STRENGTHS AND WEAKNESSES OF OPEN SOURCE DATA FOR STUDYING TERRORISM AND POLITICAL RADICALIZATION

Gary LaFree
University of Maryland
Paul Gill
University College London

STRENGTHS AND WEAKNESSES OF OPEN SOURCE DATA FOR STUDYING TERRORISM AND POLITICAL RADICALIZATION

ABSTRACT

Compared to studying other forms of illegal violence, terrorism research faces unique data challenges. Notably, consistent police data are generally lacking, victims often have no direct knowledge of attacks and are sometimes killed, and perpetrators are difficult to interview and are often unreliable sources. In the face of these data challenges, the increasing availability of open source data on terrorist attacks and perpetrators has profoundly affected the scientific study of terrorism. Over time, a growing proportion of published academic research on terrorism—especially more heavily quantitative studies—relies on open source data. Moreover, the growing availability of individual-level terrorism data has allowed empirical tests of psychological and social-psychological theories of radicalization. In this paper, we assess the impact of open source data on the study of terrorism. We begin with an overview, providing a brief history of terrorist attack and perpetrator open source data bases. We then consider the major strengths and weaknesses of both types of open source data and suggest ways of countering the weaknesses. We close with observations about the impact of open source data on terrorism research and future opportunities for improvement.

STRENGTHS AND WEAKNESSES OF OPEN SOURCE DATA FOR STUDYING TERRORISM AND POLITICAL RADICALIZATION

Compared to studying other forms of illegal violence, studying terrorism faces unique data challenges. Notably, consistent police data are generally lacking, victims often have no direct knowledge of attacks and are sometimes killed and, and perpetrators are difficult to interview and are often unreliable sources. The development of open-source databases has been a major response to these data challenges. Open source terrorism databases emerge from the synthesis of unclassified, publicly available information retrieved from print (e.g., newspapers, court records) and digital (e.g., online news reports) media. The two most common units of analyses for these databases have been terrorist events and their perpetrators.[1] Terrorist event databases provide systematized descriptive information about terrorist attacks. Individual level databases generally focus on demographic, group affiliation and contextual information about terrorism perpetrators. Most commonly, event-level databases have been worldwide in scope while individual level databases have focused on perpetrators in a single country or a small group of, typically Western, countries.

Several well-informed reviews of the terrorism and radicalization research literature note how the growing availability of open source data on terrorist attacks and perpetrators has profoundly affected the scientific study of terrorism.[2] ???Reviews invariably refer to this as a "golden age"[3] where an "overreliance on literature reviews" and conceptual modelling was overcome,[4] and "clusters of excellence" emerged.[5] A growing proportion of published academic research on terrorism and radicalization—especially more heavily quantitative studies—relies on open source data. In this paper we assess the impact of open source data on the study of terrorism and radicalization. We divide the paper into two parts examining both open source data on terrorist attacks and also open source data on terrorist perpetrators. For both we provide a brief history, consider major strengths and weaknesses. We close with observations about the impact of open source data on terrorism and radicalization research and future opportunities for improvement. As a testimony to the growing importance of open sources in this area of research, we note that a comprehensive look at all open sources is now beyond the scope of a single review article.

Using Open Source Data to Study Terrorist Attacks

Contemporary efforts to build event databases can be traced back to the late 1960s and are likely related to the introduction of satellites and portable video equipment-- technology that made it possible to instantaneously send images of conflict and violence from one place on the planet to another. Bowie identifies 60 open source data bases,[6] however, in this part of the review we concentrate on just those that focus on terrorist attacks, have been worldwide in scope and have been available to academic researchers. The ITERATE database began coverage in 1968 and has been periodically updated through 2020.[7] ITERATE contains two different types of files: quantitatively coded data on terrorist attacks and a qualitative description of each attack included in the quantitative files. The quantitative data are arranged into three files, containing: (1) information on the type of terrorist attack, including location, name of group taking responsibility, and number of deaths and injuries; (2) information on the fate of the terrorist perpetrators or terrorist group claiming responsibility; and (3) information on terrorist attacks involving hostages and skyjackings. The ITERATE data are limited to international terrorist attacks, which are a small subset of all attacks.[8]

The RAND Corporation was an early pioneer in developing terrorism event databases. With the support of the US Department of State and the Defense Advanced Research Projects Agency

(DARPA), in 1972, Brian Jenkins at RAND began to develop a "Chronology of International Terrorism" dating back to 1968. Like ITERATE, the original RAND data were generally limited to international attacks. With varying levels of support, RAND maintained the Chronology through 1997.

Another early event database that began tracking terrorist attacks in 1970 was initiated in the mid-1970s with financial support from the Pinkerton Global Intelligence Service (PGIS)—a corporate relative of the famous Scottish-American detective agency. PGIS trained researchers to identify and record terrorism incidents from wire services (including Reuters and the Foreign Broadcast Information Service [FBIS]), US State Department reports, other US and foreign government reporting, and US and foreign newspapers (e.g., the New York Times, the British Financial Times). In the 1990s, PGIS researchers increasingly relied on the Internet. Although the PGIS coding form went through several iterations, most of the items included were similar during the 28 years of data collection. About two dozen persons coded information during these years, but only two individuals supervised data collection during the entire period. The most unique aspect of the PGIS data is that from the beginning it included domestic as well as international terrorist attacks—the only early database to do so.

The US State Department began publishing an annual report on international terrorism in 1982 (reporting 1981 incidents), and in 1983, began calling the report ''Patterns of Global Terrorism.'' The Patterns Report reviews international terrorist events by year, date, region, and terrorist group and includes background information on terrorist organizations, US policy on terrorism, and counterterrorism efforts. The Patterns of Global Terrorism reports were generally issued a few months after each calendar year.

