Big Data, Selection Bias, and the Statistical Patterns of Mortality in Conflict

Megan Price, Patrick Ball

SAIS Review of International Affairs, Volume 34, Number 1, Winter-Spring 2014, pp. 9-20 (Article)

Published by The Johns Hopkins University Press
DOI: 10.1353/sais.2014.0010

SAIS Review vol. XXXVI no. 1 (Winter–Spring 2014)

Big Data, Selection Bias, and the Statistical Patterns of Mortality in Conflict

Megan Price and Patrick Ball

The notion of "big data" implies very specific technical assumptions. The tools that have made big data immensely powerful in the private sector depend on having all (or nearly all) of the possible data. In our experience, these technical assumptions are rarely met with data about the policy and social world. This paper explores how information is generated about killings in conflict, and how the process of information generation shapes the statistical patterns in the observed data. Using case studies from Syria and Iraq, we highlight the ways in which bias in the observed data could mislead policy. The paper closes with recommendations about the use of data and analysis in the development of policy.

Introduction

Emerging technology has greatly increased the amount and availability of data in a wide variety of fields. In particular, the notion of "big data" has gained popularity in a number of business and industry applications, enabling companies to track products, measure marketing results, and in some cases, successfully predict customer behavior.[1] These successes have, understandably, led to excitement about the potential to apply these methods in an increasing number of disciplines.

Megan Price is the director of research at the Human Rights Data Analysis Group. She has conducted data analyses for projects in a number of locales including Syria and Guatemala. She recently served as the lead statistician and head author of two reports commissioned by the Office of the United Nations High Commissioner of Human Rights.

Patrick Ball is the executive director of the Human Rights Data Analysis Group. Beginning in El Salvador in 1991, Patrick has designed technology and conducted quantitative analyses for truth commissions, non-governmental organizations, domestic and international criminal tribunals, and United Nations missions. Most recently, he provided expert testimony in the trial of former *de facto* President of Guatemala, Gen. José Efraín Ríos Montt.

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 this project.

Although we share this excitement about the potential power of data analysis, our decades of experience analyzing data about conflict-related violence motivates us to proceed with caution. The data available to human rights researchers is fundamentally different from the data available to business and industry. The difference is whether the data are complete. In most business processes, an organization has access to all the data: every item sold in the past twelve months, every customer who clicked through their website, etc. In the exceptional cases where complete data are unavailable, industry analysts are often able to generate a representative sample of the data of interest.[2]

In human rights, and more specifically in studies of conflict violence, we rarely have access to complete data. What we have instead are snapshots of violence: a few videos of public killings posted to YouTube, a particular set of events retrospectively recorded by a truth commission, stories covered in the local or international press, protesters' SMS messages aggregated onto a map, or victims' testimonies recorded by non-governmental human rights organizations (NGOs) are typical sources. Statistically speaking, these snapshots are "convenience samples," and they cover an unknown proportion of the total number of cases of violence.[3] It is mathematically difficult, often impossible, to know how much is undocumented and, consequently, missing from the sample.

In human rights, and more specifically in studies of conflict violence, we rarely have access to complete data.

Incompleteness is not a criticism of *data*—collecting complete or representative data under conflict conditions is generally impossible. The challenge is that researchers and advocates naturally want to address questions that require either the total number or a representative subset of cases of violence. How many people have been killed? What proportion was from a vulnerable population? Were more victims killed last week or this week? Which perpetrator(s) are committing the majority of the violence? Basing answers and policy decisions on analyses of partial datasets with unknown, indeed unknowable, biases can prove to be misleading. These concerns should not deter researchers from asking questions of data; rather, it should caution them against basing conclusions on inadequate analyses of raw data. We conclude by suggesting methods from several quantitative disciplines to estimate the bias in direct observations.

