# Challenges in using open data for information security research: An investigation in the healthcare sector

*Full-Paper Submission*


## Abstract

Several research studies have used open data to investigate information security breaches. However, these studies have used this data "as-is" without consideration of data quality issues. It is necessary to assess the quality of open data by understanding the context of its creation and its limitations. We illustrate how a data product can be created from an open dataset on security breaches in healthcare organizations provided by the Department of Health and Human Services. By focusing on the meaning of a security breach event, we applied a rigorous process to identify invalid and duplicate cases so that the final data product could be a resource for researchers and practitioners in information security.


## Introduction

Healthcare and related services organizations, a significant part of the US economy, maintain sensitive health information of the individuals they serve. The Health Insurance Portability and Accountability Act (HIPAA) regulation aims to protect the privacy and security of this information from impermissible access, use or disclosure to entities within and outside these organizations. The Office for Civil Rights (OCR), part of the U.S. Department of Health and Human Services (HHS), is responsible for enforcing the Privacy, Security, and Breach Notification Rules of HIPAA. The organizations that are covered by the HIPAA regulation include Healthcare providers, Health Plans, and Healthcare clearinghouses (henceforth referred to as Covered Entities or CEs), as well as their business associates (henceforth referred to as BAs). When a security breach occurs, i.e., an incident that exposes the security or privacy of protected health information (PHI) of individuals, the HIPAA breach notification rule mandates that healthcare organizations notify the affected individuals, and the HHS[1]. OCR uses the information provided by the reporting organizations and makes data (henceforth referred to as OCR dataset) on breaches affecting more than 500 individuals publicly available on its portal as 'open data'. Open data is any data that is available "free of charge, without registration or restrictive licenses, for any purpose whatsoever (including commercial purposes), in electronic, machine-readable formats that are easy to find, download and use[2]". The availability of such open data on a portal has several benefits including transparency, accountability (Chui et al., 2014) as well as mobility and interoperability (Leonelli, 2020).

Studies on security breaches in various industries have often relied on repositories of data such as Privacy Rights Clearinghouse (e.g., Jeong et al., 2019). Particularly, several studies investigating security breaches in the healthcare industry have used the OCR dataset given its easy availability in a machine-readable format. Among these, some have focused on descriptive analysis (Wikina, 2014; Raghupathi et al., 2023). Other studies have combined this data with other datasets to understand causal relationships that can either explain or predict security breaches or their impact such as the number of

[1] https://www.hhs.gov/hipaa/for-professionals/breach-notification/index.html

[2] *Open data challenges and opportunities for national statistical offices*. Washington, D.C. : World Bank Group. http://documents.worldbank.org/curated/en/740381468128389452/Open-data-challenges-and-opportunities-for-national-statistical-offices

individuals affected by a breach. McLeod and Dolezel (2018) combined the OCR dataset with the Healthcare Information and Management Systems Society (HIMSS) database to predict the likelihood of a breach based on the level of exposure, level of security, and organizational factors. Dolezel and McLeod (2019) analyzed the OCR dataset to predict the number of individuals affected based on the type of the breach, the location of the breach, the type of covered entity, the presence of a business associate, and the geographic region of the breach. Gabriel et al. (2018) also linked the OCR dataset with Health Information Management Systems Society (HIMSS) and the American Hospital Association Health IT Supplement databases to predict the occurrence of breach based on hospital characteristics. Ignatovski (2022) investigated the relationship of the type of healthcare entity and the number of individuals impacted by the security breaches occurring during the COVID-19 pandemic.

Open government data has been found to have data quality issues (Vetrò et al., 2016). Such data quality issues may also be found in any data associated with security breaches (e.g., Grispos,, 2016). Epistemologically, it would be important to consider the data quality issues as any knowledge claims made would depend on the quality of evidence available. Scholars have attempted to address the data quality issues associated with available data to create curated datasets (e.g., Gentry et al., 2021, Voermans and Lelli, 2024). However, the studies that have used the OCR dataset have used it "as-is" without critically questioning whether there were any limitations of this data or without paying attention to the context in which the data was generated. Studies may arrive at erroneous conclusions if the available open data either overreports or underreports the occurrence of security breaches. Information security researchers often attempt to understand the antecedents, processes, and consequences of security breaches. For researchers focused on the healthcare context who use open data such as the OCR dataset, it would be important to ascertain 1) whether a security breach incident has occurred or not, 2) whether the entity reporting the incident is covered by HIPAA, i.e., the breach is a healthcare security breach, and 3) whether it is a single incident or a set of incidents.