In April 2001, the RAND corporation, which had been collecting terrorism data since 1968, received support from the National Memorial Institute for the Prevention of Terrorism (MIPT)—an organization that had been funded by the US Congress to study terrorism in the wake of the Oklahoma City bombing. With considerably more resources devoted to the database, RAND staff verified much of the earlier data and began collecting (starting in 1998) terrorism data on domestic attacks. Funding for the RAND-MIPT data collection ended in 2008. However, shortly after, RAND received additional support and continued collecting terrorism event data, now referred to as the RAND Database of Worldwide Terrorism Incidents (RDWTI).

The Worldwide Incidents Tracking System data (WITS) collected by the National Counterterrorism Center (NCTC) began in 2004 but did not provide comprehensive annual coverage until 2005. WITS originated because of congressional dissatisfaction with the quality of the State Department's Patterns of Global Terrorism report. World-Wide Incidents Tracking System (WITS) data were collected by NCTC from open sources manually using commercial subscription news services, the US Government's Open Source Center, local news websites reported in English, and as permitted by the linguistic capabilities of their employees, local news websites in foreign languages.[9] Like PGIS and RAND-MIPT, WITS collected both international and domestic data. From its inception, a major goal of those administering WITS was to "cast a wider net on what may be considered terrorism."[10] From 2004 to 2011, WITS reported nearly 80,000 terrorist attacks—far more than any other events database for this period. These numbers stand in particularly stark contrast to the earlier US State Department Patterns series, which typically reported only several hundred international terrorist attacks per year.

In 2001, a team the lead author headed at the University of Maryland was able to secure the original hard copies of the PGIS terrorism data—which by then contained more than 67,000 cases from 1970 to 1997. With funding from the National Institute of Justice, the Maryland team completed the verification and digitization of the original PGIS data in December 2005.[11] This marked the beginning of the Global Terrorism Database (GTD). In April 2006, the GTD team received funding from the Department of Homeland Security through the National Consortium for the Study of Terrorism and Responses to Terrorism (START) to extend the GTD beyond 1997. Primary data collection of the GTD for 1998 to 2011 was completed by two different research teams and then verified and compiled by the GTD team at START.[12] The new procedures captured more than 120 variables and unlike the original PGIS data, the process included archiving the open-source texts upon which each event was based. In March 2009 the GTD research team released the extended version of the GTD through 2007. Updates were subsequently released first biennially and then annually.

Starting in 2012, the GTD team at START began to substantially increase the amount of automation used to generate the data. The team still relies on primary sources including individual news outlets such as Reuters and Agence-France Presse, as well as existing media aggregators such as Lexis/Nexis and Factiva, but these are now continually evaluated in terms of which sources make the most valid contributions to the overall data collection effort. Poorly performing sources are removed and new sources are added to improve the efficiency of manual review processes. At present, data collection for the GTD begins with a universe of 2 million articles published daily worldwide in order to identify the small subset of articles that describe terrorist attacks. The team uses customized search strings to isolate an initial pool of potentially relevant articles and then relies on natural language processing techniques to automatically identify and remove duplicate source articles by measuring similarities between pairs of documents. In addition, the team has developed a machine-learning model using feedback from trained GTD staff that classifies documents identified by the initial automated processes to determine how likely they are to be relevant to terrorism. To facilitate this iterative process, the team has developed a web-based interface used to provide continuous feedback to the system through the manual review of the source documents identifying both false-positives (source documents that appear to describe terrorist attacks but do not) and false-negatives (source documents that appear to not describe terrorist attacks but actually do).[13]

Using Open Source Data to Study Terrorist Perpetrators

Efforts to build open source terrorism perpetrator databases began later than event databases. Again, the focus in this section is on those influential databases that are open source and publicly accessible. Arguably, individual-level open source data began with the *American Terrorism Study* (ATS) conducted out of the University of Arkansas and led by Brent Smith and Kelly Damphousse. The ATS covered event and target level data as well as data focused upon the perpetrators themselves. It is the latter component that we focus on here. The project began in 1989 and covered data from 1980 onwards. The Federal Bureau of Investigation's (FBI) Terrorist Research and Analytical Center aided the identification of cases by iteratively supplying the research team lists of individuals indicted through the activities of the FBI's Counterterrorism Program. Consequently, the study strictly aligns with the FBI's definition of terrorism.[14] The initial list of perpetrators included over 200 individuals from over 20 terrorist organizations (domestic and international) active in the United States from 1980 to 1989.[15] The Department of Justice matched these names to deferral court docket numbers throughout the United States. The ATS research team initially coded a range of variables on each of these individuals by visiting each major trial site where copies of indictments, judgement/commitment orders, and other processing information on each case was available. This information was subsequently supplemented with: (1) information from the Federal Records Center at the National Archives and examined on site, and (2) mailed records from federal district courts.[16] Variables collected include basic socio-demographic markers, criminal and military histories, and terrorist recruitment dynamics. The majority of the variables in early iterations of the data base relate to criminal justice procedures and outcomes like sentencing length and type of convictions. This data collection strategy corresponded to the project's initial overarching goal of testing whether terrorist tactics shifted in relation to prosecutorial success, the identification of the types of evidence that most likely leads to successful prosecution, and testing for the presence of sentencing disparity between terrorist and non-terrorist cases. Subsequent lists provided by the FBI and funding from sponsors including the US National Institute of Justice and the US Department of Homeland Security covered further data collection through to the present day.[17]

The *American Terrorism Study* was one of several sources later used to create aspects of the US Extremist Crime Database (ECDB) which began in 2008 and covers the period 1990 to 2018.[18] The ECDB extended beyond the ATS' focus on federal terrorism cases to include state cases as many domestic violent extremism cases are prosecuted within state courts. Like ATS, the ECDB collects much more than perpetrator data (our focus in this paper). It also includes event, victim, and targeting data. The ECDB extends beyond domestic violent extremism and includes data on arsons and bombings by eco-terrorists and financial crimes across different ideologies. To meet the inclusion criteria, a homicide or plot to commit violent acts must have: (1) been attempted or committed, (2) included at least one suspect who acted to advance their extreme right-wing belief system, and (3) happened in the United States. The ECDB's definitions are also more expansive than those of the FBI and the *American Terrorism Study* which excluded a large number of extreme right-wing incidents. The ECDB also identified some cases through other event databases including the GTD.[19] The ECDB continues to collect data through to the present day via funding through the National Institute of Justice.

