BMJ Global Health

Overview of attacks against civilian infrastructure during the Syrian civil war, 2012–2018

Evan Muzzall,[1] Brian Perlman,[2] Leonard S Rubenstein,[3] Rohini J Haar [4]

To cite: Muzzall E, Perlman B, Rubenstein LS, *et al*. Overview of attacks against civilian infrastructure during the Syrian civil war, 2012–2018. *BMJ Global Health* 2021;**6**:e006384. doi:10.1136/bmjgh-2021-006384

Handling editor Seye Abimbola

Received 4 June 2021
Accepted 1 September 2021

[1]Stanford University Libraries, Stanford University, Stanford, California, USA
[2]Human Rights Center, School of Law, University of California Berkeley, Berkeley, California, USA
[3]Center for Humanitarian Health, Department of Epidemiology, School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA
[4]Division of Epidemiology and Biostatistics, School of Public Health, University of California Berkeley, Berkeley, California, USA

Correspondence to
Dr Rohini J Haar;
rohinihaar@berkeley.edu

ABSTRACT

Background Hundreds of thousands of people have been killed during the Syrian civil war and millions more displaced along with an unconscionable amount of destroyed civilian infrastructure.

Methods We aggregate attack data from Airwars, Physicians for Human Rights and the Safeguarding Health in Conflict Coalition/Insecurity Insight to provide a summary of attacks against civilian infrastructure during the years 2012–2018. Specifically, we explore relationships between date of attack, governorate, perpetrator and weapon for 2689 attacks against five civilian infrastructure classes: healthcare, private, public, school and unknown. Multiple correspondence analysis (MCA) via squared cosine distance, k-means clustering of the MCA row coordinates, binomial lasso classification and Cramer's V coefficients are used to produce and investigate these correlations.

Results Frequencies and proportions of attacks against the civilian infrastructure classes by year, governorate, perpetrator and weapon are presented. MCA results identify variation along the first two dimensions for the variables year, governorate, perpetrator and healthcare infrastructure in four topics of interest: (1) Syrian government attacks against healthcare infrastructure, (2) US-led Coalition offensives in Raqqa in 2017, (3) Russian violence in Aleppo in 2016 and (4) airstrikes on non-healthcare infrastructure. These topics of interest are supported by results of the k-means clustering, binomial lasso classification and Cramer's V coefficients.

Discussion Findings suggest that violence against healthcare infrastructure correlates strongly with specific perpetrators. We hope that the results of this study provide researchers with valuable data and insights that can be used in future analyses to better understand the Syrian conflict.

Key questions

What is already known?

- There is extensive and compelling evidence demonstrating the violence against civilians and civilian infrastructure in Syria since the conflict started in 2011. Previous research and reports already document the dramatic number of medical facilities that have been attacked by airstrikes in particular.

What are the new findings?

- This study collates these data from three sources and presents the first known overview of civilian strikes by describing the relationships between the year, location, perpetrator, weapon and affected infrastructure types. Analyses of these data indicate that, among other correlations, the Syrian government is correlated with attacks that specifically target healthcare facilities, the Russian military is correlated with violence in Aleppo, and the US Coalition is correlated with violence in Raqqa. These results provide epidemiological and data-driven support for empirical evidence of civilian violence.

What do the new findings imply?

- The study offers important contributions into how data science methods can be used to analyse complex conflict data. The associations we identified support physical and documentary evidence of civil destruction and underscore that the violence against civilians in Syria is widespread. These data are critical to supporting the Syrian people as they look towards rebuilding health systems, reconstructing civil societies and seeking accountability for violations of international humanitarian law.

INTRODUCTION

Violence against civilians remains a frequent aspect of modern conflict despite its short-term and long-term adverse consequences. This violence takes many forms and is often characterised as physical trauma inflicted on individuals but also includes the destruction of civilian infrastructure and non-physical attacks that can disrupt food chains, water and sanitation, communications and social structures that impact the lives, livelihoods, health and welfare of civilian populations. Understanding civilian violence is critical to informing relief efforts, advocating for better policies and protections and for holding parties to the conflict accountable for violations of international humanitarian law (IHL).[1] Despite its importance to understanding the scope and scale of a conflict, as well as how to mitigate the trauma, measuring

violence against civilians remains challenging. Internationally, despite the obligations of IHL to protect non-combatants and limit civilian harm, there is a limited formal documentation of the obligation of warring parties to measure or report harm they may have caused. The USA, for example, has no robust mechanism to define the civilian impacts of its military actions, measure their occurrence in any transparent way, mitigate harm or compensate the victims—and most other countries are even farther behind in these measures.[2]

Despite international protections and requirements for proportionality, violence against civilians continues unabated. The experiences of the Syrian people over the past 10 years are perhaps the most notable recent example of catastrophic and complex civilian violence.[3–12] Despite its scale and duration, civilian violence in the Syrian war is not well understood and there are no reliable estimates of civilian mortality. Estimates of population movements were limited before the Humanitarian Needs Assessment Programme began in 2016 and the United Nations stopped reporting civilian casualty data early in the conflict.[13–15]

Civil society groups including the Syrian Network for Human Rights, the Syrian Observatory for Human Rights, and the Violations Documentation Center document civilian harm by providing information about deaths and injuries but these data are limited by population movements, irregular counting on the ground and challenges in verification.[16–20] Since civilian casualties are difficult to document, examining damage to civilian infrastructure may act as an important proxy to better understand the scope and scale of the war and its impacts at the community level.[21] Several groups are documenting the destruction of water and sanitation facilities, schools, markets and social and religious structures.[22–28] Research studies are beginning to evaluate the scope of violence against particular sectors, especially healthcare which has specific protections under IHL.[6 29–32] While the data are often patchy and spread across various organisations and databases, many use investigative methods in which user-generated social media or news media reports are scrutinised and collated to identify incidents and, in some cases, verify them with multiple sources (known as open-source intelligence aka OSINT).[33] The availability of these resources and the need for more information on civilian violence together present a unique opportunity to explore how civilian infrastructure is attacked in the Syrian context and better understand the characteristics, typologies and implications of the destruction.

One of the better studied types of civilian violence in Syria is that against healthcare facilities, transports, patients and personnel. These attacks are a frequent but often under-reported part of both international and local conflicts across the globe.[34–37] Acts of violence against health result not only in loss of life and health services directly due to the violence, but also lead to long-term morbidity and mortality from disrupted health systems, the destruction of facilities, loss of supplies and the shortage of health service providers.[38–43] Since the Syrian civil war started in 2011, health services in particular have been victim to hundreds of aerial bombardments.[25] Additionally, personnel have been arrested and tortured, ambulances and vaccine convoys looted and destroyed, and patients killed and blocked from access.[44–52] While many involved parties, including the governments of Syria, Russia, the USA and multiple non-governmental groups including ISIS, al-Nusra Front and others, are implicated in such violence, by far the vast majority of strikes on health have been publicly attributed to the aligned Russian and/or Syrian government forces.[53 54]

We aim to better understand civilian violence in Syria through the lens of infrastructure attacks and explore potential relationships between civilian violence and the associated factors of time, governorate, perpetrator and weapon to help expand the available perspectives and methods used for conflict research. We implement a novel dataset compiled from three conflict databases that document violence by international actors and violence against health facilities. In addition to summary statistics, we present multivariate correlations via multiple correspondence analysis (MCA) with squared cosine variable representation measures. To investigate the reliability of the MCA, we also provide k-means cluster classification accuracies (purity scores) of the MCA row coordinates, nested k-fold cross-validated binomial lasso classification with various performance metric evaluations, and Cramer's V statistics and associated confidence intervals (CIs).

METHODS

Dataset

We use data from three databases: Airwars, Physicians for Human Rights (PHR) and Safeguarding Health in Conflict Coalition/Insecurity Insight (SHCC/II) to investigate attacks against different infrastructure categories in Syria. These data were selected based on the following criteria: availability (either publicly available online or via consultation with the organisation), level of specificity to characterise unique incidents and verification procedures for how data were collected and presented. These data contain information about attacks against various governorates, dates, infrastructure types and weapons used along with the perpetrators responsible. We included data from January 2012 through December 2018. Little data were available for later years at the time of this analysis.