We conceptualize security breaches occurring in healthcare organizations as 'events' i.e. "the point in space and time where entities or entity actions contact, encounter or meet each other", where events are "external to the perceiver" and "bounded in time and space" and are novel, critical and disruptive to the organizations (Morgeson et al., 2014). To understand the phenomenon of security breaches, we believe it is necessary to first understand what a 'security breach event' means. This relates to the ontology of the phenomenon of interest. Therefore, this research focuses on the meaning of an 'event' within the overall context of how the OCR dataset has been constructed.

We identify the data quality issues associated with the OCR dataset and describe some of the necessary preprocessing work to create a curated 'data product' (Arribas-Bel et al., 2021) that could become the basis for any meaningful analysis conducted by researchers. By focusing on the limitations of the OCR dataset and the steps needed to address some of these limitations, this research aims at helping information security researchers who intend to use public data such as the OCR dataset make better use of such data resources.

## Background literature

The growth of data in recent decades has given rise to notions of 'data-centric' science or 'open science' where data handling and dissemination practices have taken prominence (Leonelli, 2016). 'Datafication' (Cukier and Mayer-Schoenberger, 2013) represents processes by which aspects of human and social life are transformed into digital data that can be analyzed further (Mejias and Couldry, 2019) and this provides opportunities for researchers to learn about a phenomenon of interest. Wing (2019) defines data science as 'the study of extracting value from data' and presents a data lifecycle that includes the phases of generation, collection, processing, storage, management, analysis, visualization, and interpretation. In this traditional view, the role of the human actor is acknowledged to be relevant only towards the end. Further, this framework largely assumes the predominance of a linear process in extracting value from data as if data is ready for analysis once it is generated, collected and stored while interpretation occurs only in the last phase. Any feedback to previous phases is quite delayed until the interpretation phase.

On the other hand, human judgment is involved throughout the data lifecycle and decisions are made about what data is collected, how it is processed, whether and where it is stored, how it is analyzed and what visualizations are used to communicate the insights (e.g, Muller et al, 2019; Lin et al, 2022). Grolemund and Wickham (2014) present data analysis as a sensemaking process applicable to both

exploratory and confirmatory analyses. During exploratory analysis, such as the current study, the task is to find a suitable schema that matches the observed data. Tanweer et al. (2021) argue that a qualitative/interpretive approach complements computational/quantitative data analysis and can enhance the reliability and make research more comprehensive.

Data are not just found or 'given' (Kitchin, 2021) but are 'constructed', 'cooked' or 'made' (Gitelman, 2013; Leonelli, 2015; Vis, 2013)). For example, Donatz-Fest(2024) uses ethnographic fieldwork to study the data work of Netherlands Police and focused on how police employees constructed both structured and unstructured data in the form of reports that later were used elsewhere. When filling out the report, Donatz-Fest (2024) observed that while employees were presented with predefined categories, they had some discretion in terms of what codes were selected while these codes also limited their discretion to a certain extent. In reporting a given situation, the employees used their judgment and prior knowledge in identifying what code needed to be used in reporting an incident. The police employees also entered unstructured data by taking notes on mobile devices. Depending on whether the situation was an emergency, the note-taking was influenced by that and data was omitted in certain circumstances (Donatz-Fest, 2024). The ethnographic study by Donatz-Fest (2024) shows the 'datafication' of 'street-level cognitions'. In their study of bug fixing, Ekbia (2009) reported how it took significant time for software development teams to arrive at a conclusion whether a reported issue was a "bug". Thus, what was reported eventually as a bug underwent a process of interpretation. On the other hand, sometimes this process of classification or labeling may be quite rushed or non-existent and what is treated as data may have to be reviewed and revised or 'cleaned' before any analysis can be performed.

