Sexual Violence in Armed Conflict Data Project (SVAC) 2.0, 1989-2015
Codebook and Instruction Manual
November 2019

Principal Investigators:

Dara Kay Cohen
Ford Foundation Associate Professor of Public Policy
John F. Kennedy School of Government
Harvard University
Contact email: Dara_Cohen@hks.harvard.edu

Ragnhild Nordås
Assistant Professor of Political Science
University of Michigan
Senior Researcher
Peace Research Institute Oslo (PRIO)
Contact email: rnordas@umich.edu

Research Assistant:

Robert Ulrich Nagel
Postdoctoral Fellow
Georgetown Institute for Women, Peace and Security (GIWPS)
Contact email: nagel.robert@me.com

Last Updated: October 23, 2019

Table of Contents

1. Background and Acknowledgements

The original Sexual Violence in Armed Conflict (SVAC) Dataset was conceptualized in 2009 by researchers affiliated with the Centre for the Study of Civil War (CSCW) at the Peace Research Institute Oslo (PRIO), including Inger Skjelsbæk, Ragnhild Nordås, Scott Gates, and Dara Kay Cohen. The core team initially benefited from advice and feedback on the data collection from a consultative group of experts, including Mia Bloom, Christopher Butler, Amelia Hoover Green, Michele Leiby, Gudrun Østby, Håvard Strand, and Elisabeth Wood.

This project was made possible through a series of generous grants. The pilot project grant was funded by the Norwegian Ministry of Foreign Affairs. The data collection phases of the initial project were funded by a grant from the Folke Bernadotte Academy, and by a grant from the National Science Foundation (SES-1123964), and the Norwegian Research Council. The current update of the dataset, SVAC 2.0, was funded by a grant from the Norwegian Research Council as well as the Belfer Center for Science and International Affairs at the Harvard Kennedy School.

This Coding Manual is both a codebook and an instruction guide for coders.

2. Scope and Definitions of SVAC Data Project

The SVAC Dataset covers conflict-related sexual violence committed by the following types of armed conflict actors: government/state military, pro-government militias, and rebel/insurgent forces between 1989-2009 (SVAC 1.0), and government/state military and rebel/insurgent forces for 2010-2015 (SVAC 2.0).[1] We include only sexual violence by armed groups against individuals outside their own organization.

All actors listed in the SVAC 2.0 Dataset are involved in state-based conflicts as defined by the UCDP/PRIO Armed Conflict Database. Peacekeeper and civilian perpetrators are not included as actors in the dataset. We also do not include non-state actors (both rebel groups and PGMs) involved in violence that is not part of a conflict with the government.[2]

The original SVAC Dataset covered armed conflict active in the years 1989-2009, as defined by the UCDP/PRIO Armed Conflict Database. The SVAC 2.0 update adds the years 2010-2015. Therefore, the SVAC 2.0 Dataset includes all conflicts active in the years 1989-2015. The additional six-year period includes 92 conflicts in 49 countries that were either active or within five years of cessation. We collected data for all years of active conflict (defined by 25 battle deaths or more per year) and for the five years post-conflict. Beyond this post-conflict time period, "peacetime" sexual violence is outside the scope of the project. We also code "interim" years; see definition of this variable below.

[1] Pro-government militias (PGMs) are only included for 1989-2009 and not for 2010-2015 due to data availability of the years in the Pro-Government Militia Database (which ends in 2009).

[2] For example, the Liberian rebel group United Liberation Movement of Liberia for Democracy (ULIMO) is neither part of the original SVAC Dataset nor the SVAC 2.0 Dataset because ULIMO was involved in a non-state conflict with the National Patriotic Front of Liberia (NPFL) rather than a state-based conflict with the Liberian government. Similarly, non-state actors that are listed in the UCDP One-Sided Violence (OSV) Dataset might not be included in SVAC 2.0 Dataset if these actors are not part of a state-based conflict. We suspect—but have not confirmed—that many of the PGMs that we were unable to code for the 2010-2015 period are included as actors in the OSV dataset.

Conflict manuscripts, which contain additional information in the form of narratives, as well as the source information used for coding each conflict, can be found in the accompanying documentation to the SVAC dataset. Conflict manuscripts can be requested from the Principal Investigators.