The Problem of Bias

When people record data about events in the world, the records are almost always partial; reasons why the observation of violence often misses some or most of the violence are presented in the examples to follow. Most samples are partial, and in samples not collected randomly, the patterns of omission may have structure that influence the patterns observed in the data. For example, killings in urban areas may be nearly always reported, while killings

in rural areas are rarely documented. Thus, the probability of an event being reported depends on where the event happened. Consequently, analysis done directly from this data will suggest that violence is primarily urban. This conclusion is incorrect because the data simply do not include many (or at least proportionally fewer) cases from the rural areas. In this case, the analysis is finding a pattern in the documentation that may appear to be a pattern in true violence—but if analysts are unaware of the documentation group's relatively weaker coverage of the rural areas, they can be misled by the quantitative result. In our experience, even when analysts are aware of variable coverage in different areas, it is enormously difficult to draw a meaningful conclusion from a statistical pattern that is affected by bias.

Statisticians call this problem "selection bias" because some events (in this example, urban ones) are more likely to be "selected" for the sample than other events (in this example, rural ones). Selection bias can affect human rights data collection in many ways.[4] We use the word "bias" in the statistical sense, meaning a statistical difference between what is observed and what is "truth" or reality. "Bias" in this sense is not used to connote judgment. Rather, the point is to focus attention on empirical, calculable differences between what is observed and what actually happened.

In this article, we focus on a particular kind of selection bias called "event size bias." Event size bias is the variation in the probability that a given event is reported, related to the size of the event: big events are likely to be known, small events are less likely to be known. In studies of conflict violence, this kind of bias arises when events that involve only one victim are less likely to be documented than events that involve larger groups of victims. For example, a market bombing may involve the deaths of many people. The very public nature of the attack means that the event is likely to attract extensive attention from multiple media organizations. By contrast, an assassination of a single person, at night, by perpetrators who hide the victim's body, may go unreported. The victim's family may be too afraid to report the event, and the body may not be discovered until much later, if at all. These differences in the likelihood of observing information about an event can skew the available data and result in misleading interpretations about patterns of violence.[5]

Case Studies

We present here two examples from relatively well-documented conflicts. Some analysts have argued that information about conflict-related killings in Iraq and Syria is complete, or at least sufficient for detailed statistical analysis. In contrast, our analysis finds that in both cases, the available data are likely to be systematically biased in ways that are likely to confound interpretation.

Syria

Many civilian groups are currently carrying out documentation efforts in the midst of the ongoing conflict in Syria. In early 2012, the United Nations Office for the High Commissioner for Human Rights (OHCHR) commissioned

the Human Rights Data Analysis Group (HRDAG) to examine datasets from several of these groups, and in two reports, Price et al. provide in-depth descriptions of these sources.[6] In this section, we focus our attention on four sources—in essence, lists of people killed—which cover the entire length of the ongoing conflict and which have continued to provide us with updated records of victims. These sources are the Syrian Center for Statistics and Research[7] (CSR-SY), the Syrian Network for Human Rights[8] (SNHR), the Syria Shuhada website[9] (SS) and the Violations Documentation Centre[10] (VDC).