Leonelli (2015, 2020) views data as 'portable objects' and argues that in the context of scientific endeavors, data is collected, stored, and disseminated in order to be used as 'evidence for knowledge claims' about a phenomenon. Leonelli (2020) also conceptualizes data as 'lineages' - as objects that need to be transformed or "mutated" in order to travel and fit different uses or its environment and also be subject to scrutiny other than the individuals involved in its creation. Goodman et al. (2014) suggest that when data has provenance and users have access to both metadata and data about the processes that generated the data (i.e. paradata), there is greater chance for data to be reused. Metadata enables sensemaking of the data *content* and paradata is about data on *processes* that generate data (Huvila et al., 2025).

Open data that is produced must adhere to the principles of data quality in order for it to be 'portable' and reusable. The data quality framework is quite relevant to the scrutiny of data, particularly open government data in the context of information security research. The UNECE (2014) data quality framework presents three higher dimensions that account for the source (institutional factors), data (quality of the data itself), and metadata including paradata. While we did not have the benefit of directly observing the processes associated with the creation of OCR dataset or access to its paradata, our investigation and analysis reported here was able to uncover several challenges and limitations of the data which should be considered while engaging in any subsequent analysis.

## Materials and Methods

### *Data sources*

Our study uses the breach report data that the OCR makes available to the public and supplements it with data from media reports that we could locate. The OCR dataset has a description of security incidents. To clarify some of the procedures followed by the OCR, we also conducted a phone interview with an OCR official. The official confirmed that this description is based on the initial details provided by the healthcare organization submitting the breach which are revised by the OCR after their investigation of the incident is completed. HIPAA's breach notification rule requires healthcare organizations to report a breach to the HHS via an online breach report form within 60 days after the event (if 500 or more individuals were affected) or after the end of the calendar year in which the event occurred (if fewer than 500 individuals are affected) (Breach Notification Rule). In addition, they must report the breach to the affected individuals (in all cases) and to the media (if 500 or more individuals are affected). The covered entities or their business associates can report the breach by filling out a form. The form requires them to provide several pieces of information such as an estimate of the approximate number of individuals affected, the type and location of the breach, the type of protected health information (PHI) that was

affected, the safeguards that were present before the breach, and the actions taken in response to the breach.

After a security breach is reported, OCR begins its investigation to identify any potential violations of Privacy and Security rules. Based on our interview with the OCR official, we found that OCR follows essentially the same enforcement process in response to a breach notification as for complaints that they receive from affected stakeholders. The investigation may identify specific violations which OCR reports in an updated version in the description field of the breach report. OCR may require a corrective action plan (e.g., see Virtual Private Network Solutions, LLC) or levy penalty (e.g., see New England Dermatology), or perform education / outreach to enforce the HIPAA Security Rule.

The OCR dataset only includes security breaches reported from the last quarter of 2009 that affected at least 500 individuals. We downloaded the dataset in July 2023. The OCR may sometimes take as long as two years to complete its investigation and resolve a case. Most newly reported breaches in the OCR dataset have missing values in the description field which can provide more information and context for the breach. Therefore, we first filtered the data to only include the years 2010 - 2019 to ensure that OCR had likely resolved the cases and would have updated the description field which we have relied upon heavily for further processing of the data.

## *Analysis and findings*

We combined both top-down, 'theory-laden' and bottom-up 'data-driven' perspectives in order to identify patterns and specific errors in the data. In order to identify issues with data quality, we had to pay close attention to the data itself, similar to the 'close reading' approach in qualitative research. Whether the pattern found in data matched the expectations required an awareness of the expected pattern, i.e., a 'theory' about data. Adapting the UNECE (2014) data quality framework, Kitchin and Stehle (2021) present several dimensions to assess data quality. Among these we found the following to be relevant to the OCR dataset: *fidelity, cleanliness, completeness, spatial granularity, temporal granularity, metadata, changes through time,* and *methodological transparency* (paradata).

*Data Input*

The *fidelity* dimension of data quality (Kitchin and Stehle, 2021) is about accuracy, precision and bias.