Based on feedback from scholars and analysts who have used the SVAC data, we made two important decisions about the structure and coding of the SVAC 2.0 dataset:

1. **Active conflict-years dataset**: To facilitate the use of the SVAC data, we created versions of the dataset. The complete version (SVAC_complete_1989-2015) conserves the original dataset structure by including interim years and the first five post-conflict years. A second, smaller dataset (SVAC_conflictyears_1989-2015) only includes active conflict years. Because most analyses that have used the data thus far are primarily focused on sexual violence in *active conflict-years*, we created this smaller dataset that can be readily merged with other UCDP datasets. This dataset does not require the analyst to drop observations from the full dataset.

2. **Prevalence score and form variables**: Scholars' use of the original dataset indicates that the three prevalence scores and the form variable are by far the most widely used variables. Given time and budget constraints, the SVAC 2.0 dataset contains only these variables. Other variables in the original dataset, including text variables, (e.g. location; timing) were not updated.

3. Definitions

Unit of Observation

The unit of observation for the SVAC dataset is the *conflict-actor-year*: a particular conflict actor in a given calendar-year (e.g. conflict 118, LRA, 2001). Conflicts and actors are defined in the following paragraphs.

Conflicts

The SVAC 2.0 dataset includes all active armed conflicts in the period 1989-2015, as defined by the UCDP/PRIO Armed Conflict database (Gleditsch et al. 2002) and the UCDP Dyadic Dataset Version 16.1 (Harbom, Melander & Wallensteen 2008). We only include conflicts that have been active in one or more of the years 1989-2015. An armed conflict is defined as "a contested incompatibility that concerns government and/or territory where the use of armed force between two parties, of which at least one is the government of a state, results in at least 25 battle-related deaths" (Gleditsch et al. 2002).

We also code "interim years," defined as conflict-actor-years that do not reach the 25 battle-related deaths threshold if the observation in question is less than 5 years after an observation that reaches the battle-related deaths criterion. For example, if a rebel group was active in 1993, 1994, and 1996, we also code for any sexual violence that occurred in 1995.

Finally, we include the five conflict-actor-years after the last year that a conflict actor has been deemed active (i.e. involved in violence resulting in deaths surpassing the threshold of 25 battle-related deaths). Using the previous example, if a rebel group was active in 1993, 1994, and 1996, we also code whether any sexual violence was perpetrated by the group in the five years after the final active year, i.e. 1997, 1998, 1999, 2000, and 2001.

The UCDP/PRIO includes three types of conflict, which are also included in SVAC 2.0: (1) *Intrastate armed conflict*, which occurs between the government of a state and one or more internal opposition group(s) without intervention from other states; (2) *Internationalized internal armed conflict*, which occurs between the government of a state and one or more internal opposition group(s) with intervention from other states (secondary parties) on one or both sides; and (3) *Interstate conflicts* (between the governments of two states).

Actors

The SVAC dataset includes the actors present in armed conflicts as conflict parties according to the UCDP/PRIO data and the UCDP Dyadic Dataset (Harbom, Melander & Wallensteen, 2009; Harbom & Wallensteen, 2010). In SVAC 2.0 we used the UCDP dyadic dataset v. 1-2015 to establish the relevant conflict dyads for the update. We include in our dataset all government/state (Side A) and rebel/insurgent (Side B) actors as well as all other state actors (Side A2 and Side B2) in all conflict years that reached a threshold of 25 battle-related deaths. For the years 1989-2009, we include pro-government militias as given by the Pro-Government Militia Database (Carey, Mitchell and Lowe 2013). We do not include pro-government militias in the years 2010-2015, as the Carey et al. dataset ends in 2009.