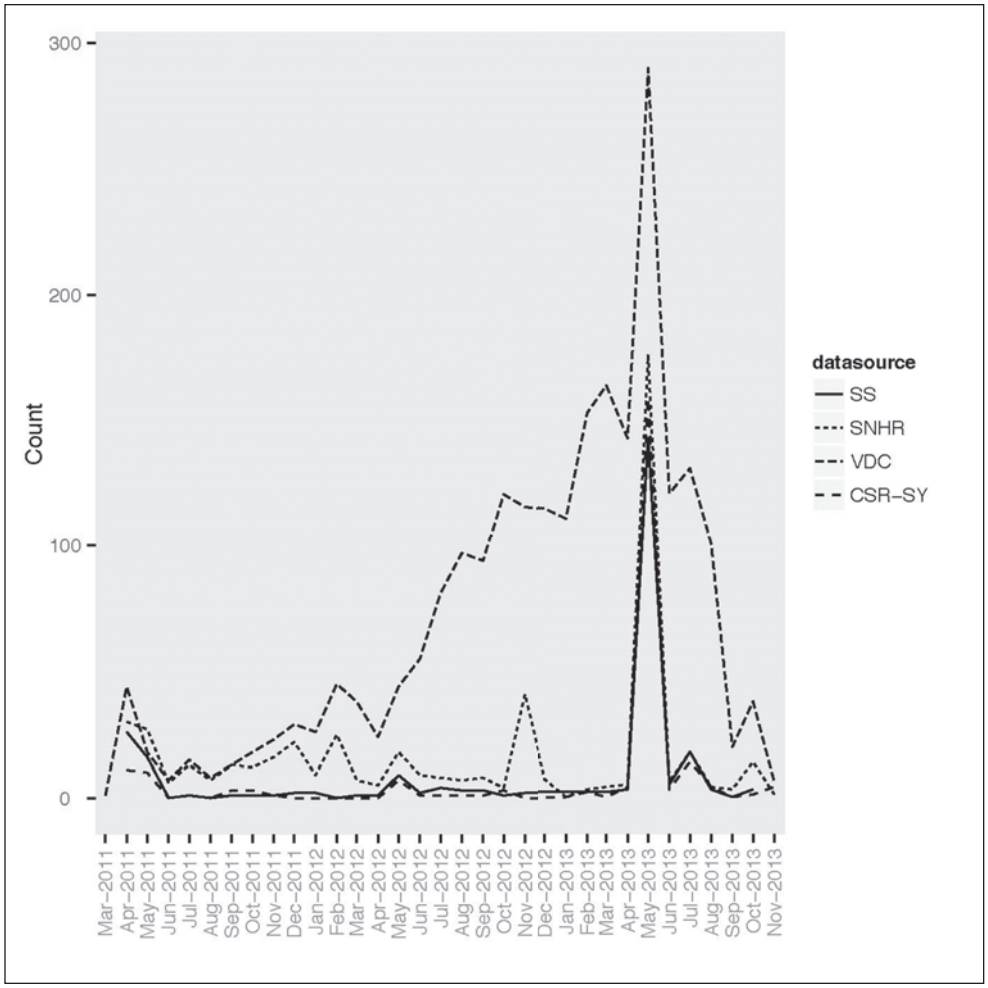
Figure 1 shows the number of victims documented by each of the four sources over time within the Syrian governorate of Tartus. The large peak visible in all four lines in May 2013 corresponds to an alleged massacre in Banias.[11] It appears that all four sources documented some portion of this event. Many victims were recorded in the alleged massacre, this event was very well reported, and all four of our sources reflect this event in their lists. However, three out of the four sources document very little violence occurring before or after May 2013 in Tartus. The fourth source, VDC, shows the peak of violence in May as the culmination of a year of consistent month-to-month increases in the number of reported killings.

When interpreting figures such as Figure 1, we should not aim 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.[12] The presence of event size bias is detectable in this particular example because all four of the sources obviously captured a similar event (or set of events) in May 2013, while at the same time one of those sources captured a very different subset of events during the preceding months. If we did not have access to the VDC data, our analysis of conflict violence in Tartus would incorrectly conclude that the alleged massacre in May 2013 was an isolated event surrounded by relatively low levels of violence.

The conclusion from Figure 1 should not be that VDC is doing a "better" job of documenting victims. VDC is clearly 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. From this figure alone we cannot conclude what other biases may be present in the observed data. For example, the relatively small peak in February 2012 could be as small as it seems, or it could be as large as the later peak in May 2013. Without a method of statistical estimation that uses a probability model to account for the undocumented events, it is impossible to know.[13]

To underline this crucial point: despite the availability of a large amount of data describing violence in Tartus, there is no mathematically sound method to draw conclusions about the patterns of violence directly from the data (though it is possible to use the data and statistical models to estimate how many events are missing). The differences in the four sources available to us make it possible to detect the event size bias occurring in May 2013, but what other biases might also be present in this observed data and hidden from view? What new events might a fifth, sixth, or seventh source document? Are there enough undocumented events such that if they were

Figure 1. Number of Victims Documented by Four Sources, Over Time, in Tartus



included, our interpretation of the patterns would change? These are the crucial questions that must be examined when interpreting perceived patterns in observed data.