The accuracy of data is affected by misclassification errors. Since the reporting organization has to rely on the pre-existing categories available in the form such as the type of breach that may have occurred or whether the reporting entity is a covered entity, it could lead to misclassification of an event due to variations in interpretations. The OCR may also revise the data eventually when there is misclassification. Our analysis of the OCR dataset indicated that several organizations reported hybrid breaches that involved more than one type of breach or indicated 'Other' as the type of breach before 2015. The multiple categories and 'Other' as a category disappears from the OCR dataset after 2015. Based on this we could infer that reporting organizations were allowed to select multiple categories or select 'Other' as an option before 2015. We also found inconsistencies in reporting the covered entity type (e.g., Healthcare Organization, Health Plan, Business Associate). Out of around 2245 covered entities/business associates who reported at least one security breach between 2010 - 2019, 228 had reported at least two breaches and when reporting these breaches, 37 had used inconsistent entity type classification when reporting a breach. The covered entity type refers to the organization suffering the breach. In one reported breach, an employee of a covered entity (Health Plan) was involved in impermissible disclosure of PHI to its business associate. However, the covered entity had classified itself as a Business Associate due to misinterpretation of the Covered Entity Type field of the breach submission form.These pertain to the accuracy aspects of data quality.

Researchers using the OCR dataset should also keep in mind the precision of the data. The reporting organizations provide only an estimate of the number of individuals affected by the breach. This estimate is sometimes revised later. Among the data we analyzed, in around 22% of the cases the word "approximately" was used in the description field to qualify the number of individuals affected by a breach. In around 68% of the cases, the number of individuals affected was also reported in the description provided by OCR thereby corroborating this value.

*Data organization*

After a breach is reported, OCR does update the data after its investigation. The initial description is empty and is updated later on, and individuals affected by the breach are also occasionally updated. While OCR introduced a breach tracking number from January 1, 2015 onwards, reporting organizations updating the breach need to remember to include this in their breach submission. Since a unique identifier was not associated with each reported breach in the OCR dataset, it leads to significant challenges in keeping track of any changes that may occur in the data as it is updated or corrected by OCR. Having a unique identifier associated with each reported breach can make it easier for researchers to compare the downloaded OCR dataset with any changes made later.

*Data Attributes*

Both *spatial* and *temporal granularity* of data would improve if the address of the reporting organization was provided or the estimated date of the security breach was provided. Currently, the data only includes the state where a reporting organization is located and indicates the date when the organization reported the breach to OCR but not when the breach may have occurred. While the breach submission form itself includes fields for reporting the date when a breach occurred (if available), the OCR dataset does not make it available.

Both *metadata* and *paradata* associated with the dataset could help researchers in contextualization when they want to use this data.

*What is a security event?*

According to OCR, the OCR dataset is a list of security breaches reported by organizations. Adopting the cognitive schema of "tidy data" (Wickham, 2014), we started with the initial assumption that each row in the dataset is an independent, mutually exclusive security breach. However, close reading of the column that provides a description by OCR about the incident indicated that sometimes OCR consolidated multiple events as a single incident. Our search for invalid and duplicate records was triggered when we noticed some case descriptions that explicitly mentioned that the incident was not a security breach affecting a HIPAA regulated entity or that it was a duplicate. This indicated the possibility that the data in the dataset could not be assumed to contain records reporting distinct valid incidents.

We were able to identify several scenarios upon close reading of the incident descriptions: 1) A security breach at a business associate affecting multiple CEs; 2) An organization or health system affected by multiple security breaches for which it may be sanctioned by OCR; 3) a healthcare system may have multiple administrative units and a breach may span these units. 4) An organization may report the incident multiple times when there is new information available that they would want to report as well 5) A reported incident was misinterpreted by the organization where OCR investigation subsequently determined that it was not fitting the criterion of a security breach; 6) An incident reported by an organization that is not regulated by HIPAA. The following illustrates Scenario 3 where the OCR portal shows that there were 23 "incidents" reported for security breaches on January 9, 2025 which appear to belong to a single organization based on the name of the reporting organization. Further inspection using media reports indicated that each of these incidents indeed may belong to a single organization named HCF, though the locations of the reporting entities were across the states of Ohio and Pennsylvania. Without having this additional information, one could erroneously conclude that these were 23 different 'events'. Further processing of this data should combine these incidents into a single event while indicating the total number of individuals affected by this event and the fact that these breaches spanned two states. It is possible that OCR may eventually combine these incidents as a single event.