In 2011, Gill's lone-actor terrorism database was funded by a one-year US Department of Homeland Security grant and later renewed with funding from the European Union and the US National Institute of Justice. At the time the lone actor database was created, academic research on lone actors was uncommon. Moreover, much of the research that existed was based on qualitative case studies and was largely atheoretical, making it of less value to the practitioner community who were becoming increasingly interested in the topic. This interest was partially event driven. In the years preceding 2011, the United States experienced a series of attacks attributed to lone actors, including the deliberate crashing of a plane into an IRS building in Texas; a shooting at the Pentagon Station in Washington, DC; a shooting of police officers in Memphis, Tennessee; a bombing attempt in the Times Square section of Manhattan; a hostage taking event at the Discovery Channel's headquarters in Silver Spring, Maryland; a series of shootings directed at military targets in Northern Virginia; and the attempted bombing of a Martin Luther King Jr. memorial march in Spokane, Washington.[20] During the same period Europe also witnessed a series of lone-actor attacks, including the attempted assassinations of the Danish cartoonist Kurt Westergaard and British MP Stephen Timms, a bombing of a Danish newspaper office in Copenhagen, a coordinated car and suicide bombing in Stockholm, and a shooting at a Frankfurt Airport targeting US Air Force personnel among others.[21] The contours of Gill's lone-actor database changed over time. The initial study was based on a dataset of 119 lone-actor terrorists from the United States and the European Union. This sample included 11 isolated dyads, 87 individuals who acted alone without command and control links, and 21 individuals with command and control links. In his subsequent 2015 book Gill drew on a sample of 111 lone-actors, including some of the previous study's subjects but excluding "isolated dyads" and individuals who acted alone "but in facilitative roles."[22] Later variations, which are publicly available, included comparison groups of mass murder perpetrators who conducted similar forms of violence in similar public settings but absent an overarching political, social, or religious cause.[23] What set the lone-actor database apart from other data sets at the time was its overt focus on feeding into threat and risk assessment practice. The number of behavioral variables (150+) was more extensive than what had been previously compiled. These variables were largely inspired by Fein and Vossekuil's *Assassination in the United States* which characterized attacks on public figures as "the end result of an understandable, and often discernible, process of thinking and action."[24] Over the years, dozens of trained coders examined open source materials (e.g. court documents, newspaper archives, manifestos, social media posts) of lone-actor terrorist cases to synthesize information on the thinking and action processes that motivated these perpetrators.

Just like later versions of the GTD, the Profiles of Individual Radicalization in the United States (PIRUS) database was built and has been maintained at the START Center at the University of Maryland.[25] It initially focused on individuals radicalized within the United States from 1948 to 2013. Outcome variables include conducting violent or non-violent ideological offenses and joining or associating with designated terrorist groups. The original data were collected between January 2013 and June 2014 and relied on similar publicly available resources as the lone-actor terrorism database described above (e.g. court documents, newspaper archives, government reports). It contained 147 variables related to each individuals' criminal activities, violent plots, relationship with extremist groups, radicalization processes, ideological attachment and personal histories.[26] Following the original creation of PIRUS, additional funding from the US National Institute of Justice has allowed the PIRUS team to expand data collection and focus on a variety of additional terrorism-related topics, including comparisons of political extremists and gang members,[27] the impact of the internet and social media on domestic radicalization,[28] and the decision-making of foreign fighters.[29] The current version of PIRUS contains over 3,200 violent and non-violent extremists from 1948 to 2021, making it one of the largest of the individual-level databases.[30] Despite its relatively large size, PIRUS represents a sample and not a universe of all US-based extremists.

Evaluating Event and Perpetrator Databases

Strengths

Although open source terrorism databases arose largely because of the inherent challenges of collecting data on terrorism from other sources, nevertheless, they have several important strengths. In this section we note four in particular. First, because there is no universally accepted legal definition of terrorism, the event databases have moved forward by providing a definition of terrorism and then with varying data collection methods, applying this definition across all countries. One of the biggest difficulties of doing cross-national research on issues like crime is that every country has its own legal definitions and unique strategies for collecting data. By creating cross-national open-source databases on terrorism, it is possible to develop a single uniform definition and apply it systematically to all countries of the world. For example, the definition of terrorism adopted by the Global Terrorism Database is "the threatened or actual use of illegal force and violence by non-state actors to attain a political, economic, religious, or social goal through fear, coercion or intimidation."[31] Because the GTD research team created their own definition, they were able to apply the same definition to all countries of the world. By contrast, most cross-national research on a specific type of crime such as homicide, must rely on officials to collect and report the data and have few options if legal definitions differ across countries.[32] As a result, international homicide data is often available for less than a third of the world's countries, while open source terrorism databases offer at least some coverage for all countries.