Iraq

We detect a subtler form of event size bias in data from the Iraq Body Count (IBC), which indexes media and other sources that report on violent deaths in Iraq since the Allied invasion in March 2003.[14] Our analysis is motivated by a recent study by Carpenter et al., which found evidence of substantial event size bias.[15] Their approach was to compare the U.S. military's "significant acts" (SIGACTS) database to the IBC records. As they report, this comparison showed that "[e]vents that killed more people were far more likely to appear in both datasets, with 94.1% of events in which ≥20 people were killed being likely matches, as compared with 17.4% of … killings [that occurred one at a time]."[16] This implies that IBC, SIGACTS, or both, capture a higher fraction of large events than small events. Carpenter et al. go on

to note that "[t]he possibility that large events, or certain kinds of events (e.g., car bombs) are overrepresented might allow attribution that one side in a conflict was more recklessly killing civilians, when in fact, that is just an artifact of the data collection process."[17]

Motivated by this analysis, we considered other ways to examine IBC records for evidence of potential event size bias. Since IBC aggregates records from multiple sources, updated IBC data already incorporates many records from SIGACTS.[18] In contrast to the work of Carpenter et al., who treated IBC and SIGACTS as two separate data sources and conducted their own independent record linkage between the two sources, we examined only records in the IBC database, including those labeled as from SIGACTS.

It should be noted that we conducted this analysis on a subset of the data after filtering out very large events with more than fifty victims. We made this choice because, on inspection, many of the records with larger numbers of reported victims are data released in batches by institutions such as morgues, or incidents aggregated over a period of time, rather than specific, individual events.

We began by identifying the top one hundred data sources; one or more of the top one hundred sources cover 99.4 percent of the incidents in IBC.[19] Given these sources, we counted the number of sources (up to one hundred) for each event. Event size was defined as the mean (rounded to the nearest integer) of the reported maximum and minimum event size values. Then the data were divided into three categories: events with one victim, events with two to five victims, and events with six to fifty victims. The analysis was performed on these groups.