These scenarios identify the cases that should be excluded as part of the data cleaning process: 1) invalid cases, and 2) duplicate cases. Invalid cases are those that involve either 1) a reported security breach that does not involve exposure of PHI, or 2) an organization that is not regulated by HIPAA. However, organizations sometimes report incidents even if one or both of these criteria are applicable to them. In these cases we relied on the determination made by the OCR to identify and exclude such cases from our study. Duplicates are multiple records that describe a single security event. We define a single security event as one that involves a specific threat affecting a specific organization. Duplicate records are created in the OCR dataset when 1) a business associate is affected by an incident, and the business associate and the covered entity served by that business associate both report the same incident, 2) a business associate is affected by an incident, the business associate serves multiple covered entities, and more than one of this group of organizations report the same incident, 3) a given organization experiences different

incidents over a time period, each with a separate record in the dataset, and the OCR creates a new record summarizing some or all of those incidents, 4) a given organization submits more than one report of the same incident (e.g., one soon after the incident and another a few weeks later with updated details).

*Data cleaning*

We used both automated and manual processes for data cleaning. Once a pattern could be found by close reading of data, we tried to incorporate it into a script to identify all cases that matched the pattern while not resulting in inclusion of cases that do not belong.

Our search for the invalid and duplicate records was conducted in the following phases:

Phase 1: Identify invalid cases

To systematically identify invalid and duplicate records we used an R script (R Core Team, 2024) to exclude records based on the information provided by OCR in the "web description" field of the OCR dataset. We used the following phrases to identify such records (see Table 1):

**Table 1: Words / phrases to identify records to be excluded**

| Reason for exclusion | Words / phrases |
|---|---|
| Organization was not a regulated healthcare organization | "no longer a covered entity"; "no longer a CE"; "no longer a business associate"; "not a covered entity"; "not covered by HIPAA" |
| Not a security breach of PHI | "no breach occurred in this case"; "not a breach" |
| Cases that are consolidated by OCR into other cases to avoid duplication | "has been consolidated"; OCR has consolidated"; "was consolidated"; "and consolidated"; "duplicate case" |
| Cases closed administratively by OCR | "administratively closed"; "is closed" |

Phase 2: Initial identification of duplicates

Since there can be variations in the recording of covered entity names, a consistent format was applied by converting the names to uppercase, removing any commas or periods, and then removing any text strings that were not part of the entity name (e.g., "Incorporated", "Inc", "LLC", "PLLC", "PC", "LTD", "LLP"). Cases where the name of the covered entity and the state was the same were identified as 'possible' duplicates. Where the number of affected individuals was also the same for records with the same name of covered entity and state, we marked such records as 'likely' duplicates.

We then closely read the descriptions of such cases to verify if the records in each such record set were indeed duplicates. We considered the following factors in determining whether records in a set were duplicates: the date of report submission, the date of incident if available, and the type of incident. One of the authors reviewed each of these sets of records, marked records for exclusion or retention based on a review of each set of records, and flagged doubtful cases. Both authors then jointly reviewed these sets of doubtful cases. In our individual and joint reviews, we considered a particular set of records as duplicates only if there was compelling evidence either in the description field in the OCR dataset itself or in external sources of information such as media reports or information published on the website of the affected organization. For example, two records about "Virginia Mason Medical Center" were deemed to reflect two separate events since they were separated in time and involved different types of events. In another example, two records for "American Medical Response" in Texas were determined to be duplicates as their respective descriptions noted that "this report was a duplicate of a previous breach notification to OCR".

During this review of records, we observed cases of multiple covered entities being affected by a breach at a business associate who served these organizations. We also suspected that multiple organizations reporting the same type of breach (e.g., Hacking/IT Incident, or Theft) having the same location of breach information (e.g., Server or Email), and on the same day may be affected by the same breach. This prompted us to further search for any duplicate records based on these patterns.