Sometimes government/state actors with special status will not be specifically named in the dataset; examples include special police, special units, treasury police, presidential guards, presidential units, and security forces. We include all government actors with special status as representatives of the state (unless that actor has been previously assigned a separate ID code). Activities by actors such as domestic police, interrogators, border patrol, border police, and checkpoint police are coded as committed by the government/state side (Side A) if coders find explicit evidence that the sexual violence is conflict-related and/or directed at an insurgent or suspected member of an insurgent group, a close relative of a member of an insurgent group, and/or undertaken for the purpose of collecting intelligence related to the conflict. Additionally, in cases where the incident of sexual violence is perpetrated in a conflict territory, such as at a border or a checkpoint in a clearly defined conflict area, the incident of sexual violence perpetrated by one of the aforementioned actors is considered conflict-related.

Sexual Violence

Following the definition used by the International Criminal Court (ICC)[3], we use a definition of crimes of sexual violence which includes (1) rape,[4] (2) sexual slavery,[5] (3) forced prostitution,[6] (4) forced pregnancy,[7] and (5) forced sterilization/abortion.[8] Following Elisabeth Wood (2009), we also include (6) sexual mutilation,[9] and (7) sexual torture.[10] This definition does not exclude the existence of female perpetrators and male victims. We focus on behaviors that involve direct force and/or physical violence. We exclude acts that do not go beyond verbal sexual harassment and abuse, including sexualized insults or verbal humiliation.

4. Variables

For compatibility and ease of integration with widely used existing datasets, we include a number of general variables on region, country, year, actor ID, type of actor, and conflict ID mostly from the UCDP/PRIO data.

We use a monadic conflict-actor-year data structure; we rejected a dyadic structure because many of the victims of sexual violence in armed conflicts are civilians, which does not lend itself easily to a dyadic logic. By including variables with dyad ID and conflict ID, however, analysts may create a dyadic structure.

---

[3] International Criminal Court, Elements of Crimes, U.N. Doc. PCNICC/2000/1/Add.2 (2000). Article 8 (2)(e). Last downloaded, 15 November 2011 from: http://wfrt.net/humanrts/instree/iccelementsofcrimes.html#_ftn64

[4] Rape is defined as: The perpetrator invaded the body of a person by conduct resulting in penetration, however slight, of any part of the body of the victim or of the perpetrator with a sexual organ, or of the anal or genital opening of the victim with any object or any other part of the body. The invasion was committed by force, or by threat of force or coercion, such as that caused by fear of violence, duress, detention, psychological oppression or abuse of power, against such person or another person, or by taking advantage of a coercive environment, or the invasion was committed against a person incapable of giving genuine consent.

[5] The perpetrator exercised any or all of the powers attaching to the right of ownership over one or more persons, such as by purchasing, selling, lending or bartering such a person or persons, or by imposing on them a similar deprivation of liberty. The perpetrator caused such person or persons to engage in one or more acts of a sexual nature.

[6] The perpetrator or another person obtained or expected to obtain pecuniary or other advantage in exchange for or in connection with the acts of a sexual nature.

[7] The perpetrator confined one or more women forcibly made pregnant, with the intent of affecting the ethnic composition of any population or carrying out other grave violations of international law.

[8] The perpetrator deprived one or more persons of biological reproductive capacity.

[9] Permanent disfiguration, including but not limited to cutting/severing of breasts or genitals. The conduct caused death or seriously endangered the physical or mental health of such person or persons.

[10] In general, "torture means any act by which severe pain or suffering, whether physical or mental, is intentionally inflicted on a person for such purposes as obtaining from him or a third person information or a confession, punishing him for an act he or a third person has committed or is suspected of having committed, or intimidating or coercing him or a third person, or for any reason based on discrimination of any kind, when such pain or suffering is inflicted by or at the instigation of or with the consent or acquiescence of a public official or other person acting in an official capacity." (UN Convention against torture: http://www.hrweb.org/legal/cat.html). In the SVAC project, we also code torture committed by non-state actors.

Below we present the general variables of the dataset (not pertaining to sexual violence), followed by the sexual violence variables and how they are coded.