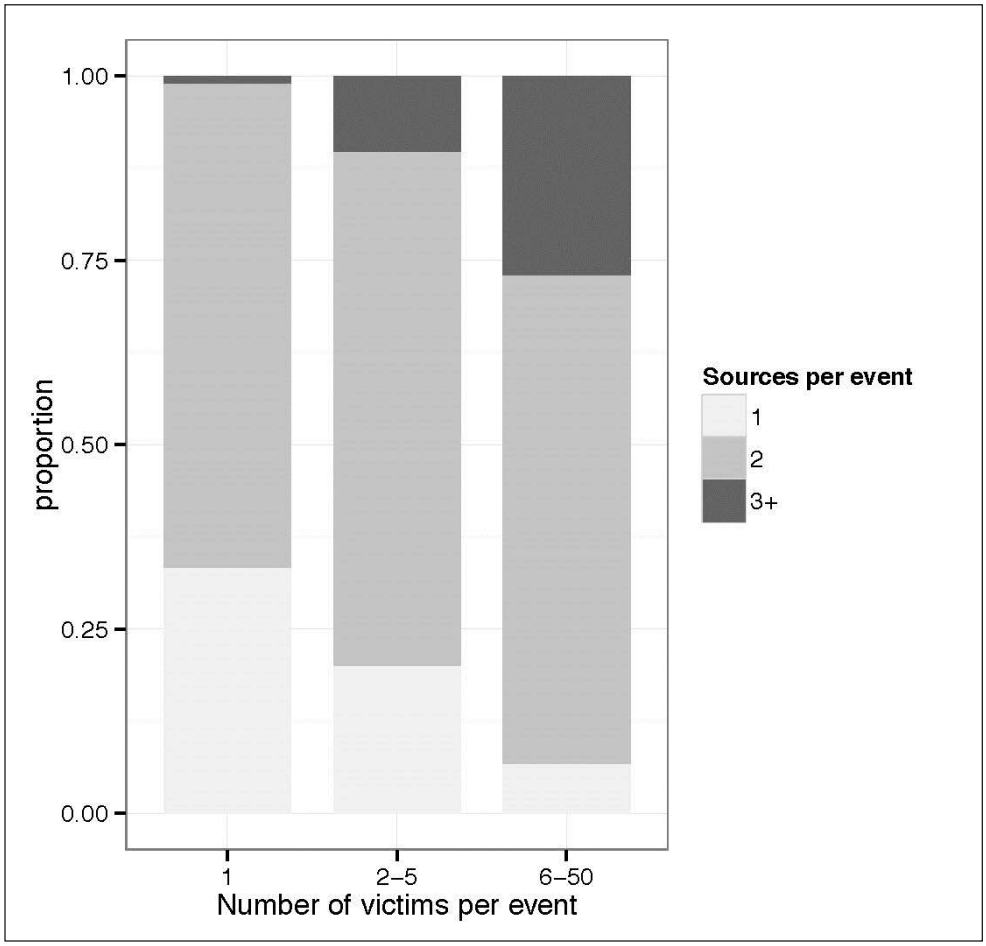
Figure 2 summarizes our findings. The shading of each bar in Figure 2 indicates the proportion of events of that size reported by one, two, or three or more sources. For each category of event sizes, most events have two sources. For events of size one, the second most frequent number of sources is one, accounting for nearly a third of all events of this size; almost no single-victim events have three or more sources. The number of events with three or more sources increases quickly in medium-sized events and in large events. Relatively few of the largest events are reported by a single source. Thus there seems to be a relationship between event size and the number of sources: larger events are captured by more sources. This reinforces the finding by Carpenter et al. that larger events are more likely to be captured by both IBC and SIGACTS. We have generalized this finding to the top one hundred sources; larger events are more likely to be captured by multiple sources.

The number of sources covering an event is an indicator of how "interesting" an event is to a community of documentation groups—in this case, media organizations. The pattern shown in Figure 2 implies that media sources are more interested in larger events than smaller events. Greater interest in the larger events implies that larger events are more likely to be reported (observed) by multiple sources relative to smaller events. Since a larger proportion of small events are covered by only a single source, it is likely that more small events are missed, and therefore excluded from IBC.[20]

As noted by Carpenter et al., the correlation between event attributes and the likely reporting of those events can result in highly misleading interpretation of apparent patterns in the data. As a relatively neutral example, analysts might erroneously conclude that most victims in Iraq were killed in large events, whereas this may actually be an artifact of the data collection. A potentially more damaging, incorrect conclusion might be reached if large events are centered in certain geographic regions or attributed to certain perpetrators; in these cases, reading the raw data directly would mistake the event size bias for a true pattern, thereby misleading the analyst. Inappropriate interpretations could result in incorrect decisions regarding security measures, intervention strategies, and ultimately, accountability.

The correlation between event attributes and the likely reporting of those events can result in highly misleading interpretation of apparent patterns in the data.

Figure 2. Proportion of Events Covered by One, Two, or Three or More Sources



Discussion

Event size bias is one of many kinds of selection and reporting biases that are common to human rights data collection. It is important to recall that we refer here to biases in the statistical sense: a measurable difference between the observed sample and the underlying population of interest. The biases that worry us here affect statistics and quantitative analyses; we are not implying that the political goals of the data collection groups have influenced their work.