Phase 3 (Identification of duplicate records based on breach submission date)

Surmising that reports about the same incident from different affected organizations (e.g., covered entities affected by an incident at their business associate) could possibly have the same dates of breach submission or dates that are quite close to each other, we sorted the records by breach submission date and reviewed their descriptions along with other fields. We selected a stratified sample of records across the dataset at different time periods based on the submission date of reported security breaches: 1) 110 records during January - July 2010; 2) 122 records during June - October 2014; 3) 101 records during November 2017 - February 2018; 4) 110 records during April - June 2020; 5) 100 records during February - May 2022. Thus, we examined 543 records (12.86%) out of 4223 total records (spanning years 2010 - 2022). One of the authors reviewed each of these records and flagged possible duplicate records. Then we jointly reviewed such records and relied on media reports to determine which records were duplicates. Out of these 543 records, we found 11 records (2.03%) that were duplicates.

Phase 4 (Identification of duplicate records based on number of affected individuals)

Next, we considered the possibility that duplicate records could have the same number of individuals affected by a breach. So we sorted the records by the number of affected individuals. One of the authors reviewed each set of two or more records by reading the description field of those records to assess whether they were duplicates and flagged those sets that were doubtful. We jointly reviewed these doubtful cases and also used media reports to mark records that were duplicates. Some of the factors that we used to decide were whether the organizations were based in the same state, how similar were the descriptions, and how close were the reporting dates or the incident dates for the breach in each of the records. As part of this review, we found several cases of a security breach at one business associate that affected one or more covered entities, and where two or more of the parties reported the same incident.

Phase 5 (Identification of duplicate records based on business associate - covered entity relationship)

Since we found a few duplicate records based on an incident at a business associate affecting several organizations, we reviewed the entire OCR dataset for records where a business associate was mentioned in the description field. In records where we found mention of a business associate, we searched for the name of the business associate in the entire OCR dataset. As part of this search, we found several cases where the same business associate was mentioned in more than one record, usually in the description field. There were also records in which one organization was doing business as entities with different names. As in the previous phases, one of the authors flagged cases for removal where it was clear from the description that they were duplicates of another record.

Phase 6 (Identification of duplicate records based on description of events)

Finally, we filtered records in which two or more records had the exactly the same description. We reviewed all such records jointly to determine whether they were duplicates based on factors such as reporting dates of incidents, state where the organization was based, and media reports.

*Other limitations of the dataset*

There are some other limitations of the OCR dataset. First, In many cases the business associate of a covered entity was not identified and so it was not possible to ascertain if a business associate also had reported the same breach along with the covered entity. Second, without gathering more information about the organization reporting a security breach, it was not possible to identify if multiple administrative units of a health system had reported the same breach. Sometimes there are mergers of healthcare organizations: Advocate Health Care and Aurora Health Care merged to create Advocate Aurora Health in 2017 becoming the 10th largest health care system in the United States at that time. In 2022 Advocate Aurora Health after merger with Atrium was renamed Advocate Health, becoming the fifth largest healthcare system in the United States. Thus, any security breaches reported by an organization that is part of the merger should be attributed to the merged organization after the merger event. Third, we also found some records where a covered entity may report a possible breach while stating that it may not be a breach at all and OCR had not documented its determination. This leads to ambiguity about whether the reported breach was indeed a breach or not.

The *completeness* dimension of data quality (Kitchin and Stehle, 2021) refers to attribute coverage and whether the data is exhaustive. It should be recognized that the OCR data covers only those incidents that

reported that 500 or more were affected in each incident. The Annual Compliance Report to Congress by the HHS Department shows that such incidents comprise less than 1% of all security breaches in healthcare organizations. Hence, any findings from the data in this paper (and those reported in prior published work) only reflect large-scale breaches as compared to small-scale breaches. In our interview with the OCR official, we were informed that data on breaches that affected fewer than 500 individuals is not available to the public.

*Finished dataset*

After cleaning the OCR data we were able to create a 'data product' (Arribas-Bel et al., 2021) which we could use for further analysis in our own projects as well as share with the information security research community. Following open science practices, we intend to provide data, metadata and paradata as a package so that users of this cleaned OCR dataset can recontextualize and reuse this data.