Airwars is a non-profit monitoring and advocacy group that tracks and assesses claims of civilian non-combatant casualties from international military actions based on open-source investigations from a wide range of sources as well as military reports.[55] Airwars shared data about Russian strikes in Syria beginning on 30 September 2015 through 12 October 2018, and on US-led Coalition strikes from 23 September 2014 to 19 October 2018.[55] PHR is a non-profit human rights organisation and has developed a publicly available registry and mapping of

attacks on health in Syria since March 2011 using open-source information from press reports and social media corroborated by field sources.[25 26] We manually extracted data available on the public online map for this study. For the year 2018, we supplemented this dataset with data that we received electronically from SHCC/II, a coalition that collects and collates data on security incidents in humanitarian settings by collaborating directly with humanitarian and human rights organisations and conducting open-source investigations .[23] For its attacks on health data, SHCC/II identified and consolidated data from sources provided by PHR, the Syrian American Medical Society and others.[24 28] Airwars' event-based data is publicly available in incident reports at https://airwars.org along with more detailed methodologies.[55] PHR's incident data is available as brief reports on an interactive map at https://syriamap.phr.org/#/en along with its methodologies.[26] SHCC/ datasets on health attacks are available for download on the Humanitarian Data Exchange at https://data.humdata.org/dataset/shcchealthcare-dataset by year.

The dataset is represented by four categorical variables (year, governorate, perpetrator and weapon) along with five binary present/absent variables for infrastructure type (health, private, public, school and unknown) coded as 1 or 0. It contains 2502 rows that contain 2689 total attacks. PHR and SHCC only document attacks on healthcare facilities while Airwars documents civil infrastructure attacks in qualitative incident reports from which the target structure was extracted and classified into one of the category types we developed, a priori. The dataset contains many strikes against 'Unknown' infrastructure types, defined as a strike that was recorded by one of the agencies but with undetermined specificity of the locations and degrees of destruction and potentially executed in an indiscriminate manner. The dataset used in our analyses is available online (https://github.com/EastBayEv/syria_overview_bmjgh2021).

Statistical methods

We included a multifaceted statistical framework to investigate characteristics of the types of attacks against the different infrastructure types: summary statistics and visualisations, MCA, k-means clustering of the MCA row coordinates, binomial lasso classification and Cramer's V statistics. Summary statistics consist of bar plots with the total number of attacks by year, governorate, perpetrator and weapon along with frequencies and proportions for the five different infrastructure categories. All statistical analyses were performed using R V.4.1.0 and the stats package[56] along with the others cited later.

Multiple correspondence analysis

MCA is an unsupervised (ie, it does not seek to predict an outcome) dimension reduction and data mining technique to investigate latent structures not readily observable in tabular data. It is similar to principal component analysis but is ideal for nominal categorical data and was thus ideally suited for our dataset. Although strikes were recorded as 1s and 0s, and can be interpreted mathematically as such, they translated to 'present' or 'absent' categories for the MCA and Cramer's V analyses.[57–60] Abdi and Valentin succinctly summarised MCA:

> Correspondence analysis (CA) is a generalization of principal component analysis tailored to handle nominal variables. CA is traditionally used to analyze contingency tables, but is also often used with data matrices that comprise only nonnegative data. CA decomposes the chi-square statistics associated to the data table into two sets of orthogonal components that describe, respectively, the pattern of associations between the elements of the rows and between the elements of the columns of the data table. When the data table is a set of observations described by a set of nominal variables, CA becomes multiple correspondence analysis (MCA).[61]

We use MCA to distill variation between the different categorical groups within columns of the dataset that can be reinterpreted as single point coordinates along the new linearly combined axes of variation. These coordinates can be plotted to illustrate exploratory relationships that would be otherwise difficult to ascertain and that might not be well suited to significance testing due to their categorical and often imbalanced nature. Squared cosine distance was used to measure the confidence of the point placements in the reduced dimensional space.[61 62] This value ranges between 0 and 1 and the larger the squared cosine value the more confident we are about a particular point's placement, as Rodriguez-Sabate and colleagues explained:

> The quality of a variable (cosine2) represents the distribution of the inertia of this variable across dimensions (normalized between 0 and 1 which represents the whole inertia of the variable; a value lower than 0.1 indicates a poor representation of the variable in the computed dimensions). The term cosine-squared refers to the fact that this value is also the squared cosine value of the angle the point makes with the specific dimension (it may also be interpreted as the correlation of the respective point with the respective dimension) … Eigenvalues represent the relative relevance of each dimension to the total inertia (it is normalized to 1 which represents all the information of all the variables in all the dimensions). The highest eigenvalue was always in the first dimension, progressively decreasing across the following dimensions.[63]

Inertia is the total variation within the dataset and a scree plot is traditionally used to examine which reduced dimensions of variation to include in analyses. We focus exclusively on the first and second dimensions because variation was more evenly dispersed across the many remaining dimensions. This is not to say that other relationships do not exist in the other remaining dimensions, so we also provide a figure that provides squared cosine values across the first 20 dimensions. However, interpreting results of unsupervised techniques such as the MCA is challenging. It is hard to be sure if point locations are the result of the nature of the dataset itself, its

highly dispersed, sparse and categorical nature, and/or imbalanced class distributions. To gain a better understanding of the results we also implement purity scores across multiple k-means clusters on the row coordinates from the MCA to examine classification accuracy of the different variables, binomial lasso classification to look at related variables for predicting attacks on healthcare infrastructure, and Cramer's V statistics to try and better understand the relationships illustrated by the MCA. MCA was coded using the FactoMineR, factoextra, corrplot and ggplot2 packages.[58 60 64–66]

k-means clustering

We perform clustering on the row coordinates produced by the MCA. k-means clustering is an approach to divide row observations into smaller and smaller subsets of like types to form point clusters based on their shared distances to the centre of a cluster, called a centroid. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable k. k-means clustering requires the user to specify the number of clusters and eight was the optimal number of clusters based on the silhouette method (binary distance dissimilarity measure and k-means analysis method).[67] Cluster homogeneity was evaluated using the purity score, an external evaluation metric for assessing cluster classification accuracy, for two through nine clusters.[68]

Binomial lasso classification

We then focused specifically on the potential validity of attacks against healthcare because of their far-reaching and devastating impacts on civilian life. We used binomial lasso classification (least absolute shrinkage and selection operator, a form of penalised regression) to predict whether or not a strike occurred against civilian healthcare infrastructure in Syria.[69] Lasso penalises the beta coefficients to zero for independent variables that are not related to the outcome, thus removing them from the model and ostensibly producing a model with less predictors that is easier to interpret. In our case, the outcome is binomial because we predict whether or not a strike occurred against healthcare infrastructure.

Like other supervised machine learning algorithms (ie, ones that seek to predict an outcome variable), binomial lasso classification relies on data splitting procedures to assess model performance. In the simplest cases, data are divided into training and test sets. Normally this entails partitioning a (often majority) portion of the original data into the training set so that the algorithm can build a model that learns the relationships between the outcome (in our case whether a healthcare facility was struck or not) and the characteristics of the other independent variables as predictors. Performance is evaluated through confusion matrix derivation metrics such as classification accuracy, sensitivity, specificity, precision, false discovery rate and so on. Then, the trained model is applied to the remaining portion of the withheld/test data, in order to predict the outcome on data that the model has not yet seen. In this manner, the algorithm can be used to evaluate how well the trained model could learn not just the relationships in the known data but also predict unknown outcomes and generalise to new data without knowing the ground-truth labels in advance.

Performance of the lasso model was evaluated using a variety of metrics based on nested 20-fold cross-validation, which repeats the train/test splitting process described above 20 times in a stratified manner and that rotates different portions of the data for training and testing to ensure that all observations are part of a test set at least once.[70] The lasso allows for further utility through model selection. A model can be selected by choosing values of lambda from the ordinary least squares solution that includes coefficients for all independent variables, to a simpler model that includes only the non-zeroed-out coefficients, but that is within one standard error of the minimum error value.

To account for potential imbalance between attacks on healthcare (601 instances) versus non-healthcare attacks (2088 instances), we calculated area under the curve-receiver operator characteristic (AUC-ROC). We also oversampled the data so that the number of attacks against healthcare were approximately the same as the number of non-healthcare attacks through generation of 'synthetic positive instances using ADASYN algorithm. The number of majority neighbors of each minority instance determines the number of synthetic instances generated from the minority instance'[71] and fit the algorithm a second time using the mikropml R package[72] to produce 15 additional performance metrics for comparison with the original model. For both fitting processes, the categorical variables year, governorate, perpetrator and weapon were one hot encoded to indicator variables; the five infrastructure type variables were already represented by 1s and 0s and represented categorically to indicate if a strike was recorded as present or absent, respectively.