This strength is also evident in the case of terrorist perpetrator databases where no official data exists beyond descriptive prevalence rates of charges/arrests/convictions or referrals to different countering violent extremism/preventing violent extremism initiatives. Just like the case of terrorism event databases, this strength is particularly pronounced in those few initiatives that sought to collect data cross-nationally. In those cases, uniform definitions allow researchers to traverse the difficult landscape of how different criminal justice systems work, their propensities to seek convictions for homicide or attempted murder in clearly ideologically influenced cases, and their different inclinations for pursuing prosecutions for terrorism versus hate crimes. For example, in May 2010, Roshonara Coudhry, a 21-year old British former student and an Islamist extremist, was found guilty of attempted murder rather than terrorism for the stabbing of Stephen Timms, a Labour MP for East Ham. In the case of lone-actor terrorism databases, the official terms and definitions differ greatly even across ostensibly similar Western countries. For example, the Danish Security and Intelligence Service distinguishes "solo terrorists" from "lone wolf terrorists" because the latter "has no contact to a terror groups (not even historically) or any other radicalized individuals and consequently the individual acts completely isolated and without instruction from any other militant individual." By contrast, the Dutch National Counterterrorism Strategy 2011 to 2015 divides "menacing loners" across two types: "radicalized individuals" and "fixated persons." Radicalized individuals are defined as "individuals who are inspired, motivated, and sometimes directed by (virtual) networks of a more or less known ideology or religion." This definition therefore incorporates both individuals lacking guidance and control and those directed by others – a distinction that the Danish Security and Intelligence Service is also keen to stress but with different boundaries and terms which makes official comparison difficult.[33]

Second, because of the compelling interest that terrorists have in media attention, open-source information may be uniquely useful in the study of terrorism.[34] Terrorists, unlike most common criminals, actively seek media attention. Terrorism expert Brian Jenkins has argued that "terrorism is theatre" and points out that terrorist attacks are often carefully choreographed to attract the attention of the electronic media and the international press.[35] The fact that terrorists are specifically seeking attention through the media means that compared to coverage of common crimes, coverage of terrorism is likely to be more complete. Thus, while few researchers would suggest tracking burglary or fraud rates by studying electronic and print media, it is likely more defensible to track terrorist attacks in this way. For example, it is hard to imagine that it is possible today for an aerial hijacking or politically motivated assassination—even in remote parts of the world—to completely elude the attention of the global media.

Although open source information is typically bountiful on the vagaries of an attack in terms of who was attacked, the mode of attack, and casualty rates – the degree of information on individual perpetrators often varies greatly across different sources.[36] The same is also true for the volume of information available on different types of cases. The general rule of thumb from open source databases that include perpetrators is that the rarer the event, the greater the volume of open source data that is typically available about the perpetrator. For example, in the past lone-actor terrorist perpetrators have been far less common than perpetrators who offend as part of a group. Accordingly, the amount of available behavioural data is likely to be far greater for lone-actor perpetrators than perpetrators who operate on behalf of a prolific group. Given the low base rate of lone-actor terrorism, the volume of reporting tends to be much higher compared to campaigns of violence where trials and convictions are a weekly occurrence. It is very difficult to obtain from open sources much more than very basic socio-demographic information (e.g. gender, age) on individual members of extremely active terrorist groups. For example, the Gill lone actor terrorist data introduced above had specific educational attainment information for 65 percent of the sample.[37] This is compared to less than ten percent of Gill and Horgan's sample of Provisional Irish Republican Army (PIRA) members, for whom level of education could generally only be inferred from the individual's occupational status.[38]

Third, the utility and sophistication of terrorism open-source databases is improving over time. The amount of media coverage of events and the available tools for capturing these events and turning them into analyzable data has grown enormously in the last few decades along with the growth of the electronic media. Event databases created in the late 1960s began by asking human experts to review print media, identify promising cases and then physically record the details. The information available and the process of creating event databases is considerably more complicated today. As the size and scope of the print and electronic media continues to grow the breadth and coverage of data has increased dramatically. In addition, researchers progressively have greater access to new automated process tools to deal with what has become an example of big data. One recent study that compared closed-source data from police and intelligence files to open source data largely confirmed the validity of open-source data.[39] Based on police data, the authors found broadly similar prevalence rates as identified in open-source lone-actor terrorism research. As such, we might conclude that outcomes from open source data, when properly collected and handled, may be of considerable value. Although we hasten to add that more comparative research is needed. This study was based on lone actors and as we noted above, open source data on lone actors may be more complete than that from perpetrators in group-based plots or terrorist-related activities that did not produce convictions (e.g., those who successfully travelled to Syria to fight for ISIS).

Finally, because the sources underpinning open-source data are freely available to the public, open-source collection initiatives often offer the bonus of research process replicability—wherein other researchers are able to replicate the exact methods of data collection and analysis used in another study—rather than just attempt to test or replicate research outcomes. Replicating research processes can be much more difficult when using a completely closed-source data collection protocol, given that access to the data necessary to replicate the study is heavily restricted, difficult to access, and often derived through a wide range of tactically sensitive means.

Weaknesses

Not surprisingly, open source terrorism databases also have important weaknesses and in this section we elaborate on nine of these. First, they lack a generally accepted definition of terrorism. Each event and perpetrator database answers the question "what constitutes terrorism" differently, and accordingly, this informs their respective data collection criteria. This is not a trivial issue. An early review by Schmid and Jongman found 109 different research definitions of terrorism.[40] Sheehan examined the terrorism definitions used by five prominent event databases along 12 different dimensions.[41] He finds that RAND included nine of these dimensions in their terrorism definition, ITERATE and GTD included seven, and WITS included six.[42] Only two dimensions were included in the terrorism definitions of all the databases (the act is outside the context of warfare or *coup d'état* and the perpetrators are sub-state actors or groups). Two additional dimensions were included in all but one of the five databases (the act was intended to influence an audience and the act included the use of violence). The remaining eight elements were divided between different combinations of three or fewer of the five databases. It is apparent that if the creators of event databases start with different definitions of the issue being studied, they will end with different results.

Second, apart from the challenges raised by lack of agreement on operational definitions, terrorism databases are ultimately a function of what journalists, wire services, and criminal justice systems choose to report on, and likewise, what they may refer to as acts of political violence.[43] Perhaps the most obvious of the open source data weaknesses that follow from this fact is the differential worldwide freedom of the press. When the press in countries like Eritrea, North Korea or Turkmenistan report few terrorist attacks, we cannot know for certain that it is because they actually have few attacks, or whether the leadership in these countries are successful at muzzling the press from reporting on attacks. Drakos and Gofas caution that research on the frequency of terrorist attacks may be biased by under-reporting of attacks in countries with low freedom of the press.[44] However, the potential for bias is not limited to countries with low levels of press freedom. In high press-freedom countries the print and electronic media are mostly businesses which have strong incentives to focus on the deadliest, most serious, and therefore newsworthy incidents. This means that coverage may be less complete for attacks in remote areas with less coverage, especially if there are no casualties.