General Variables

Variable Name	Source	Description
year		Year
actor	UCDP/PRIO	Name of country if the actor is a government, otherwise name of organization if this is a rebel organization or militia
actorid	UCDP/PRIO	Old UCDP Non-State Actor ID (before version 17.1)
actorid_new	UCDP/PRIO	In version 17.1 of all UCDP data ID the system for conflicts, actors and dyads was changed in order to make them unique across all UCDP core datasets and all UCDP types of violence.
actor_type	SVAC	A coding for the type of actor. More specifically, we employ the following scheme: 1: State or incumbent government (in UCDP dyadic, this actor type is called 'Side A') 2: State A2 (in UCDP dyadic, this actor type is called 'Side A2nd'). These are states supporting the state (1) involved with conflict on its territory. 3: Rebel (in UCDP dyadic, the actor type is called 'Side B') 4: State supporting 'Side B' in other country (in UCDP dyadic, this actor type is called 'SideB2nd'). 5: Second state in interstate conflict (in UCDP dyadic, this actor is called 'Side B'). 6: Pro-government militias (PGMs)
conflictid_old	UCDP/PRIO	Old UCDP/PRIO Conflict ID (before version 17.1)
conflictid_new	UCDP/PRIO	New UCDP/PRIO Conflict ID (starting with version 17.1 of UCDP data)
type	UCDP/PRIO	Nominal variable with three categories: 2: Interstate Conflict 3: Intrastate Conflict 4: Internationalized Internal Armed Conflict
Incompatibility	UCDP/PRIO	A general coding of the conflict issue 1: Territory 2: Government 3: Government and territory
Region	UCDP/PRIO	Numerical coding of the geographical region

Location	UCDP/PRIO	The name(s) of the country/countries of fighting and whose government(s) have a primary claim to the territory in dispute.
gwnoloc	Gleditsch/Ward	Gleditsch/Ward country ID of location variable
conflictyear	UCDP/PRIO	Dummy indicating active conflict-year. See below for a detailed description of each year type.
interm	UCDP/PRIO	Dummy indicating an interim conflict-year. See below for a detailed description of each year type.
postc	UCDP/PRIO	Dummy indicating a post-conflict-year. See below for a detailed description of each year type.

Sexual Violence Variables

The sexual violence variables aim to capture data on two dimensions, prevalence and form.

(1) Prevalence

The prevalence measure gives an estimate of the relative magnitude of sexual violence perpetration was by the conflict actor in the particular year. This is coded according to an ordinal scale, adapted from Cohen (2010; 2016) and discussed in Cohen and Nordås (2014). Note that the coding is primarily based on the qualitative description; only secondarily do we rely on a count of estimated incidents. The SVAC dataset cannot be used as a means to estimate the numbers of victims.

Prevalence = 3 (Massive) Sexual violence is likely related to the conflict, and:

- Sexual violence was described as "systematic" or "massive" or "innumerable"
- Actor used sexual violence as a "means of intimidation," "instrument of control and punishment," "weapon," "tactic to terrorize the population," "terror tactic," "tool of war," on a "massive scale"

Note: Absent these or similar terms, a count of 1000 or more reports of sexual violence indicates a prevalence code of 3.

Prevalence = 2 (Numerous) Sexual violence is likely related to the conflict, but did not meet the requirements for a 3 coding, and:

- Sexual violence was described as "widespread," "common," "commonplace," "extensive," "frequent," "often," "persistent," "recurring," a "pattern," a "common pattern," or a "spree"
- Sexual violence occurred "commonly," "frequently," "in large numbers," "periodically," "regularly," "routinely," "widely," or on a "number of occasions;" there were "many" or "numerous instances"

Note: Absent these or similar terms, a count of 25-999 reports of sexual violence indicates a prevalence code of 2.

Prevalence = 1 (Isolated) Sexual violence is likely related to the conflict, but did not meet the requirements for a 2 or 3 coding, and:

- There were "reports," "isolated reports," or "there continued to be reports" of occurrences of sexual violence

Note: Absent these or similar terms, a count of less than 25 reports of sexual violence indicates a prevalence code of 1.

Prevalence = 0 (None) Report issued, but no mention of rape or other sexual violence related to the conflict

Note: For example, a coder finds a report covering a country in a given year but within the report there is no mention of rape or other sexual violence related to the conflict.

Prevalence = -99 (BOTH No Report AND No Information) No report found and no data available from subsequent years, and consequentially no data. This code should be used as infrequently as possible.