Cramer's V

Finally, we measured the strength of associations between the columns of the dataset via Cramer's V coefficient (phi or φ_c). This value ranges between 0 and 1 and is similar to a χ^2 test of independence for effect size but is interpreted like a correlation coefficient for large tables.[73] Values of 0–0.30 indicate a weak correlation, 0.30–0.70 indicate a moderate correlation and 0.70–1.00 indicate a strong correlation. Confidence intervals were computed using non-central χ^2 method at 95% confidence.

Cramer's V statistics were calculated using the DescTools R package.[74]

RESULTS

The US Coalition, Russian military and the Syrian government account for perpetrating 2446 out of 2689—or 91%—of the attacks in the dataset. Barplots showing total number of attacks by year, governorate, perpetrator and

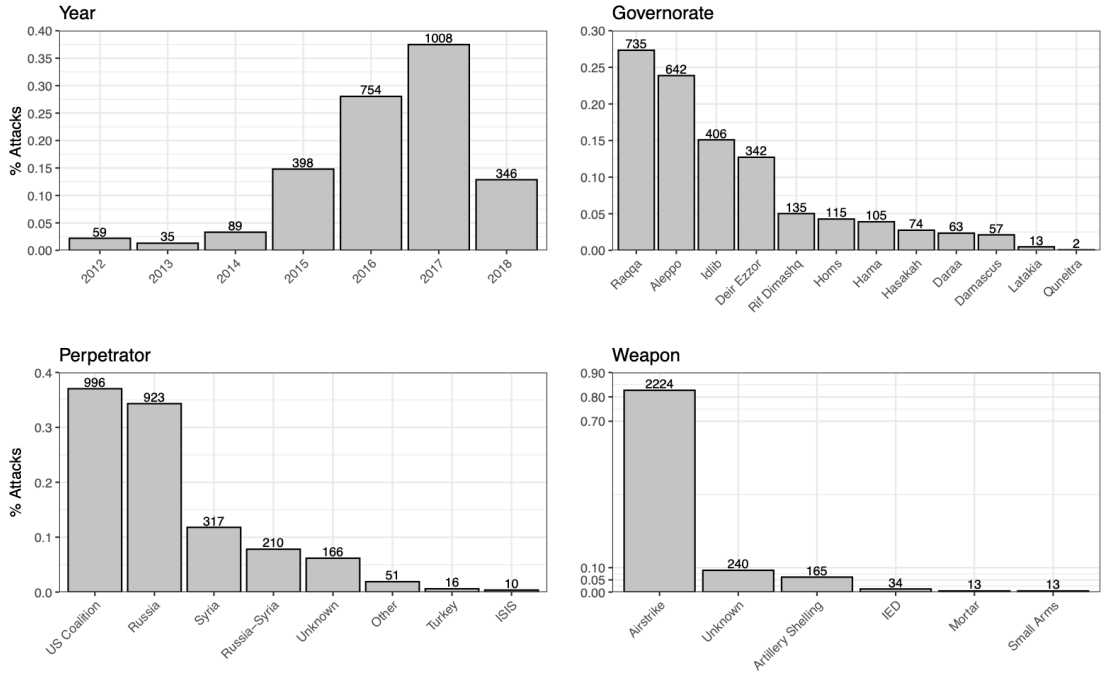

Figure 1 Barplots showing total number of attacks by year, governorate, perpetrator and weapon. Category is shown on the x-axis and percentage of total attacks is shown on the y-axis. Attack frequency is the number on top of each bar.

weapon are shown in figure 1. Frequencies and proportions of infrastructure attacks by year, governorate, perpetrator and weapon are presented in table 1. The sum of each infrastructure type column by the four row subgroups (year, governorate, perpetrator and weapon) all sum to the same values shown in the Total and Total of proportion rows, thereby contributing to the same overall percentages of the 2689 strikes although the frequencies vary by the four row subgroups. Unknown infrastructure types were the most frequently struck (37%), followed by health (22%), public (20%) and private (18%) types. Schools (3%) were targeted much less often based on these data.

The row totals provide an informative glimpse into the distribution of these attacks on civilian infrastructure. According to these data, 15% and 28% of attacks occurred during 2015 and 2016, respectively, as the conflict was escalating. Thirty-seven per cent of attacks occurred during 2017, the height of the conflict, and followed by 13% of attacks in 2018. By governorate, the hardest hit were Raqqa (27%), Aleppo (24%), Idlib (15%) and Deir Ezzor (13%), with the other governorates being struck less than 10% of the time. Furthermore, the US Coalition was reportedly responsible for 37% of these attacks, while Russia was responsible for 34%, Syria for 12%, and the Russia–Syria Coalition responsible for 8% of attacks; the other perpetrators were responsible for less than 8% of attacks. Airstrikes were the overwhelming method of violence employed and were used for 83% of strikes.

Multiple correspondence analysis

A multivariate view elucidates additional patterns in these data. The first two dimensions of variation from MCA analysis contain 9.3% and 6.5% of variation explained, respectively. Relationships between the nine variables in the dataset are shown in the MCA variable plot in figure 2. This figure shows that governorate, year and perpetrator contribute most to variation along the x-axis and y-axis together, while health infrastructure contributes to variation only along the x-axis. Weapon contributes considerably less to variation along both axes, while unknown infrastructure type contributes only slightly along the x-axis.

More profound relationships between year, governorate, perpetrator and affected infrastructure type are shown in figure 3. For a particular infrastructure type, a '1' indicates that this infrastructure class was targeted and a '0' means it was not targeted (eg, 'Health 1' vs 'Health 0', 'School 1' vs 'School 0', etc). Here, the closeness of the text demonstrates the degree of association between the variables, while the colour of the text indicates confidence in their placement with darker text representing greater confidence as measured by squared-cosine similarity. Four topics of interest emerge: (1) US Coalition led strikes against Raqqa in 2017 (upper-left portion of the plot), (2) Russian campaigns in Aleppo in 2016 (bottom left), (3) Syrian government attacks against healthcare facilities (right side) and (4) airstrikes across non-health infrastructure (left side). Interestingly, attacks attributed to the Russia–Syria Coalition are found in between the perpetrators Russia and Syria in the lower-right portion of the plot, although with lesser confidence. Variable representation quality for 20 axes of variation from the MCA is shown in figure 4. Larger and darker circles indicate a higher squared cosine value along that axis.

Table 1 Affected infrastructure type by year, governorate, perpetrator and weapon, 2012–2018

Variable	Health	Private	Public	School	Unknown	Total	Percentage of attacks
Year							
2012	58	0	0	1	0	59	0.02
2013	35	0	0	0	0	35	0.01
2014	77	0	9	0	3	89	0.03
2015	121	82	79	14	102	398	0.15
2016	108	159	145	22	320	754	0.28
2017	49	201	221	32	505	1008	0.37
2018	153	37	91	6	59	346	0.13
Governorate							
Aleppo	148	104	109	15	266	642	0.24
Damascus	28	6	5	3	15	57	0.02
Daraa	37	4	9	1	12	63	0.02
Deir Ezzor	38	69	105	10	120	342	0.13
Hama	43	12	17	1	32	105	0.04
Hasakah	6	27	13	5	23	74	0.03
Homs	31	24	17	2	41	115	0.04
Idlib	138	45	86	16	121	406	0.15
Latakia	11	1	0	0	1	13	0.00
Quneitra	1	0	1	0	0	2	0.00
Raqqa	21	184	166	22	342	735	0.27
Rif Dimashq	99	3	17	0	16	135	0.05
Perpetrator							
ISIS	7	1	1	0	1	10	0.00
Other	21	3	9	0	18	51	0.02
Russia	54	188	163	36	482	923	0.34
Russia–Syria	156	6	36	0	12	210	0.08
Syria	274	3	27	1	12	317	0.12
Turkey	3	0	3	1	9	16	0.01
Unknown	62	7	30	3	64	166	0.06
US Coalition	24	271	276	34	391	996	0.37
Weapon							
Airstrike	396	447	469	69	843	2224	0.83
Artillery Shelling	72	14	27	2	50	165	0.06
IED	19	1	11	0	3	34	0.01
Mortar	9	1	2	0	1	13	0.00
Small Arms	2	2	6	0	3	13	0.00
Unknown	103	14	30	4	89	240	0.09
Total	601	479	545	75	989	2689	1
Percentage of attacks	0.22	0.18	0.20	0.03	0.37	1	

k-means classification

Results of the k-means cluster purity scores show that the school variable had the highest mean classification value (0.97), followed by health (0.92), weapon (0.84), private (0.81) and perpetrator (0.81). Public, unknown, year and governorate all have scores lower than 0.80. Overall, these mean scores are similar to the ones shown for the supposed optimum number of eight clusters (table 2). These scores suggest that several variables in the data, including attacks against healthcare, cluster similarly even when parameters are modified.