Third, as terrorism databases have evolved the size of the media "fire hose" upon which they are based continues to increase. For example, the media sources that the GTD relies on has moved from one million per day to over two million per day in less than ten years.[45] This can be seen as a positive development in that more extensive information is available over time. However, these increases also raise the possibility that trends in terrorist attacks over time are affected by increasing access to information rather than an escalation in the total number of attacks. As the size of the media fire hose increases, there are also a rising number of duplicate cases. A high-profile case like the Fishmonger's Hall stabbing in London on November 29, 2019, can generate thousands of news stories worldwide,[46] greatly complicating the difficulties of collecting valid open-source data.[47] Increasingly there is also a "fog of war" issue in relation to perpetrator data collection whereby early reporting on an incident later proves to be incorrect. Recent high profile attacks in Europe have misattributed ideological motivation (e.g., early reporting on Breivik's attack speculated that it was Islamist inspired) and perpetrator identity (e.g., the UK media regulator found Channel 4 News seriously breached the broadcasting code by identifying the incorrect individual as the perpetrator of the 2017 London Bridge attack in the hours after the attack).

Fourth, despite the fact that we live in an increasingly interconnected world where the depth of available worldwide information continues to increase, open sources frequently have great difficulty distinguishing between negative responses and missing data. Given the nature of the source material, it is unrealistic to expect each story of a terrorist perpetrator or attack to contain lengthy passages that detail every variable the offender or event in question did *not* exhibit. Thus, creators of open source data frequently face the challenge of correctly assessing whether no mention of substance abuse or mental health issues or a criminal record means that the perpetrator did not have these characteristics or these characteristics were simply not mentioned. Exacerbating the challenge is that each open source data base comes up with its own decisions about the circumstances under which no information on a particular variable can be assumed to mean that the variable should be coded as a "no."

Fifth, beyond these general issues, there are subtler biases related to the media itself. Fishman describes a well-known tendency for news sources to develop "waves" of information following specific themes, like “Islamist extremism” or “improvised explosive devices.”[48] Over the past decade, different “risk factors” or explanations for terrorist engagement have come in and out of vogue within mainstream media reporting. These include various radicalizing settings (e.g., geographic hotspots, prisons), online affordances (e.g., particular social media applications), mental health problems, and domestic violence perpetration. Each have often been treated as simple mono-causal explanations at different times. When interest in such characteristics within the media are at their height, there may be a bias within reporting to go seek these factors out as new terrorist incidents emerge, thus artificially inflating their prevalence rate when other push and pull factors might be more relevant.[49] Once reporters start focusing on these themes it can give the appearance that a particular form of behavior is increasing even if the increases are being driven all or in part by selective attention to a specific type of event. For example, Kearns, Betus and Lemieux examined news coverage from LexisNexis Academic and CNN.com for all terrorist attacks in the United States between 2006 and 2015 and controlling for a variety of rival explanations found that attacks by Muslim perpetrators received, on average, 357 percent more coverage than other attacks.[50]

Sixth, challenges to collecting terrorism event data are especially great in war-torn, high conflict societies—and of course, such societies are especially likely to suffer from terrorist attacks.[51] Miller reports that 38 percent of worldwide terrorist attacks reported by the GTD in 2019 were connected with just three countries: Afghanistan, Yemen and Iraq—all three of which were experiencing a great deal of violence and conflict at the time, and none of which have reliable open-source perpetrator databases.[52] Instead, to the present, perpetrator databases have been most common in Western democracies with relatively strong freedom of the press.

Seventh, adequately covering terrorist attacks reported in all of the world’s languages has always been a challenge to researchers in general and open source databases in particular. In response, all of the early event databases made efforts to include researchers who were fluent not only in English but in several other strategic languages (especially, Arabic, French, Mandarin, Russian and Spanish). Although this was a logical response, its effectiveness depended largely on the idiosyncratic staffing decisions of teams collecting event data. When the GTD team started collecting terrorism data in 2012, it still included analysts with non-English language skills. However, over time the GTD team has increasingly relied on supplementing the English-language content in the LexisNexis Metabase news feed with content from Open Source Enterprise (OSE)—a US government media aggregation and translation service originally established in 1941. However, on June 3, 2019 the US government decided that the OSE platform would no longer be available to those who are not US government employees with Joint Worldwide Intelligence Communications System credentials. In response, the GTD team now uses BBC Monitoring, a media aggregation and translation platform, to pick up foreign language sources. In a recent review of this change, the GTD team concluded that the switch has reduced the total number of foreign language stories available but the extent to which this change has affected total event counts is unknown.[53] The value of supplementing additional language skills into perpetrator databases has also been demonstrated within recent studies whereby much richer data was gleaned using local sources reporting in native languages.[54]

Eighth, theoretical developments are just as important for the field of terrorism studies as those of empirical developments. Event databases allow for competing theories about the so-called “root causes” of terrorism to be tested against one another. This line of research has proliferated widely due to open source databases like the GTD because such databases provide the capability of matching event level data with state-level socio-political data collated elsewhere (e.g., World

Bank data). However, the same is not currently true for perpetrator-level databases. For individual level databases, the inclusion of variables is often driven by pragmatic concerns rather than fully developed theorizing. [55] Increasingly, there is a realization that many common theories remain untested on perpetrators of terrorism.[56] Arguably, this disparity in theory testing is because event-level analyses can design analyses that vary on the dependent variable (e.g., number of terrorist attacks in a year within a given country). The development of adequate control groups for perpetrator-level analyses is still an on-going methodological challenge.[57]