Note: For example, if a coder finds no HRW or AI annual report and no special report for a conflict-actor-year, this is given a code -99.

Prevalence disaggregated by source

Prevalence scores are coded separately for three different sources in the variables "Prev_ State," "Prev_ HRW," and "Prev_ AI." "Prev_ State" scores are assigned using information from US State Department annual reports. "Prev_ HRW" scores are assigned using information from Human Rights Watch annual and special reports. "Prev_ AI" scores are assigned using information from Amnesty International annual and special reports.[11]

There are two important conventions for coding prevalence scores: First, in some cases, a coder may find evidence in a report that supports multiple prevalence scores. For example, in one section of the report, sexual violence is described using a keyword such as "reports" while in another section of the report sexual violence is described using a keyword such as "numerous." When evidence exists for coding prevalence = 1 (based on "reports") and coding prevalence = 2 (based on "numerous"), coders chose the highest prevalence score supported by the evidence.

Second, in some cases, a coder may find conflicting text and numerical evidence in a report. For example, in one section of the report sexual violence is described numerically as "under 25" while in another section of the report sexual violence is described using a keyword such as "widespread." When disagreement exists between numerical evidence and qualitative keyword evidence (text), coders should base coding decisions on keyword evidence (text). In the aforementioned example, a coder would code prevalence = 2 (based on "widespread").

[11] When both annual and special reports exist, there should be score agreement between the reports. If agreement does not exist, coders informed one of the principal investigators to adjudicate the disagreement.

(2) Form

Nominal, numerical variable listing the forms of conflict-related sexual violence committed by the armed conflict actor. Coders list all forms of conflict-related sexual violence including (and limited to):

1 Rape
2 Sexual Slavery
3 Forced Prostitution
4 Forced Pregnancy
5 Forced Sterilization/Abortion
6 Sexual Mutilation
7 Sexual Torture

Note: Coder should only include forms of sexual violence committed by actors included in the dataset (i.e. not sexual violence forms by actors that are *not* defined as actors in the dataset). Sexual abuse and sexual molestation are considered forms of sexual torture. The form variables are not mutually exclusive, as there can be various types of sexual violence committed in a conflict-actor-year.

A forthcoming dataset by Logan Dumaine, Professor Elisabeth Jean Wood, Professor Ragnhild Nordås, and Maria Gargiulo builds on these data to compile information about the seven different forms of sexual violence perpetrated by conflict actors in armed conflicts between 1989 and 2015, including a measure of reported prevalence for each individual form. That dataset is called the *Repertoires of Sexual Violence in Armed Conflict Dataset* (RSVAC).

Year Variables

(1) Active Conflict Years (conflictyear)

This variable is coded 1 for all years where the observation (conflict-actor-year) is in an active conflict (reaching the UCDP battle-deaths threshold in the particular year), and 0 otherwise**.**

(2) Interim Years (interim)

Interim years have been added to the dataset, and they are, therefore, observations that are not in the UCDP dyadic dataset but follow logically from that dataset. These are observations (actor-years) where there has been 1, 2, 3, or 4 years of inactivity in the dyad and then the dyad becomes active again. All observations that have been added to the SVAC dataset using this rule have the value 1 for the "interim" variable and 0 otherwise.

(3) Post conflict years (postc)

Post-conflict years are actor-years for the five years after the last year the dyadID is included in the UCDP dyadic dataset. These observations are coded 1 for postc, and 0 otherwise.

Note: The post-conflict logic is based on dyads, not entire conflicts. For example, suppose that a Conflict ID involves two active dyads: State A fights rebels X (dyad 1) and rebels Y (dyad 2). Dyad 1 is active in 1990, 1991, and 1992. Dyad 2 is active in 1991, 1992, 1993, 1994. The conflict does not reignite. In this case, the state is in active conflict in years 1990, 1991, 1992, 1993, and 1994, and is post-conflict in years 1995, 1996, 1997, 1998, and 1999. Rebels X are post-conflict (postc=1) in 1993, 1994, 1995, 1996, 1997, and rebels Y are post-conflict (postc=1) in 1995, 1996, 1997, 1998, 1999 (i.e. in the five years after their last respective active dyad year). In conflict dyads that

ended in the last five years of the dataset, we let the "postc" logic trump the "interim" logic until we have information that a conflict has reignited.