Binomial lasso classification

AUC-ROC=0.985 for the most regularised lasso model with nested 20-fold cross-validation that is within one SE of the minimum (lambda=0.054, log lambda=−2.923). AUC-ROC=0.972 for prediction for the holdout/test dataset (750 randomly selected observations from the

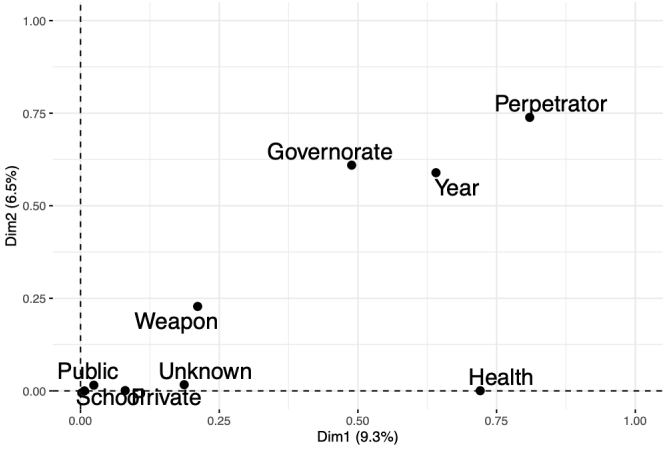

Figure 2 Multiple correspondence analysis variables plot to illustrate the relationships of the variables from the dataset along the first two dimensions of variation.

original dataset). Absolute values for the simplest model within one SE of the minimum contained seven non-zero coefficients (out of the 43 total) and are shown in table 3. Attacked unknown, public and private infrastructure types have the highest coefficients, followed by perpetrators Syria, Russia–Syria Coalition and US Coalition. One thousand eight hundred and seventy-seven synthetic attacks against healthcare (compared with the actual 601) were used for the balanced comparison against the actual 1901 attacks *not* against healthcare infrastructure in the oversampled data cross-validated lasso model. Performance metrics for this second model are found in table 4. The binomial lasso classification models suggest that certain infrastructure types and specific perpetrators tend to characterise attacks on healthcare.

Cramer's V

Cramer's V coefficients and confidence intervals are shown in table 5. The coefficient is displayed in the upper triangle. Coefficients greater than 0.30 are denoted in bold text to indicate a moderate or strong correlation. Their CIs are shown in the lower triangle; grey shading indicates a CI for a moderate or strong coefficient from the upper triangle. Of the 36 total pairwise comparisons, 24 are weak, 11 are moderate, 1 is strong. Perpetrator correlates strongly (V=0.75) with attacks against health infrastructure, while attacks on health infrastructure correlate moderately with year (V=0.56) and governorate (V=0.46). Unknown infrastructure type correlates with a few of the other variables but is difficult to interpret due to its unknown nature.

DISCUSSION

This exploratory study illustrates the potential of conflict data, both publicly available and collected and shared by non-profit organisations, to contribute to the evidence base of civilian harm and on violations of IHL in times of conflict. Through the presentation of novel, aggregated data specifically about attacks on civilian infrastructure and violence against healthcare in the Syrian civil war, we provide a new lens on the many incidents that caused grave civilian harm and profound disrespect for

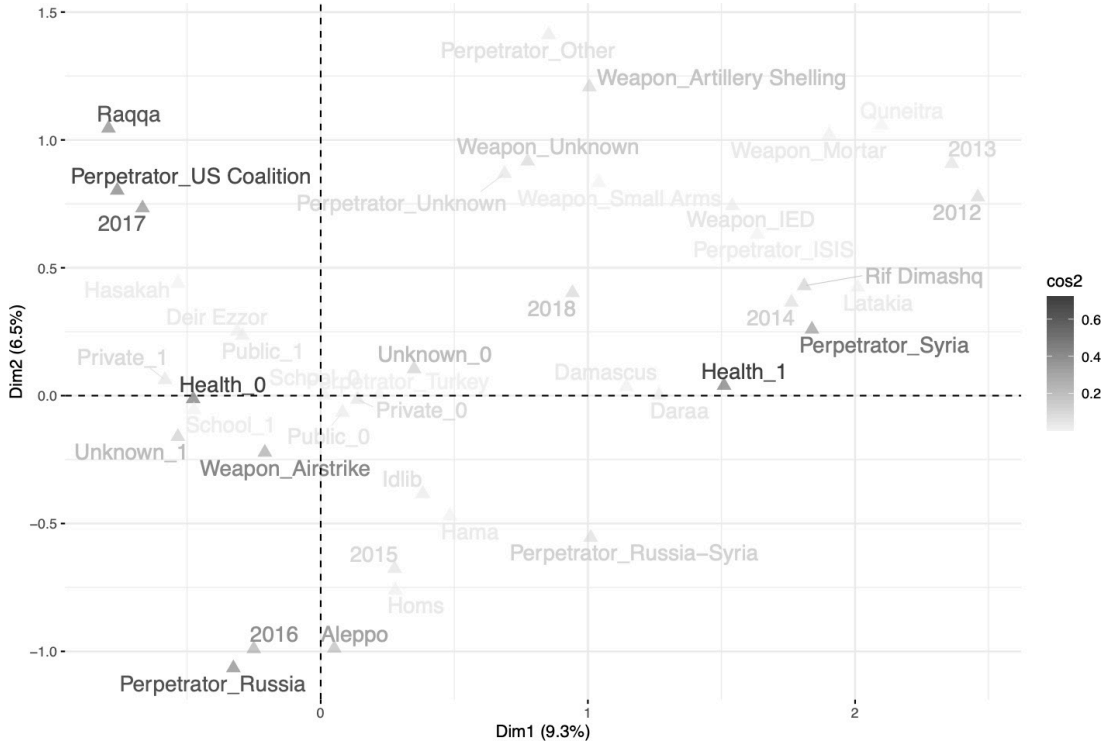

Figure 3 Multiple correspondence analysis squared cosine distance (cos2) plot. Very high confidence of point placement is labelled with darker text, high confidence placement with grey and moderate confidence with light grey text. Note disparate locations of whether healthcare was attacked (Health 1) or not (Health 0) along the x-axis.

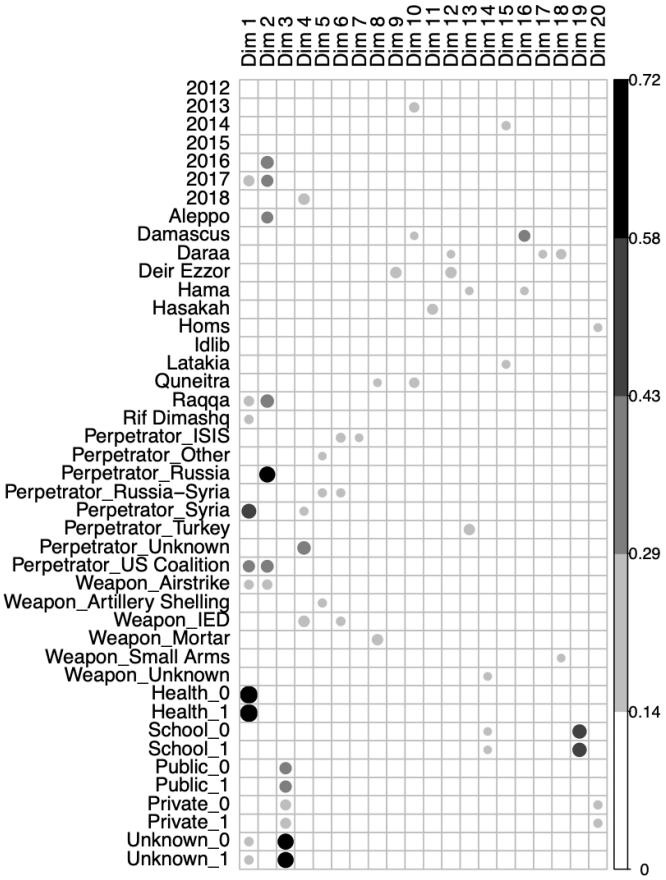

Figure 4 Squared cosine distance plot showing degree of representation for all variable categories across 20 axes of variation defined by the multiple correspondence analysis. Larger and darker circles indicate better variable representation on that particular axis.

principles of proportionality and protection of civilians enshrined in IHL. As data collection in conflict settings advances across the globe, standards for data collection and coordination must be strengthened to fully leverage the promise of data science methodologies for analysis of conflict data. This study underscores several correlations of interest, particularly that the Syrian and/or Russian governments are primarily responsible for the majority of attacks against health in Syria (81%). These findings highlight that these strikes may not simply be the collateral impact of modern conflict or of civilian strikes.