In the context of conflict violence, meaningful statistical analysis involves comparisons to answer questions such as: Did more violence occur this month or last month? Were there more victims of ethnicity A or B? Did the majority of the violence occur in the north or the south of the country? The concern about bias focuses on how the data collection process may more effectively document one month relative to another, creating the appearance of a difference between the months. Unfortunately, the apparent difference is the result of changes in the documentation process, not real changes in the patterns of violence.

. . . the apparent difference is the result of changes in the documentation process, not real changes in the patterns of violence.

To make sense of such comparisons, the observed data must in some way be adjusted to represent the true rates. There are a number of methods for making this adjustment if the observed data were collected at random, but this is rarely the case. There are relatively few models that can adjust data that were collected because it was simply available.

In order to compare nonrandom data across categories like months or regions, the analyst must assume that the rate at which events from each category are observed is the same. For example, 60 percent of the total killings were collected in March, and 60 percent of the total killings were collected in April. This rate is called the coverage rate, and it is unknown, unless somehow the true number of events were known or estimated. If the coverage rates for different categories are not the same, the observed data tell only the story of the documentation; they do not indicate an accurate pattern. For example, if victims of ethnicity A are killed in large-scale violent events with many witnesses, while victims of ethnicity B are killed in targeted, isolated violent events, we may receive more reports of victims of ethnicity A and erroneously conclude that the violence is targeted at ethnicity A. Until we adjust for the event size bias resulting in more reports of victims of ethnicity A, we cannot draw conclusions about the true relationship between the number of victims from ethnicity A versus B.

There are many other kinds of selection bias. As an example, when relying on media sources, journalists make decisions about what is considered newsworthy. Sometimes their decisions may create event size bias, as large

events are frequently considered newsworthy. But the death of individual, prominent members of a society are frequently also considered newsworthy. Conversely, media "fatigue" may result in under-documentation later in a conflict, or when other newsworthy stories may limit the amount of time and space available to cover victims of a specific conflict.[21] Many other characteristics of both the documentation groups and the conflict can result in these kinds of biases such as logistical or budgetary limitations, trust or affinity variations within the community, and the security and stability of the situation on the ground.[22] As each of these factors changes, coverage rates are likely to change as well.

The fundamental reason why biases are so problematic for quantitative analyses is that bias often correlates with other dimensions that are interesting to analysts, such as trends over time, patterns over space, differences compared by the victims' sex, or some other factor. As in the example of ethnicities A and B above, the event size bias is correlated with the kind of event. Failing to adjust for the reporting bias leads to the wrong conclusion. As another example, consider the Iraq case described above: If event size is correlated with the events' perpetrators, then bias on event size means bias on perpetrator, and a naïve reading of the data could lead to security officials trying to solve the wrong security problems. Or, in the Syria case, if decisions about resource allocation to Tartus were made on the basis of the observed information, without taking into account the patterns of killings that were not observed, researchers may have inaccurately concluded that violence documented in May 2013 represented an isolated event. One could imagine that such a conclusion could lead to any number of incorrect decisions: sending aid groups into Tartus under the erroneous assumption of relative security, or failing to send aid and assistance before or after May 2013, assuming that such resources were more in need elsewhere.

It is important to note that these challenges frequently lack a scientific solution.[23] We do not need to simply capture more data. What we need is to appropriately recognize and adjust for the biases present in the available data. Indeed, as indicated in the Iraq example, where multiple media sources appear to share similar biases, the addition of more data perpetuates and in some cases amplifies the event size bias.