5. Sources and Data Collection Strategy

Our data collection strategy relies on the most commonly used sources in the quantitative human rights literature: *United States State Department* annual reports, *Amnesty International* annual and special reports; and *Human Rights Watch* annual and special reports.

State Department reports are published annually and are called "Country Reports on Human Rights Practices." The reports are published during the spring following the calendar year covered in the reporting. For example, the 2010 Country Report on Human Rights Practices is published in April 2011 and covers the period January 2010 through December 2010. State Department reports are available online at http://www.state.gov/g/drl/rls/hrrpt/ for calendar years 1999 – 2016. Older reports can be accessed online through http://www.unhcr.org/refworld (search by publisher) or through www.heinonline.org.

Amnesty International (AI) publishes two types of reports that are used as sources for the project. They publish an annual report called "Annual Report: The State of the World's Human Rights." Within the annual report, one can search for general reports, country reports, and special (topical) reports. Both types of reports are available online at https://www.amnesty.org/en/ for the periods 2007-2016. Note that there are no annual reports available for 2013 as AI changed its publishing cycle. The annual report 2013 published in May 2013 includes events for the calendar year 2012. The next available report is the annual report 2014/2015 published in February 2015 reporting events of 2014. AI also publishes "News and Publications," including special reports by country and/ or by human rights topic. Coders review annual and special reports and include data from both resources in Conflict Manuscripts and coding sheets.[12,13]

Human Rights Watch (HRW) publishes a variety of reports that are used as sources for the project. Annual reports are called "World Reports" and are available online at http://www.hrw.org/en/node/79288 for the period 1989-2016. The reports are organized by country. HRW also publishes special reports organized by human rights issue and/or country. Special reports are available on the HRW website and can be located using the report search function. As with AI, coders should review annual and special reports and include data sourced from both resources in conflict manuscripts and coding sheets.

6. Data Reliability Measures

To ensure high quality, reliable data collection and coding, the team discussed ambiguous cases and coding rules, as well as any issues related to data collection, data coding, data format, project scope, or necessary adjustments to the Coding Manual. To further increase transparency and

[12] AI publishes annual reports for most countries in most years and special reports for few countries in most years. One should also note that special reports often contain data for multiple years and sometimes multiple conflicts and/ or actors. Special reports are sometimes lengthy but should be reviewed carefully as many of the reports have quality data for SVAC variables of interest.

[13] Coders noted in the Country Manuscript any years where AI annual and special reports report conflicting information about the variable "Prevalence".

information flow, the core project team and research assistants use web-based document sharing software.

Intercoder Reliability Tests

Before the release of SVAC 2.0, the team reviewed the newly collected 1,827 observations from 2010-2015. We found high intercoder reliability (kappa value > .81) for the 55% of observations that two independent coders had coded. Results of intercoder reliability testing are available upon request.

7. Frequently Asked Questions

(1) Sources

"Why does the project limit sources to State Department, Amnesty International, and Human Rights Watch?"

First, the SVAC data project relies on sources that publish credible human rights reports covering each year and location included in the dataset.[14] By collecting data from sources that are publicly available for each year included in the dataset, we are able to build a comprehensive dataset with a limited number of missing values due to a lack of reporting coverage. Our strategy also limits the introduction of data biases associated with the availability of reporting for some countries in some years and not others (for additional discussion of potential biases, see Cohen and Nordås 2014). Additionally, our strategy allows us to capture data from both low and high prevalence years and better illustrate variance within countries between periods.

Second, pilot studies suggested that including a more comprehensive set of sources did not yield enough additional codeable data to warrant the large additional number of research hours required.

"Why not also include data from health surveys or other data projects focused on gender or violence against women?"

There are several other data projects that have collected related data, including WomanStats (http://womanstats.org), Gender-Based Violence Information Management System (http://gbvims.org), and the Demographic and Health Surveys (http://www.measuredhs.com).