Importantly, visualising conflict data in multidimensional space through unsupervised, exploratory methods such as MCA is useful to gain high-level clarity of variables of interest. In this manner, the human eye can potentially better understand the many complicated relationships from war contexts in effective and preattentive ways that might be missed by both univariate and more complex modelling approaches. The four topics of interest identified by the MCA squared-cosine plot (figure 3) support previous findings that the Syrian government is largely responsible for strikes on healthcare facilities in their own country as well as links between the Syrian and Russian governments to attacks on Syrian health infrastructure.[32 75 76] While many of these initiatives used analysis of detailed information regarding a single or several specific instances of violence, the aggregated data used in this study strongly supports these assumptions. Recent research and human rights reporting present findings consistent with those in this analysis: that violence against

Table 3 Non-zero coefficients for the one SE nested cross-validated lasso model

Variable	Coefficient
unknown_1	2.48
public_1	1.62
private_1	1.43
perpetrator_syria	1.38
perpetrator_russia_syria	0.88
perpetrator_us_coalition	0.24
(Intercept)	0.11

Larger coefficients indicate that a variable stayed in the model for a longer time before being zeroed out by the penalisation and are thus more closely related to the outcome that healthcare infrastructure was attacked.

Table 2 k-means cluster purity scores (classification accuracies) for varying numbers of clusters (2 through 9) based on row coordinates produced by multiple correspondence analysis

Variable	2	3	4	5	6	7	8	9	Mean score
School	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
Health	0.94	0.94	0.92	0.91	0.92	0.92	0.92	0.9	0.92
Weapon	0.82	0.82	0.82	0.83	0.83	0.86	0.85	0.87	0.84
Private	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.83	0.81
Perpetrator	0.48	0.78	0.83	0.92	0.92	0.84	0.88	0.8	0.81
Public	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.81	0.78
Unknown	0.64	0.65	0.66	0.65	0.65	0.65	0.66	0.84	0.68
Year	0.46	0.58	0.61	0.57	0.57	0.58	0.57	0.64	0.57
Governorate	0.33	0.5	0.52	0.49	0.5	0.49	0.5	0.52	0.48

Eight was the optimal k as identified by the silhouette method.

Table 4 Performance metrics for the adaptive synthetic (ADASYN) oversampled cross- validated lasso model for attacks on healthcare

Performance metric	Healthcare infrastructure
cv_metric_AUC	0.9467
logLoss	0.2963
AUC	0.9595
prAUC	0.9590
Accuracy	0.8976
Kappa	0.7952
F1	0.8973
Sensitivity	0.9005
Specificity	0.8947
Pos_Pred_Value	0.8942
Neg_Pred_Value	0.9011
Precision	0.8942
Recall	0.9005
Detection_Rate	0.4475
Balanced_Accuracy	0.8976

To balance the samples, 1877 attacks against healthcare were compared against the original 1901 non-healthcare attacks.
See the mikropml vignette for definitions of these metrics along with the R coding walkthrough (https://cran.r-project.org/web/packages/mikropml/vignettes/introduction.html).

healthcare has been significant in the Syrian conflict, and that the vast majority of airstrikes on health facilities were perpetrated by Syrian government and Russian military forces. PHR and Human Rights Watch investigations point to Syrian and Russian government forces in eyewitness testimony[77–79] and several open-source investigations verify claims of Russian involvement in airstrikes in Syria.[80] With regard to US-led Coalition strikes on civilians in Raqqa in 2017, Amnesty International has detailed how the coalition conducted its air campaign in 2017 to oust Islamic State from its de-facto capital, resulting in thousands of civilian deaths.[45]

Results of statistical analyses support these assertions, especially for attacks against health. First, MCA offers an overview into the multivariate correlations of the variables in the data. Notably, it reveals that while perpetrator, year and governorate have strong discriminatory power along the first two dimensions of variation, health is only discriminatory along the first dimension (figure 3). Furthermore, we see attacks against health (Health_1) and the state actor (Perpetrator_Syria) not just closely aligned in the positive direction of the first dimension in, but also with high confidence in their placement as represented by the squared cosine score to indicate the appropriateness of this relationship. This can also be roughly understood from the summary statistics in table 1. Second, support is offered by the k-means cluster accuracy scores in table 2, which indicate that perpetrator clusters at 0.88 accuracy at the optimal eight number of clusters (along with a 0.81 mean score). Here, only School (0.97) and Health (0.92) have higher accuracies for the eight cluster model. Third, binomial lasso classification indicates that whether or not healthcare was attacked is readily predictable, but that using a penalised model highlights variables that are most related to the outcome as shown in table 3. Importantly, we see that struck unknown, public and private infrastructures correlate the most with affected health infrastructure, followed closely by the Syrian government, the Syria–Russia Coalition and the US Coalition and suggests that these variables are strongly tied to attacks on health. Finally, the lone strong Cramer's V coefficient (V=0.75) in table 5 is found between health infrastructure and perpetrator. It is difficult to ascertain how attacks against unknown infrastructure types might influence correlations between perpetrator and attacks against known infrastructure.

In September 2015, the Russian government approved a request to begin airstrikes in Syria and these first instances of Russian strikes appear in our dataset in September 2015.[81] Ultimately, in December 2016, the Russian-backed Syrian government fully recaptured Aleppo from the opposition.[55 82] Russian involvement

Table 5 Cramer's V coefficients and CIs for correlations between variables in the dataset

	Year	Governorate	Perpetrator	Weapon	Health	School	Public	Private	Unknown
Year	1	**0.33**	**0.37**	0.24	**0.56**	0.05	0.13	0.17	**0.34**
Governorate	0.31 to 0.34	1	**0.35**	0.17	**0.46**	0.08	0.15	0.21	0.22
Perpetrator	0.35 to 0.38	0.33 to 0.36	1	**0.34**	**0.75**	0.10	0.18	0.29	**0.39**
Weapon	0.22 to 0.25	0.14 to 0.17	0.32 to 0.35	1	0.25	0.05	0.10	0.15	0.10
Health	0.52 to 0.60	0.41 to 0.49	0.71 to 0.78	0.20 to 0.28	1	0.05	0.25	0.25	**0.45**
School	0.00 to 0.08	0.00 to 0.09	0.04 to 0.13	0.00 to 0.07	0.02 to 0.09	1	0.08	0.08	0.14
Public	0.08 to 0.16	0.10 to 0.18	0.13 to 0.21	0.05 to 0.13	0.21 to 0.29	0.04 to 0.12	1	0.02	**0.42**
Private	0.12 to 0.20	0.16 to 0.24	0.25 to 0.33	0.10 to 0.18	0.21 to 0.29	0.04 to 0.11	0.00 to 0.06	1	**0.39**
Unknown	0.30 to 0.38	0.17 to 0.25	0.35 to 0.43	0.05 to 0.13	0.42 to 0.49	0.10 to 0.18	0.38 to 0.46	0.35 to 0.43	1

Coefficients are shown in the upper triangle, with bold text to represent >0.3. Shaded grey cells in the lower triangle show CIs that correspond to the coefficients >0.3 from the upper triangle.

in the air campaign to assist the Syrian government in retaking Aleppo during 2016 is also well-documented.[55] Our results also ostensibly confirm that Russian government forces conducted direct airstrikes on Syrian soil, in addition to militarily and financially supporting the Syrian government, which carried out its own strikes. Our findings represent the first attempt to quantify how much infrastructure has been destroyed with prejudice across Syria. While much of the attention has, for good reason, focused on the protection of health, the destruction of life and other infrastructure (eg, water/sanitation and power facilities, rail, roads, parks, ports, telecommunications, religious, cultural, and entertainment facilities, and schools) potentially highlights the systematic and catastrophic breakdown of the social order.[83] These areas of civilian violence require more documentation and analysis. Although the US Coalition is responsible for just 4% of attacks against health, it is important to note that it is responsible for 37% of all attacks in our dataset. Finally, the destruction of non-health infrastructure across Syria during the conflict is well-documented in numerous reports.[32 84 85]

This research supports the voices of many human rights advocates that contend violence against civilians in Syria must be investigated under the Rome Statute of the International Criminal Court's definitions of war crimes and crimes against humanity. The frequency of strikes on health, and the strength of correlation of the perpetrators of this violence, does not provide hard evidence of strategic violence but does suggest that there are temporo-spatial relationships between the perpetrators and type of infrastructure that was afflicted. While violations of IHL are typically seen through a legal and case-by-case lens, the statistical analyses used in this study can be a valuable tool to highlight the value of data analysis of aggregate data on many incidents for the protection of human rights. Triangulating these data with witness and documentary evidence can strengthen calls for accountability and support for victims. Stronger data collection protocols and transparency of air and ground strikes by all actors would allow for stronger methodologies to provide more opportunities to identify trends and correlations among conflict indicators.