Finally, given the substantial financial and logistical requirements, creating and keeping a trained research team capable of maintaining high quality data collection for either event or perpetrator-level data bases is challenging. In general, support for event databases on terrorism has been reactive, based on specific political crisis rather than proactive efforts to establish policy priorities. Thus, following the Oklahoma City bombing by Timothy McVeigh and Terry Nichols on April 19, 1995, US funds were made available to greatly strengthen RAND's terrorism data. Similarly, following the 9/11 attacks, US funding was made available for the collection of WITS and GTD data. Given these changing reactions it is unsurprising that many of the terrorism databases that began in the late 1960s are no longer collecting new data. In fact, of the terrorism event data bases summarized above only ITERATE and GTD are still active: ITERATE is limited to international attacks, and the GTD, which at the time this article was being prepared was available through 2020, has faced periodic financial challenges. Unfortunately, perpetrator databases, have been even more difficult to sustain than event databases. At present there is no individual-level database of which we are aware that has guaranteed long-term funding, although several prominent databases (including ATS, ECDB and PIRUS) have managed to continue collecting data through a series of individual grants.

Improving Open Source Terrorism Databases

As noted, one of the most important weaknesses of event data is the possibility that terrorist attacks are likely to be more difficult to accurately identify in some countries than others. One strategy for countering this challenge is to develop rankings for the quality of data being used. For example, starting in 2012, the GTD team developed a ranking system for all sources used to create the database. To reduce error, the GTD only includes attacks that are reported by at least one high-validity independent news outlet: those that are independent (free of influence from the government, political perpetrators, or corporations), those that routinely report externally verifiable content, and those that are based on eyewitness accounts.[58]

Another strategy for responding to inevitable data challenges in producing open sources is to give the users of the data a set of options for how to conduct analysis. For example, to provide maximum flexibility to analysts, the GTD errs on the side of inclusiveness in determining whether a specific incident is terrorism, but includes in the database filtering mechanisms through which users can modify the data set according to their specific analytic goals. Thus, cases are included in the GTD that meet at least two of the following three criteria: (1) the act must be aimed at attaining a political, economic, religious, or social goal; (2) there must be evidence of an intention to coerce, intimidate, or convey some other message to a larger audience than the immediate victims; and (3) the action must be outside the context of legitimate warfare activities. In addition, the GTD includes a category that flags cases that fit the GTD criteria but also resemble categories other than terrorism; most notably, insurgency, political conflict or warfare or more conventional crimes such as homicide or kidnapping. This allows researchers that use the GTD to remove cases where inclusion decisions are less definitive. Similar strategies are possible for perpetrator databases. For example, Gill's database on lone actors allows researchers to filter for competing definitions of lone-actor terrorism that exclude or include individuals with any connections to a broader network,[59] or allow for some form of accomplices.[60]

Despite the reference to open source data, some of the open source databases that have been collected are not in fact open to researchers or are kept behind a pay wall that limits their accessibility. The motives for being protective with regard to data that in many cases has taken a good deal of time and energy to collect are obvious. However, it is just as obvious that from a scientific stand point, making data available to the research and policy communities—perhaps after a reasonable delay--is extremely valuable. Some of the open source data bases reviewed here (e.g., ATS) have made data available through the Inter-university Consortium for Political and Social Research (ICPSR) library at the University of Michigan. Others make their data available through online websites (e.g., PIRUS). The history of the GTD provides an example of the importance of making open source data available to other researchers. From the time the GTD team completed digitizing the original data in December 2005, the data have always been publicly available.[61] By 2010 the GTD was being used in 200 books and articles each year.[62] By 2015 this number had more than tripled to over 600 citations per year. GTD citations hit 1,200 per year for the first time in 2018 and have remained at about that level through 2021. Reliance on the GTD in published research increased six-fold in less than a decade. Not all of these citations reflect original analysis, but many do. Making data sharing following the completion of a project the norm could greatly increase the scope and speed of scientific discovery in this area.

In addition to a greater willingness to share open source data, it is critical that researchers be transparent in the coding decisions that were used to create databases. Making detailed codebooks and coding decisions publicly available is essential.

More empirical comparisons between open source databases would also be useful. Efforts to compare open source databases have been slow to develop, hampered especially by major differences between definitions, scope and years of coverage. Nevertheless, such comparisons could increase knowledge about the extent to which individual databases support similar conclusions. This could be especially useful for perpetrator databases, since most of these are limited to a particular country or a small number of countries.

Conclusion

Contemporary terrorism event databases have become more feasible over time, first with the availability of satellite technology and portable cameras in the late 1960s and later with the explosive growth of the internet. The major strengths of open source terrorism databases include the fact that they allow researchers to skirt definitional differences across countries by applying the same definition to all cases, they take advantage of the fact that terrorism depends on media coverage, their increasingly frequent use has created a growing corpus of research evidence, and the fact that they are not dependent on classified governmental sources makes them easier to access and replicate. They also have a number of important weaknesses, including the fact that the media may report misinformation, there may be conflicting information or false, multiple or no claims of responsibility and despite improvements, coverage still relies more on Western than non-Western news organizations. It is worth pointing out that all of these weaknesses are also frequently mentioned as drawbacks of official data on terrorism. Certainly, censorship and disinformation affects not only media sources but official government sources. As open source databases continue to multiply and expand, it is important that researchers do more to improve them. Improvement strategies could include paying greater attention to the validity of sources, providing analysts with greater flexibility to select cases for inclusion depending on their analytic goals, encouraging creators of open source databases to make them publicly available, and undertaking more empirical comparisons between different databases. Open source databases on terrorism are likely to persist because they provide information that is not easily gained from any other existing source.