WomanStats is a comprehensive compilation of information on the status of women in the world. The project collects data on variables relevant to the SVAC data project, such as the "physical security of women" scale and includes variables that capture the existence and enforcement of laws on rape and sexual violence. While we recognize WomanStats as a comprehensive resource covering a variety of topics related to the security of women and girls, the data available are not appropriate for the SVAC project for the following reasons:

- WomanStats data is not available for each country/ year covered in the SVAC dataset.

[14] Primary sources typically publish reports covering all locations/ conflict years in the dataset but on occasion skip a location/ conflict year usually due to the publication of a special report or due to a crisis in country that limits the organization's access.

- WomanStats metrics that are applicable to the SVAC project cover the existence/ enforcement of laws protecting the physical security of women and cover domestic sexual violence against women. The SVAC project collects data covering conflict-related sexual violence. To be included in the SVAC dataset, the data must be traceable to the conflict-actor-year level.
- WomanStats metrics are focused on women and girls, while the SVAC project includes data describing sexual violence perpetrated against women, girls, men, and boys.

The GBVIMS was created to harmonize data collection on GBV in humanitarian settings, to provide a simple system for GBV project managers to collect, store and analyze their data, and to enable the safe and ethical sharing of reported GBV incident data. The intention of the GBVIMS is to assist service providers to better understand the GBV cases being reported as well as to enable actors to share data internally across project sites and externally with agencies for broader trends analysis and improved GBV coordination. The primary service provided by the system is data compilation and statistical analysis (data is focused on incident details, survivors, and to a lesser extent perpetrators). At the time of coding the SVAC dataset, GBVIMS was active in 27 countries: Burundi, Chad, Central African Republic, Colombia, Côte d'Ivoire, Democratic Republic of Congo, Ethiopia, Greece, Guinea, Haiti, Iraq, Jordan, Kenya, Lebanon, Liberia, Mali, Nepal, Niger, Nigeria, Pakistan, Philippines, Sierra Leone, South Sudan, Tanzania, Thailand, Uganda, and Yemen. The SVAC project is focused on a much wider universe of cases than the GBVIMS.

Demographic and Health Surveys, and other similar projects, provide rich population, health, and nutrition data about populations/ countries. However, most health survey data are not appropriate for the SVAC data project. For example, the DHS collects data on women's empowerment and status for over 70 countries but covers gender-based violence primarily in the context of domestic violence. The main limitation is however that the DHS surveys do not provide information about perpetrators, and the reported gender-based violence is therefore not codeable in a conflict-actor-year setup. An additional constraint of health-based surveys (specific to the SVAC project) is that data are usually only collected periodically and for a limited set of countries.

(2) Methodology

The following questions are summaries of coding instructions provided to individual coders given frequently asked questions.

"Can coders use keyword searches to quickly identify codeable information?"

Keyword searches are an effective way to quickly identify potentially data rich areas of long reports. The following is a list of commonly used keywords:

Rap*; Sex*; Mutil*; Sodom*; Abus*; Castra*; Slave*; Forced; Steril*; Traffic*; Prostit*; Molest*; Breast; Genit*; Anus; Testic*; Groin; Penis; Vagina; Rectum; Wife; Wive*; Girl; Detain*

One cautionary note is that coders cannot rely on keyword searches alone to identify potential data. A best practice methodology for reviewing reports is to begin reviewing text by searching for keywords and then carefully reading sentences with keywords and adjacent text. It is sometimes necessary to read several paragraphs before and after the keyword to collect all relevant data and to understand the context of the sexual violence. It should also be noted that keyword searches are not always an effective methodology when reviewing special reports, especially when special

reports contain a high quantity of codeable data. In these circumstances, it is likely a more productive strategy to skim reports to locate potential data.

"How are Conflict Manuscripts organized?"

All Conflict Manuscripts are organized by source-year and contain the following:

- searchable headers for source and year (i.e. **State Department** 2010)
- supporting documentation (including direct quotations from sources)
- research assistant comments that explain the coding decision and logic in terms of identifying actor, form of sexual violence, and prevalence for each included quotation.

"What should I do if I find documented sexual violence by an actor not included in the UCDP/PRIO data?"