Limitations

Data limitations

We acknowledge that this initial exploratory study has significant weaknesses. Most importantly, key parts of conflict data are missing from our analysis because these data are not publicly available and accessible by researchers. These include, but are not limited to, ground strikes against non-health infrastructure and airstrikes perpetrated by non-international actors. Hopefully, organisations that already collect these data can share their methodologies and/or be given more resources to expand their scope of purpose to include data on additional indicators (ie, ground strikes, local armed groups, conflict timelines, local perceptions of events, etc). Augmenting the data collected in conflict in such ways will be an important next step in this sector. Furthermore, especially during a war, extracting complex qualitative information into categorical data designed for statistical analysis can never fully capture the true complexity of a conflict and the dynamic interplay between the myriad factors and variables at play. We hope that this information will be of use to other researchers for more complex and nested analyses to investigate certain components of conflict that are not directly visible.

Additionally, our multi-organisation dataset depends on the work of data collectors and personnel at civil society institutions, sometimes at great personal risk. The intrinsic limitations and potential inaccuracies of the various organisations' methodologies are embedded in our data and collaboration with additional organisations could augment the research in more informative ways. Collating, cleaning and deduplicating data that came with differing details and methodologies was challenging and also potentially introduced error. Additionally, our analyses included data that was sent to us by various organisations updated up to April 2019. Some incident information that was added to publicly available databases after that date is not in our analysis and should be used for further research.

Statistical limitations

Results were impacted by biases from the framing of our study itself, potential class imbalances in the data, and the statistical methods chosen for analysis. Historical and representation biases are hard to truly ascertain, especially because data collection in conflict settings can be dangerous. The challenges of data collection in such scenarios thus impact analysis, interpretation and the conclusions able to be drawn. First, computation of summary statistics will be biased based on the impossibilities of on-the-ground data collection in war zones. This is inevitable due to the lack of support that these agencies receive and again underscores the need for more urgency around data collection protections and safety. Second, these biases will also produce potentially deceptive correlations as identified by MCA and Cramer's V. It is important to understand where in the dataset sparsity occurs and why; in conflict it is difficult to discern if data are missing completely at random (that it is a random missing subset of the larger dataset, data generative process or data ecosystem) or missing conditionally at random (that it is missing based on conditional relationships with other variables) versus unrecorded/not able to be recorded. Two variables in our dataset—year and unknown infrastructure type—might be particularly susceptible to this. The positively skewed number of attacks by year could potentially provide false discriminatory power. However, when removed from the analysis results did not change considerably and were thus left in to provide the reader with a temporal component that would not exist if it were removed. Only those collecting data within the wars themselves have the ability to better speak about these endeavours, and must continue to be an integral part of the research process.

Finally, the impacts on exploratory analysis will permeate any validation technique if distorted baseline measures are present. In the case of lasso classifications that can produce sparse solutions (ie, only a few variables from the dataset are actually used in the model to predict the outcome), the emphasis on interpretation of the non-zero coefficients could be misleading. Or, the coefficient for a particular independent variable might be removed not because it is unrelated to the outcome, but potentially because it is correlated with another independent variable whose coefficient was penalised and removed. However, we feel that lasso is better suited for this type of analysis compared with multinomial logistic regression because the latter is more likely to poorly model the tails of the distribution. However, other dimension reduction, clustering and machine learning techniques should be incorporated to continue to better understand the human impact of conflict data.[86 87]

CONCLUSIONS

This study underscores the alarming number of incidents that impacted civilians in Syria from January 2012 through December 2018 and provides further evidence of the grave nature of civilian harm and destruction of infrastructure—and more specifically, health systems—in Syria. The findings support growing calls for investigation of and accountability for war crimes by the UN Security Council and International Criminal Court and International Court of Justice. Our research highlights the potential for other correlated factors in the Syrian conflict—especially between perpetrators and strikes on health. Our results emphasise the need for more reliable data collection, data accessibility, and better cross-agency collaboration to document more accurately and consistently these acts of violence. Future data science research should continue to include more complex statistical modelling when more informative data become available. Additionally, it is unclear how reliable correlations of health, school, private and public infrastructure are to the vague 'Unknown' category.

While it is unknown whether or how data analyses such as those reported in this paper could influence legal accountability mechanisms, the direct evidence from interviews, physical evidence and documentation, embedded in each attack analysed in this work combined together, provides a trove of evidence. The findings of this research may strengthen the overall claims regarding the attacks, the perpetrators and the trends. In addition to the data we report on, we hope that the novel methods contributions of this study have a lasting effect on understanding conflict and its consequences. Funders should better support additional data collection as well as better coordination among civil society initiatives and militaries to create well-designed, transparent and robust databases that foster more collaboration and analysis.

Twitter Rohini J Haar @rohinihaar

Acknowledgements We especially acknowledge the Syrian health workers, first responders, humanitarian actors and advocates on the front lines of this conflict who go unnamed but not unappreciated. We also thank Chris Woods of Airwars, the MENA research team at Physicians for Human Rights, Christina Wille and Helen Buck of Insecurity Insight and Safeguarding Health in Conflict Coalition for providing access to the data. Jae Yeon Kim provided helpful advice about CI calculations for the Cramer's V statistics. Sayaka Ri assisted with extraction of data from PHR's online interactive map. We thank the Human Rights Center at UC Berkeley's Open Source Investigation lab students for conducting data extraction. We also thank Stephanie Croft, Alexa Koenig, Andrea Lampros and Félim McMahon at UC Berkeley for feedback, guidance and review on open-source investigations. We thank Hazem Rihawi, Diana Rayes, Naser AlMhawish and Ahmad Tarakji for sharing their knowledge and experience about Syria and contextualising this work. We are grateful for the two anonymous reviewers who helped us considerably improve the quality of this manuscript.

Contributors LSR and RJH conceived of the project. BP led the data collation and extraction. EM conducted the analysis. EM and BP wrote the draft. RJH and LSR reviewed the draft. All authors reviewed and approve of the final manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors. Publication made possible in part by support from teh Berkeley Research Impact Initiative (BRII) sponsored by the UC Berkeley Library.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Ethics approval No identifying or personal information was obtained for this study and all information is aggregated at the governorate level by year to avoid disclosure of vulnerable data. This study was reviewed and exempted by the University of California, Berkeley's Institutional Review Board. Data from Airwars and SHCC/II were generously shared and the organisations contributing these data consented to the analysis and publication. PHR data were extracted from the public site. Of note, data shared by Airwars were a one-time snapshot and additional revisions/amendments that may be available on the Airwars site are not part of our dataset. Per our data guidelines, data presented in this study must not be used for commercial gain, nor sold or passed on to any third party without our express written consent.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. Data are available upon reasonable request. Data may be obtained from a third party and are not publicly available. The dataset used in our analyses is available at: https://github.com/EastBayEv/syria_overview_bmjgh2021. Given the political sensitivity of the research, information on individual incidents compiled by the research team using open source technology are available upon request to researchers. Data aggregated from third party sources must be requested directly from the source institutions via their websites. Data presented in this study must not be used for commercial gain, nor sold or passed on to any third party without the express written consent of Airwars, Physicians for Human Rights and Safeguarding Health in Conflict Coalition and Insecurity Insight.