ENDNOTES

[1] Although far less common, a third type of open source terrorism database examines terrorist organizations (see for example, Asal, Victor, Park, H.H., Rethemeyer, R.K. and Ackerman, G., "With Friends Like These… Why Terrorist Organizations Ally." *International Public Management Journal*, 19(2016): 1-30).
[2] LaFree, Gary, *Toward a Criminology of Terrorism* (Elements in Criminology, Cambridge University Press, 2023); LaFree, Gary and Joshua Freilich, *The Handbook of the Criminology of Terrorism* (Wiley Blackwell, 2016).
[3] Silke, Andrew, and Jennifer Schmidt-Petersen. "The golden age? What the 100 most cited articles in terrorism studies tell us." *Terrorism and Political Violence* 29, no. 4 (2017): 692-712.
[4] Schuurman, Bart. "Research on terrorism, 2007–2016: A review of data, methods, and authorship." *Terrorism and Political Violence* 32, no. 5 (2020): 1011-1026.
[5] Neumann, Peter, and Scott Kleinmann. "How rigorous is radicalization research?." *Democracy and Security* 9, no. 4 (2013): 360-382.
[6] Bowie, Neil G. "Terrorism Events Data: An Inventory of Databases and Data Sets, 1968-2017," *Perspectives on Terrorism* 11(2017): 50–72.
[7] Mickolus, Edward F. "How Do We Know We're Winning the War Against Terrorists? Issues in Measurement." *Studies in Conflict & Terrorism* 25(2002): 151–60; Mickolus, Edward F., Todd Sandler, Jean M. Murdock, and Peter Flemming. *International Terrorism: Attributes of Terrorist Events (ITERATE)*" (Dunn Loring, VA: Vinyard Software, 2010).
[8] We define international attacks as those in which the nationality of the perpetrator, the target or the location are different (LaFree, Gary, Dugan, Laura and Erin Miller. *Putting Terrorism in Context: Lessons from the Global Terrorism Database*. (London: Routledge, 2014, p. 150).
[9] Wigle, John. "Introducing the Worldwide Incidents Tracking System (WITS)." *Perspectives on Terrorism* 4(2010): 21.
[10] Ibid (p. 5).
[11] LaFree, Gary, and Laura Dugan. "Introducing the global terrorism database," *Terrorism and Political Violence* 19(2007): 181-204.
[12] LaFree, Dugan and Miller, Putting Terrorism in Context.
[13] Jensen, Michael. Discussion Point: The Benefits and Drawbacks of Methodological Advancements in Data Collection and Coding: Insights from the Global Terrorism Database (GTD). (2013).
[14] The FBI defines terrorism as "The unlawful use of force or violence against persons or property to intimidate or coerce a government, the civilian population, or any segment thereof, in furtherance of political or social objectives".
[15] Smith and Damphousse 2007.
[16] Ibid.
[17] Ibid.
[18] Freilich, Joshua D., Steven M. Chermak, Roberta Belli, Jeff Gruenewald, and William S. Parkin. "Introducing the United States extremis crime database (ECDB)." *Terrorism and Political Violence* 26, no. 2 (2014): 372-384.
[19] See Note 13 above Freilich et al 2014, Appendix 4
[20] Paul Gill, *Lone-actor Terrorists: A Behavioural Analysis* (Routledge, 2015).
[21] Gill, 2014; Gill, *Lone-actor Terrorists.*
[22] Gill 2015.

[23] Gill, Paul, James Silver, John Horgan, Emily Corner, and Noemie Bouhana. "Similar crimes, similar behaviors? Comparing lone-actor terrorists and public mass murderers." *Journal of Forensic Sciences* 66, no. 5 (2021): 1797-1804.
[24] Fein, Robert A., and Bryan Vossekuil. "Assassination in the United States: an operational study of recent assassins, attackers, and near-lethal approachers." *Journal of Forensic Sciences* 44 (1999): 321-333.
[25] LaFree, Dugan and Miller, *Putting Terrorism in Context*, ch. 2.
[26] Jensen and LaFree 2016.
[27] Pyrooz, David, Gary LaFree, Scott Decker and Patrick James. "Cut from the same cloth? Comparing gangs and Violent Political Extremists." *Justice Quarterly* 35(2017): 1-32.
[28] LaFree, Gary. Terrorism and the Internet. *Criminology and Public Policy* 16(2017): 93-98.
[29] Jensen, Michael, Neil Ferguson, Elizabeth Sheehan and Gary LaFree. "Do Social Networks Distinguish American ISIS Foreign Fighters from ISIS-Inspired Terrorists?" *Journal of Conflict Resolution* (2023): Mar 26:00220027231164925.
[30] Wolfowicz, Michael, Yael Litmanovitz, David Weisburd, and Badi Hasisi. "A field-wide systematic review and meta-analysis of putative risk and protective factors for radicalization outcomes." *Journal of Quantitative Criminology* 36 (2020): 407-447.
[31] LaFree, Dugan and Miller, *Putting Terrorism in Context*, p. 13.
[32] LaFree, *Toward a Criminology of Terrorism*.
[33] Gill, 2014; Gill, *Lone-actor Terrorists*.
[34] See also Moskalenko, Sophia, Joshua D. Freilich, Steven Chermak, Jeff Gruenewald, and Clark McCauley. "Growth opportunities in American and British terrorism research." *Dynamics of Asymmetric Conflict* 16, no. 1 (2023): 44-69.
[35] Jenkins, Brian. "Will Terrorists Go Nuclear?," January (1975). https://www.rand.org/pubs/papers/P5541.html.
[36] Chermak, Steven M., Joshua D. Freilich, William S. Parkin, and James P. Lynch. "American terrorism and extremist crime data sources and selectivity bias: An investigation focusing on homicide events committed by far-right extremists." *Journal of Quantitative Criminology* 28, no. 1 (2012): 191-218.
[37] Gill, Paul, and John Horgan. "Who were the volunteers? The shifting sociological and operational profile of 1240 Provisional Irish Republican Army members." *Terrorism and Political Violence* 25, no. 3 (2013): 435-456.
[38] Ibid,
[39] Gill, Paul, Emily Corner, Amy McKee, Paul Hitchen, and Paul Betley. "What do closed source data tell us about lone actor terrorist behavior? A research note." *Terrorism and Political Violence* 34, no. 1 (2022): 113-130.
[40] Schmid, Alex P. and Albert J. Jongman. *Political Terrorism: A New Guide to Actors, Authors, Concepts, Data Bases, Theories, and Literature*. (Transaction Publishers, 1988, 5).
[41] Sheehan, Ivan Sascha. "Assessing and comparing data sources for terrorism research." In *Evidence-Based Counterterrorism Policy*, edited by Cynthia Lum and Leslie W. Kennedy, 13–40. Springer Series on Evidence-Based Crime Policy. (New York, NY: Springer, 2012).
[42] The fifth database reviewed was the Terrorism in Western Europe: Event Data (TWEED). We excluded TWEED from the earlier discussion because it is limited to domestic terrorism attacks in 18 West European countries [Engene, Jan Oskar. "Five Decades of Terrorism in Europe: The TWEED Dataset," *Journal of Peace Research* 44(2007): 109–21].
[43] See Moskalenko et al., Note 31.