Coders should include the reported sexual violence in the Country Manuscript but not add rows to the dataset. The current practice is to add data to the Manuscript under a searchable header at the end of the relevant conflict- year.

If a researcher discovers a pattern of multiple years of sexual violence by an armed conflict actor not in the dataset, the researcher should inform one of the principle investigators to confirm the exclusion (or possibly inclusion) of the armed conflict actor.

Coders should also check that an armed conflict actor is not identified in the dataset by an alternate name.

It is also possible that a particular actor has been included in the UCDP dyadic dataset in more recent versions than what was used to generate the SVAC dataset. To check for this, the UCDP documentation of known revisions and errata can be consulted. However, the SVAC dataset is not updated annually and the UCDP dataset is, so some such discrepancies could occur. Researchers using the SVAC data should handle such instances at their own discretion should they arise.

"What should I do if I find documented sexual violence by an active armed actor but the year is not part of the dataset?"

If the actor is included in the dataset but the year is not included in the dataset, coders should include data in the Manuscript but not add rows to the dataset.

"To help capture all years of the conflict including the 5 years post-conflict, should coders add conflict-years to the dataset?"

The 5 years post battle-related activity should be included in the dataset already. It should therefore not be necessary to add lines. If coders suspect that there is a specific mistake, they should bring it to the attention of one of the principle investigators.

(3) Prevalence of Sexual Violence

"How do coders determine the magnitude of prevalence in the absence of a numerical estimate of the number of victims?"

In some cases, coders will observe data with few and/or unrecognized keywords. When a coder encounters a difficult case, s/he should consult one of the principal investigators.

"Can I code data that describes sexual violence over a period of years. For example, a report states that an armed conflict actor has kidnapped and sexually abused girls for the past two decades."

General descriptions like the example above are not codeable on the level of the conflict-actor-year. It is often not clear if the description means literally sexual violence at the same prevalence level for 20 consecutive years, or if it is more of a rhetorical devise. However, if the sexual violence is reported for a specific range it is codeable. If a report stated, for example, that an actor had kidnapped and sexually abused girls from 1992-1995, this can and should be coded.

(4) Forms of Sexual Violence

"What should coders do if they find evidence of sexual violence that does not fit into the categories of rape, sexual mutilation, sexual slavery, forced prostitution, forced pregnancy, forced sterilization/ abortion, or sexual torture?"

If coders observe data that describes sexual violence that they consider does not fit into the SVAC definition of sexual violence, they should bring it to the attention of one of the principal investigators.

"How is sexual violence that occurs post-mortem coded?"

In general, we do not code post-mortem sexual violence. In ambiguous cases, coders should describe the event in the Conflict Manuscript but not include the data in the coding sheet (i.e. do not code the data).

"Is it considered sexual torture when a victim is stripped naked and tortured but the victim suffers no physical harm to his or her sexual organs?"

For the purposes of the SVAC data project, stripping a victim naked is not coded as sexual violence, but penetration and/or physical harm to sexual organs is.

"How are threats of rape coded?"

Threats of rape or any other form of threatened sexual violence are not included in the project's definition of sexual violence. Coders should code actualized sexual violence meeting the SVAC definition.

(5) Changes in the updated dataset

"Why are there fewer variables?"

The updated dataset is restricted to the core variables of interest, i.e. reported prevalence and forms of sexual violence. The updated version no longer includes the following variables of the original dataset: pgm_id, selection, selection_ethnicity, selection_nationality, selection_religion, selection_age, selection_actor, selection_other, male, child, detainee, refugee, timing, timing_month, timing_military, timing_political, timing_errands, timing_search, location_text, location_camp, location_checkpoint, location_detention, location_private, location_school, public_public, public_semipublic, public_private, witness_family, witness_victims, witness_soldiers, witness_other, gang, byproxy.

"Why are there different datasets available?"

To facilitate the use of the SVAC data we created two datasets. The complete version (SVAC_complete_1989-2015) conserves the original dataset structure by including interim years and the first five post-conflict years. The smaller dataset (SVAC_conflictyears_1989-2015) only includes active conflict years. As most analysts are primarily interested in the role of sexual violence in active conflict-years, this smaller dataset should facilitate the use of the data.