ORCID iD

Rohini J Haar http://orcid.org/0000-0002-9049-984X

REFERENCES

1 ICRC. Geneva Conventions of 1949 and Additional Protocols, and their Commentaries [Internet]. International Committee of the Red Cross, 2016. Available: https://www.icrc.org/applic/ihl/ihl.nsf/vwTreaties1949.xsp [Accessed cited 2016 Apr 26].
2 Holewinski S. The progress not made on protecting civilians, 2021. Just Security. Available: https://www.justsecurity.org/74363/the-progress-not-made-on-protecting-civilians/
3 Aburas R, Najeeb A, Baageel L, *et al*. The Syrian conflict: a case study of the challenges and acute need for medical humanitarian operations for women and children internally displaced persons. *BMC Med* 2018;16:65.

4 Akbarzada S, Mackey TK. The Syrian public health and humanitarian crisis: A 'displacement' in global governance? *Glob Public Health* 2018;13:914–30.
5 Elamein M, Bower H, Valderrama C, *et al*. Attacks against health care in Syria, 2015–16: results from a real-time reporting tool. *The Lancet* 2017;390:2278–86.
6 Fouad FM, Sparrow A, Tarakji A, *et al*. Health workers and the weaponisation of health care in Syria: a preliminary inquiry for the Lancet –American University of Beirut Commission on Syria. *The Lancet* 2017;390:2516–26.
7 Guha-Sapir D, Schlüter B, Rodriguez-Llanes JM, *et al*. Patterns of civilian and child deaths due to war-related violence in Syria: a comparative analysis from the violation documentation center dataset, 2011-16. *Lancet Glob Health* 2018;6:e103–10.
8 Guha-Sapir D, Rodriguez-Llanes JM, Hicks MH, *et al*. Civilian deaths from weapons used in the Syrian conflict. *BMJ* 2015;351:h4736.
9 Heisler M, Baker E, McKay D. Attacks on health care in Syria--Normalizing violations of medical neutrality? *N Engl J Med* 2015;373:2489–91.
10 Lund A. The UN made a list of hospitals in Syria. Now they're being bombed, 2019. The century Foundation. Available: https://tcf.org/content/report/un-made-list-hospitals-syria-now-theyre-bombed/
11 Syrian American Medical Society. The failure of un security Council resolution 2286 in preventing attacks on health care in Syria, 2017. Available: https://www.sams-usa.net/reports/failure-un-security-council-resolution-2286-preventing-attacks-healthcare-syria/
12 Syrian American Medical Society. Saving lives underground: the case for underground hospitals in Syria, 2017. Available: https://d1p8u7ytnu6qui.cloudfront.net/Saving%20Lives%20Underground%20report.pdf
13 Ohlheiser A. The U.N. has stopped counting the deaths in Syria, 2014. The Atlantic. Available: https://www.theatlantic.com/international/archive/2014/01/un-stopped-updating-its-syria-death-toll/356758/
14 The News-Herald. The UN doesn't know how many people have died in Syria's civil war, 2014. Available: https://www.news-herald.com/news/the-un-doesnt-know-how-many-people-have-died-in-syrias-civil-war/article_64201c79-0c72-5a42-990e-3814b5bd3b09.html
15 OCHA. IDPs Tracking & Humanitarian Response, 2020. Available: https://www.humanitarianresponse.info/en/operations/stima/idps-tracking
16 Price M, Gohdes A, Ball P. Updated statistical analysis of documentation of killings in the Syrian Arab Republic 2014;26.
17 Airwars. Airwars: monitoring and assessing civilian harm from airpower, 2020. Available: https://airwars.org/
18 Time Statistics. Syrian center for statistics and research, 2020. Available: https://www.csr-sy.com/?id=475&sons=redirect&l=2&
19 Syrian Network for Human Rights. Violations against health sector. Available: http://sn4hr.org/blog/category/report/monthly-reports/violations-against-health-sector/
20 Syrian Observatory for Human Rights. The Syrian Observatory for human rights. Available: https://www.syriahr.com/en/
21 Ri S, Blair AH, Kim CJ, *et al*. Attacks on healthcare facilities as an indicator of violence against civilians in Syria: an exploratory analysis of open-source data. *PLoS One* 2019;14:e0217905.
22 ACLED. The Armed Conflict Location & Event Data Project: Bringing Clarity to Crisis, 2019. Available: https://acleddata.com/
23 Humanitarian Data Exchange (HDX), Insecurity Insight. Attacks on health care in countries in conflict, 2021. Available: https://data.humdata.org/dataset/shcchealthcare-dataset
24 Insecurity Insight. Health care in conflict program, 2021. Available: http://insecurityinsight.org/projects/healthcare
25 Physicans for Human Rights. A map of attacks on health care in Syria, 2021. Available: http://syriamap.phr.org/#/en
26 Physicians for Human Rights. Anatomy of a crisis MAP methodology, 2017. Available: https://s3.amazonaws.com/PHR_syria_map/methodology.pdf
27 Violations Documentation Center. Center for documentation of violations in Syria, 2021. Available: https://www.vdc-sy.info/index.php/en/about
28 Safeguarding Health in Conflict. Impunity remains: attacks on health care in 23 countries in conflict, 2019. Available: https://www.safeguardinghealth.org/sites/shcc/files/SHCC2019final.pdf
29 Amnesty International. Syria: "Nowhere is safe for us': Unlawful attacks and mass displacement in North-west Syria, 2020. Available: https://www.amnesty.org/en/documents/document/?indexNumber=MDE24%2f2089%2f2020&language=en
30 Footer KHA, Clouse E, Rayes D, *et al*. Qualitative accounts from Syrian health professionals regarding violations of the right to health, including the use of chemical weapons, in opposition-held Syria. *BMJ Open* 2018;8:e021096.
31 Haar RJ, Risko CB, Singh S, *et al*. Determining the scope of attacks on health in four governorates of Syria in 2016: results of a field surveillance program. *PLoS Med* 2018;15:e1002559.
32 Human Rights Watch. Syria/Russia: strategy targeted civilian infrastructure, 2020. Available: https://www.hrw.org/news/2020/10/15/syria/russia-strategy-targeted-civilian-infrastructure
33 Human Rights Center, University of California. Berkeley protocol on digital open source investigations, 2020. Available: https://humanrights.berkeley.edu/programs-projects/tech-human-rights-program/berkeley-protocol-digital-open-source-investigations
34 Raghavan S, Booth W, Eglash R. The U.N. is accusing Israel of breaking international law with an attack on a school in Gaza. Washington Post, 2014. Available: https://www.washingtonpost.com/world/israel-presses-attack-16-killed-at-un-school/2014/07/30/4a643588-17a5-11e4-85b6-c1451e622637_story.html
35 Briody C, Rubenstein L, Roberts L, *et al*. Review of attacks on health care facilities in six conflicts of the past three decades. *Confl Health* 2018;12:1–7.
36 Haar RJ, Footer KH, Singh S, *et al*. Measurement of attacks and interferences with health care in conflict: validation of an incident reporting tool for attacks on and interferences with health care in eastern Burma. *Confl Health* 2014;8:1–12.
37 Safeguarding Health in Conflict. Health workers at risk: violence against healthcare, 2020. Available: https://www.safeguardinghealth.org/sites/shcc/files/SHCC2020final.pdf
38 Chukwuma A, Ekhator-Mobayode UE. Armed conflict and maternal health care utilization: evidence from the Boko Haram Insurgency in Nigeria. *Soc Sci Med* 2019;226:104–12.
39 Footer KHA, Meyer S, Sherman SG, *et al*. On the frontline of eastern Burma's chronic conflict--listening to the voices of local health workers. *Soc Sci Med* 2014;120:378–86.
40 Hemat H, Shah S, Isaakidis P, *et al*. Before the bombing: high burden of traumatic injuries in Kunduz trauma center, Kunduz, Afghanistan. *PLoS One* 2017;12:e0165270.
41 Rytter MJH, Kjaeldgaard A-L, Brønnum-Hansen H, *et al*. Effects of armed conflict on access to emergency health care in Palestinian West bank: systematic collection of data in emergency departments. *BMJ* 2006;332:1122–4.
42 Trelles M, Stewart BT, Hemat H, *et al*. Averted health burden over 4 years at Médecins Sans Frontières (MSF) Trauma Centre in Kunduz, Afghanistan, prior to its closure in 2015. *Surgery* 2016;160:1414–21.
43 Chi PC, Bulage P, Urdal H, *et al*. Perceptions of the effects of armed conflict on maternal and reproductive health services and outcomes in Burundi and Northern Uganda: a qualitative study. *BMC Int Health Hum Rights* 2015;15:7.
44 Amnesty International. Death everywhere: war crimes and human rights abuses in Aleppo, Syria, 2015. Available: http://www.amnestyusa.org/sites/default/files/embargoed_5_may_aleppo_report_death_everywhere.pdf
45 Amnesty International. War of annihilation. devastating Toll on civilians, Raqqa, Syria, 2018. Available: https://www.amnesty.org/download/Documents/MDE2483672018ENGLISH.PDF
46 Çelikel A, Karbeyaz K, Kararslan B, *et al*. Childhood casualties during civil war: Syrian experience. *J Forensic Leg Med* 2015;34:1–4.
47 Fardousi N, Douedari Y, Howard N. Healthcare under siege: a qualitative study of health-worker responses to targeting and besiegement in Syria. *BMJ Open* 2019;9:e029651.
48 Guha-Sapir D, Rodriguez-Llanes JM, Hicks MH, *et al*. Civilian deaths from weapons used in the Syrian conflict. *BMJ* 2015;351:h4736.
49 Kimball D, Davenport K. Timeline of Syrian chemical weapons activity, 2012-2021, 2020. Arms control association. Available: https://www.armscontrol.org/factsheets/Timeline-of-Syrian-Chemical-Weapons-Activity
50 Torture in Syria's hospitals. *The Lancet* 2011;378:1606.
51 Rodriguez-Llanes JM, Guha-Sapir D, Schlüter B-S, *et al*. Epidemiological findings of major chemical attacks in the Syrian war are consistent with civilian targeting: a short report. *Confl Health* 2018;12:1–6.
52 Wong CH, Chen CY-T. Ambulances under siege in Syria. *BMJ Glob Health* 2018;3:e001003.
53 Amnesty International. Syrian and Russian forces targeting hospitals as a strategy of war, 2016. Available: https://www.amnesty.org/en/press-releases/2016/03/syrian-and-russian-forces-targeting-hospitals-as-a-strategy-of-war/
54 Hill E, Triebert C. 12 hours. 4 Syrian hospitals bombed. one culprit: Russia. the new York times, 2019. Available: https://www.nytimes.com/2019/10/13/world/middleeast/russia-bombing-syrian-hospitals.html