[44] Drakos, Konstantinos, and Andreas Gofas. “The devil you know but are afraid to face: Underreporting bias and its distorting effects on the study of terrorism,” *Journal of Conflict Resolution* 50(2006): 714–35.
[45] LaFree, Gary. Terrorism open-source databases,” Pp. 113-134 in *Contemporary Terrorism Studies*. edited by Diego Muro and Tim Wilson (2022: New York: Oxford University Press).
[46] Grierson, Jamie. “Islamist extremism remains dominant UK terror threat, say experts,” *The Guardian*, December 10, 2019, sec. Politics. http://www.theguardian.com/politics/2019/dec/10/islamist-extremism-remains-dominant-uk-terror-threat-say-experts.
[47] King, Gary and Will Lowe. “An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design,” *International Organization* 57 (2003): 617–42; Schrodt, Philip A. and David Van Brackle. 2013. “Automated Coding of Political Event Data.” In *Handbook of Computational Approaches to Counterterrorism*, edited by V.S. Subrahmanian, 23–49. (New York, NY: Springer, 2013).
[48] Fishman, Mark. *Manufacturing the News*. (University of Texas Press, 1998).
[49] Corner, Emily, and Paul Gill. "Is there a nexus between terrorist involvement and mental health in the age of the Islamic State?" *The CTC Sentinel* 10, no. 1 (2017): 1-10.
[50] Kearns, Erin M., Allison E. Betus, and Anthony F. Lemieux. “Why Do Some Terrorist Attacks Receive More Media Attention Than Others?” *Justice Quarterly* 36 (2019): 985–1022.
[51] Findley, Michael G. and Joseph K. Young. “More Combatant Groups, More Terror?: Empirical Tests of an Outbidding Logic, *Terrorism and Political Violence*, 24 (2012): 706-721.
[52] Miller, Erin. 2020. "Global Terrorism Overview: Terrorism in 2019." College Park, MD: START (2020: July. https://www.start.umd.edu/pubs/START_GTD_GlobalTerrorismOverview2019_July2020.pdf
[53] Ibid., 8.
[54] See for example: Lindekilde, Lasse, Francis O’Connor, and Bart Schuurman. "Radicalization patterns and modes of attack planning and preparation among lone-actor terrorists: an exploratory analysis." *Behavioral Sciences of Terrorism and Political Aggression* 11, no. 2 (2019): 113-133; Gattinara, P. C., O’Connor, F., & Lindekilde, L. (2018). Italy, no country for acting alone? Lone actor radicalisation in the Neo-fascist milieu. *Perspectives on Terrorism*, *12*(6), 136-149.
[55] Gill, Paul. "The data collection challenge: Experiences studying lone-actor terrorism." *Resolve Network: Researching Violent Extremism Series* (2020).
[56] LaFree, Gary and Yesenia Yanez. “Criminological Perspectives on Extremist Radicalization and Mobilization.” In *Handbook on Radicalization and Countering Radicalization*. London: Routledge (2023); Fisher, Daren, and Erin M. Kearns. "The Theorizing of Terrorism within Criminology." *Journal of Research in Crime and Delinquency* (2023).
[57] Clemmow, Caitlin, Sandy Schumann, Nadine L. Salman, and Paul Gill. "The base rate study: Developing base rates for risk factors and indicators for engagement in violent extremism." *Journal of forensic sciences* 65, no. 3 (2020): 865-881.
[58] Global Terrorism Database. *Codebook: Methodology, Inclusion Criteria and Variables*. (2021): 10. https://www.start.umd.edu/publication/global-terrorism-overview-terrorism-2019
[59] Fred Burton and Scott Stewart, "The ‘Lone Wolf’ Disconnect," *STRATFOR*, January 30, 2008; Edwin Bakker and B. A. de Graaf, "Preventing Lone Wolf Terrorism: Some CT Approaches Addressed," *Perspectives* 5, no. 5-6 (2011): 8.

[60] Ramon Spaaij, *Understanding Lone Wolf Terrorism: Global Patterns, Motivations and Prevention* (Dordrecht: Springer, 2011); Petter Nesser, "Single Actor Terrorism: Scope, Characteristics and Explanations," *Perspectives on Terrorism* 6, no. 6 (2012); Matthew Feldman, "Comparative Lone Wolf Terrorism: Toward a Heuristic Definition," *Democracy and Security* 9, no. 3 (2013): 270–86.

[61] LaFree, Dugan and Miller 2015:17-18.

[62] LaFree, Gary. Recent Developments in the Quantitative Study of Terrorism. Vienna, Austria: *European Institute for Counter Terrorism and Conflict Prevention* (forthcoming 2023).