pairs examined in detail by our Syrian expert. He labeled each pair in the training set as referring to the same person (a match) or to different people (a non-match). This data was used to fit a model of how the numerical comparison scores determine whether a pair matches or not. This model was then applied to the full set of candidate pairs to label each of them as matching or non-matching.

We then combined the matched pairs into groups of records which all refer to the same person. 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 we believe refer to the same death. Finally, we merged the records in each group into a single record containing the most precise information available from each of the individual records.

C.2 Matching technical details

Matching databases using partial information has a long history, first formulated by Dunn (1946) and Newcombe et al. (1959), and approached the-

oretically by Fellegi and Sunter (1969).[22] Our method for transliterating and phonetically comparing names written in Arabic and English is based on Freeman et al. (2006).

For matching, HRDAG uses a model-based iterative supervised learning procedure similar to the method described in Sarawagi and Bhamidipaty (2002). In this method, the model is refined iteratively, with the pairs of highest uncertainty under one iteration's model chosen for expert attention to expand the training set used to train the model in the next iteration. For an overview of machine learning techniques for classification and clustering, as well as a description of the software HRDAG used for modeling, the Weka software, version 3-7-4 (Hall et al., 2009), see Witten et al. (2011). The ADTree software is documented at Weka's website. The algorithm for ADT was first described by Freund and Mason (1999) and optimized by Pfahringer et al. (1996).

The model classified all the possible pairs of records from all eight datasets as referring to the same person (a *match*) or to different people (a *non-match*). When tested against the training set, the model classified 86% of the training pairs accurately, averaged across the partitions. The most important test of the matching is whether the model can classify correctly pairs it has never seen before, called 'cross-validation.' In a stratified 10-fold cross-validation, the kappa statistics for the four partitions were 0.79, 0.77, 0.81, and 0.81 respectively.

The matching pairs were organized into larger groups of records that refer to the same person by Hierarchical Agglomerative Clustering (Manning et al., 2008), specifically complete linkage clustering with a hand-tuned stopping distance.

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).

[22] See the reviews of the problem, called variously "record linkage," "matching," and "database deduplication" in Winkler (2006) and Herzog et al. (2007). A key method is approximate string distance (Levenshtein, 1966).

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

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

Freeman, A. T., Condon, S. L., and Ackerman, C. M. (2006). Cross Linguistic Name Matching in English and Arabic: A "One to Many Mapping" Extension of the Levenshtein Edit Distance Algorithm. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 471–478.

Freund, Y. and Mason, L. (1999). The Alternating Decision Tree Learning Algorithm. In *Sixteenth International Conference on Machine Learning, Slovenia*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations Newsletter*, 11:10–18.

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

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10(8):707–710.

Manning, C. D., Raghavan, P., and Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge.

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

Pfahringer, B., Holmes, G., and Kirkby, R. (1996). Optimizing the Induction of Alternating Decision Trees. In *Fifth Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*.

Sarawagi, S. and Bhamidipaty, A. (2002). Interactive Deduplication Using Active Learning. In *KDD '02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 269–278. ACM Press.

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/.

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

Witten, I., Frank, E., and Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufman.

About HRDAG

The Human Rights Data Analysis Group is a non-profit, non-partisan organization[23] 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.

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