55 Airwars. Airwars: our monitoring of civilian harm, 2021. Available: https://airwars.org/
56 R Core Team. R: the R project for statistical computing, 2021. Available: https://www.r-project.org/
57 Greenacre M. Tying up the loose ends in simple, multiple, joint correspondence analysis. In: Rizzi A, Vichi M, eds. *Compstat 2006 - Proceedings in Computational Statistics*. Heidelberg: Physica-Verlag HD, 2006: 163–85.
58 Lê S, Josse J, Husson F. FactoMine*R* : An *R* Package for Multivariate Analysis. *J Stat Softw* 2008;25.
59 Tenenhaus M, Young FW. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika* 1985;50:91–119.
60 Husson F, Josse J. Multiple correspondence analysis. In: Blasius J, Greenacre M, eds. *Visualization and Verbalization of data*. 1st. New York: Chapman and Hall/HRC, 2014. ISBN: 9780429167980. https://www.taylorfrancis.com/books/edit/10.1201/b16741/visualization-verbalization-data-jorg-blasius-michael-greenacre?refId=df14d590-f242-4d82-aac1-6b487d8c86f3
61 Abdi H, Valentin D. Multiple Correspondence Analysis. In: Salkind N, ed. *Encyclopedia of measurement and statistics*. Thousand Oaks, California: Sage, 2007.
62 Ambrogi F, Biganzoli E, Boracchi P. Multiple correspondence analysis in S-PLUS. *Comput Methods Programs Biomed* 2005;79:161–7.
63 Rodriguez-Sabate C, Morales I, Sanchez A, *et al*. The multiple correspondence analysis method and brain functional connectivity: its application to the study of the non-linear relationships of motor cortex and basal ganglia. *Front Neurosci* 2017;11:345 https://www.frontiersin.org/articles/10.3389/fnins.2017.00345/full
64 Kassambara A, Mundt F. Extract and Visualize the Results of Multivariate Data Analyses [R package factoextra version 1.0.7], 2020. Comprehensive R Archive Network (CRAN). Available: https://CRAN.R-project.org/package=factoextra
65 Wickham H. *ggplot2: elegant graphics for data analysis*. New York: Springer-Verlag, 2009. https://www.springer.com/gp/book/9780387981413
66 Wei T. Visualization of a Correlation Matrix V0.90 R Package "corrplot", 2021. Available: https://github.com/taiyun/corrplot
67 Yuan C, Yang H. Research on k-Value selection method of k-means clustering algorithm. *J* 2019;2:226–35.
68 Manning C, Raghavan P, Schutze H. Evaluation of clustering. In: *Introduction to Information Retreival*. Cambridge University Press, 2008. https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html
69 Hastie T, Qian J, Tay K. An Introduction to 'glmnet', 2021. Available: https://glmnet.stanford.edu/articles/glmnet.html
70 Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am Stat* 1983;37:36–48.
71 Siriseriwan W. A collection of Oversampling techniques for class imbalance problem based on SMOTE, 2019. Available: https://cran.r-project.org/web/packages/smotefamily/smotefamily.pdf
72 Topçuoğlu BD, Lapp Z, Sovacool KL, *et al*. mikropml: user-friendly R package for supervised machine learning pipelines. *J Open Source Softw* 2021;6:3073.
73 Sheskin DJ. *Handbook of Parametric and Nonparametric Statistical Procedures*. 2nd edition. Boca Raton: Chapman and Hall/CRC, 2000: 1016p.
74 Signorell A. Tools for Descriptive Statistics [R package DescTools version 0.99.41] . Comprehensive R Archive Network (CRAN), 2021. Available: https://CRAN.R-project.org/package=DescTools
75 Syrian Archive. Russian airstrikes database, 2021. Available: https://syrianarchive.org/en/datasets/russian-airstrikes
76 Triebert C, Hill E, Browne M. How times reporters proved Russia bombed Syrian hospitals, 2019The New York Times. Available: https://www.nytimes.com/2019/10/13/reader-center/russia-syria-hospitals-investigation.html
77 HRW. Syrian and Russian strikes on civilian infrastructure. Available: https://www.hrw.org/report/2020/10/15/targeting-life-idlib/syrian-and-russian-strikes-civilian-infrastructure [Accessed 12 Aug 2021].
78 Case Study [Internet]. Physicians for human rights. Available: https://phr.org/issues/health-under-attack/attacks-in-syria/case-study-number-3/
79 Human Rights Watch. Targeting life in Idlib, 2020. Available: https://www.hrw.org/report/2020/10/15/targeting-life-idlib/syrian-and-russian-strikes-civilian-infrastructure
80 Khan A, Gopal A. The uncounted, 2017The New York Times Magazine. Available: https://www.nytimes.com/interactive/2017/11/16/magazine/uncounted-civilian-casualties-iraq-airstrikes.html
81 Roth A, Murphy B, Ryan M. Russia begins airstrikes in Syria; U.S. warns of new concerns in conflict, 2015. Washington Post. Available: https://www.washingtonpost.com/world/russias-legislature-authorizes-putin-to-use-military-force-in-syria/2015/09/30/f069f752-6749-11e5-9ef3-fde182507eac_story.html
82 Domonoski C. As East Aleppo falls, accounts of carnage — then a cease-fire, 2016. Available: https://www.npr.org/sections/thetwo-way/2016/12/13/505378173/as-east-aleppo-falls-desperate-accounts-of-carnage-and-civilian-deaths
83 Rogin J. Opinion | The catastrophe in Syria is getting worse. It's time to act, 2020. Washington Post. Available: https://www.washingtonpost.com/opinions/global-opinions/why-americans-should-care-about-syria/2020/03/05/550e88bc-5f2a-11ea-b29b-9db42f7803a7_story.html
84 Syrian American Medical Society. Civilians, civilian infrastructure in northwest Syria face heightened attacks. ten schools, one hospital attacked today, 2020. Available: https://www.sams-usa.net/press_release/civilians-civilian-infrastructure-in-northwest-syria-face-heightened-attacks-ten-schools-one-hospital-attacked-today/
85 Security Council. Deadly attacks on civilians in Syria's Idlib region 'Happening in Plain Sight, Under Our Watch', Under-Secretary-General Tells Security Council | Meetings Coverage and Press Releases. UN Meetings and Press Coverage. SC 14132(8738th meeting), 2020. Available: https://www.un.org/press/en/2020/sc14132.doc.htm
86 Allaire JJ. *Deep learning with R*. 1st edition. Manning Publications, 2018: 360 p.
87 van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol* 2007;6:Article25.