Detection of, and adjustment for, bias requires statistical estimation. A wide variety of statistical methods can be used to adjust for bias and estimate what is missing from observed data. In our work we favor multiple systems estimation, which has been developed under the name capture-recapture in ecology, and used to study a variety of human populations in research in demography and public health. Analysts more familiar with traditional survey methods often prefer adjustments based on post-stratification or "raking," each of which involves scaling unrepresentative data to a known representative sample or population.[24] Each method has limitations and requires assumptions, which may or may not be reasonable, but formal statistical models provide a way to make those assumptions explicit, and in some cases, to test whether they are appropriate. Comparisons from raw data implicitly but necessarily assume that such snapshots are statistically representative. This assumption may sometimes be true, but only by coincidence.

Conclusions

Carpenter et al. warn that "press members and scientists alike should be cautious about assuming the completeness and representativeness of tallies for which no formal evaluation of sensitivity has been conducted. Citing partial tallies as if they were scientific samples confuses the public, and opens the press and scholars to being manipulated in the interests of warring parties." In a back-of-the-envelope description elsewhere, we have shown that small variations in coverage rates can lead to an exactly wrong conclusion from raw data.[25]

Groups such as the Iraq Body Count, the Syrian Center for Statistics and Research, the Syrian Network for Human Rights, the Syria Shuhada website, and the Violations Documentation Centre collect invaluable data, and they do so systematically, and with principled discipline. These groups should continue to collate and share it as a fundamental record of the past. The data can also be used in qualitative research about specific cases, and in some circumstances, in statistical models that can adjust for biases.

It is tempting, particularly in emotionally charged research such as studies of conflict-related violence, to search available data for answers. It is intuitive to create infographics, to draw maps, and to calculate statistics and draft graphs to look for patterns in the data. Unfortunately, all people—even statisticians—tend to draw conclusions even when we know that the data are inadequate to support comparisons. Weakly founded statistics tend to mislead the reader.

Statistics, graphs, and maps are seductive because they seem to promise a solid basis for conclusions. The current obsession with using data to formulate evidence-based policy increases the pressure to use statistics, even as new doubts emerge about whether "big data" predictions about social conditions are accurate.[26] When calculations are made in a way that enables a mathematical foundation for statistical inference, these statistics deliver on the promise of an objective measurement in relation to a specific question. But analysis with inadequate data is very hard even for subject matter experts to interpret. In the worst case, it offers a falsely precise view, a view that may be completely wrong. In the best case, it invites speculation about what's missing and what biases are uncontrolled, creating more questions than answers, and ultimately, a distraction. When policymakers turn to statistical analysis to address key questions, they must assure that the analysis gives the right answers.

Statistics, graphs, and maps are seductive because they seem to promise a solid basis for conclusions.

Notes

[1] One extreme example includes Target successfully predicting a customer's pregnancy, as reported in the *New York Times* and *Forbes*. In particular, Target noticed that pregnant women buy specific kinds of products at regular points in their pregnancy, and the company used this information to build marketing campaigns.
[2] However it is certainly worth noting that even in these contexts sometimes big data are not big enough and may still be subject to the kinds of biases we worry about in this paper. See Kate Crawford's keynote at STRATA and Tim Harford's recent post on *Financial* Times for examples.
[3] Specifically, "convenience samples" refer to data that is non-randomly collected, though collecting such data is rarely convenient.
[4] Another common kind of bias that affects human rights data is *reporting* bias. Whereas selection bias focuses on how the data collection process identifies events to sample, reporting bias describes how some points become hidden, while others become visible, as a result of the actions and decisions of the witnesses and interviewees. For an overview of the impact of selection bias on human rights data collection, see Jule Krüger, Patrick Ball, Megan Price, and Amelia Hoover Green (2013). "It Doesn't Add Up: Methodological and Policy Implications of Conflicting Casualty Data." In *Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict*, ed. by Taylor B. Seybolt, Jay D. Aronson, and Baruch Fischhoff. Oxford UP.
[5] Christian Davenport and Patrick Ball. "Views to a Kill: Exploring the Implications of Source Selection in the Case of Guatemalan State Terror, 1977–1996." *Journal of Conflict Resolution* 46(3): 427–450. 2002.
[6] Megan Price, Jeff Klingner, Anas Qtiesh, and Patrick Ball (2013). "Full Updated Statistical Analysis of Documentation of Killings in the Syrian Arab Republic." Human Rights Data Analysis Group, commissioned by the United Nations Office of the High Commissioner for Human Rights (OHCHR). Megan Price, Jeff Klingner, and Patrick Ball (2013). "Preliminary Statistical Analysis of Documentation of Killings in the Syrian Arab Republic." The Benetech Human Rights Program, commissioned by the United Nations Office of the High Commissioner for Human Rights (OHCHR).
[7] http://www.csr-sy.com
[8] http://www.syrianhr.org
[9] http://syrianshuhada.com
[10] http://www.vdc-sy.info
[11] See reports in the *LA Times*, *BBC*, and the *Independent*, among others.
[12] Price. et al. 2013.
[13] See https://hrdag.org/mse-the-basics/ for the first in a series of blog posts describing Multiple Systems Estimation (MSE) or Kristian Lum , Megan Emily Price and David Banks (2013). Applications of Multiple Systems Estimation in Human Rights Research. *The American Statistician*, 67:4, 191–200. DOI: 10.1080/00031305.2013.821093
[14] http://www.iraqbodycount.org
[15] Carpenter D, Fuller T, Roberts L. "WikiLeaks and Iraq Body Count: the sum of parts may not add up to the whole—a comparison of two tallies of Iraqi civilian deaths." *Prehosp Disaster Med*. 2013;28(3):1–7. doi:10.1017/S1049023X13000113
[16] Ibid.
[17] Ibid.
[18] We downloaded the ibc-incidents file on 14 Feb 2014, and processed it using the pandas package in python.
[19] The top 100 sources include, for example, AFP, AL-SHAR, AP, CNN, DPA, KUNA, LAT, MCCLA, NINA, NYT, REU, VOI, WP, XIN, and US DOD VIA WIKILEAKS.
[20] These assumptions can be formalized and tested within the framework of 'species richness,' which is a branch of ecology that estimates the number of different types of species within a geographic area and/or time period of interest using models for data organized in a very similar way to the IBC's event records. See Wang, Ji-Ping. "Estimating species richness by a Poisson-compound gamma model." *Biometrika* 97.3 (2010): 727–740.

[21] A research question to address this might be: Do media-reported killings in a globally-interesting conflict like Iraq or Syria decline during periods when other stories attract interest? Do reported killings decline during the Olympics?

[22] Krüger et al. (2013)

[23] Bias issues can sometimes be resolved with appropriate statistical models, that is, with better scientific reasoning about the specific kind of data involved. However, we underline that bias is not solvable with better technology. Indeed, some of the most severely biased datasets we have studied are those collected by semi- or fully-automated, highly technological methods. Technology tends to increase analytic confusion because it tends to amplify selection bias.

[24] For a description of multiple systems estimation, see Lum et al. 2013. For methods on missing data in survey research which might be applicable to the adjustment of raw, non-random data if population-level information is available, see Brick, J. Michael, and Graham Kalton. "Handling missing data in survey research." *Statistical methods in medical research* 5.3 (1996): 215–238. For an overview of species richness models which might be used to estimate total populations from data organized like the IBC, see op. cit Wang. For an analysis of sampling issues in "elusive" populations, see Johnston, Lisa G., and Keith Sabin. "Sampling hard-to-reach populations with respondent driven sampling." *Methodological Innovations Online* 5.2 (2010): 38–48.

[25] https://hrdag.org/why-raw-data-doesnt-support-analysis-of-violence/

[26] Lazer, David and Kennedy, Ryan and King, Gary and Vespignani, Alessandro, Google Flu Trends Still Appears Sick: An Evaluation of the 2013–2014 Flu Season (March 13, 2014). Available at SSRN: http://ssrn.com/abstract=2408560