# Discussion

The easy availability of open data can be an asset for researchers. However, if data are to serve as a 'research object' (Aaltonen, et al., 2023) or as 'evidence for claims made' (Leonelli, 2015), the role of contextual factors must also be considered such as how the data has been generated, what are the data quality issues associated with the data including completeness, cleanliness, fidelity, and granularity. Our study identifies several data quality issues that relate to the input of data itself. As the study by Donatz-Fest (2024) indicates, the design of the breach submission form itself may impose a particular cognitive schema on those submitting the form which may also give rise to certain misclassifications in reporting of the breach. In the context of public health, the gradual implementation of the pregnancy checkbox in 2003 to ascertain the causes of maternal death led to overreporting of maternal deaths due to misclassification errors (Joseph et al., 2024). Our study also found several misclassification errors.

The 'data product' that we have created based on the source OCR dataset has the benefit of having information on how the data was processed (as reported in this paper here) and when used along with the code and notes that can help other researchers understand how the data was 'constructed' and whether this data could be used in their own context or any further work may be needed.

We have several recommendations both for data providers such as the OCR and researchers who intend to use open data. Data providers could provide both metadata and paradata along with data which can help researchers in contextualizing data and use data as evidence appropriately. While the OCR has introduced a tracking number for security breaches reported after January 1, 2015, this number is not included in the OCR dataset. Following the best practices of database design, if the OCR can create unique identifiers for organizations reporting the breach and for the security breaches reported, it will make the task of keeping track of security breaches easier, particularly when there are changes to data such as mergers of organizations or updates in the data about the breach.

Researchers who use open data such as the OCR dataset should avoid using the data "as-is". Researchers would benefit from using supplementary resources such as media reports and websites of the organizations reporting the breach and by conducting interviews with organizations that suffered a breach.

# Conclusion

Information systems security researchers and practitioners rely on data to study past incidents and develop solutions to minimize their occurrence in the future. In comparison with other sectors, data on security breaches among healthcare organizations is more easily available to the public due to HIPAA regulations. However, such ease of accessibility carries with it the burden of cleaning the data based on a deep understanding of its context. This paper presents an initial exploration of this data by focusing on the meaning of a security "event" and creates a new data product for use in future research by all scholars.

# References

Aaltonen, A., Alaimo, C., Parmiggiani, E., Stelmaszak, M., Jarvenpaa, S. L., Kallinikos, J., & Monteiro, E. (2023). What is Missing from Research on Data in Information Systems? Insights from the Inaugural

Workshop on Data Research. Communications of the Association for Information Systems, 53, 475-490. https://doi.org/10.17705/1CAIS.05320

Arribas-Bel, D., Green, M., Rowe, F., & Singleton, A. (2021). Open data products-A framework for creating valuable analysis ready data. *Journal of geographical systems*, *23*(4), 497–514. https://doi.org/10.1007/s10109-021-00363-5

Chui, M., Farrell, D. & Jackson, K. (2014) How Government can Promote Open Data and Help Unleash Over $3 Trillion in Economic Value. *McKinsey & Co.*, April 2014.

Cukier, K. & Mayer-Schoenberger, V. (2013) The Rise of Big Data: How It's Changing the Way We Think about the World. *Foreign Affairs*, 92, 28-41.

Dolezel, D., & McLeod, A. (2019). "Cyber-Analytics: Identifying Discriminants of Data Breaches," *Perspectives in Health Information Management*, 16(Summer), 1a.

Donatz-Fest, I. (2024). The 'doings' behind data: An ethnography of police data construction. *Big Data & Society*, 11(3). https://doi.org/10.1177/20539517241270695

Ekbia, H. R. (2009). Digital artifacts as quasi-objects: Qualification, mediation, and materiality. *Journal of the American Society for Information Science and Technology*, 60(12), 2554–2566. doi: http://doi.org/10.1002/asi.21189

Gabriel, M. H., Noblin, A., Rutherford, A., Walden, A., & Cortelyou-Ward, K. (2018). Data breach locations, types, and associated characteristics among US hospitals. *The American journal of managed care*, *24*(2), 78–84.

Gentry, R. J., Harrison, J. S., Quigley, T. J., & Boivie, S. (2021). A database of CEO turnover and dismissal in S&P firms, 2000-2018. *Strategic Management Journal* (42), 968-991.

Gitelman L. (ed) (2013) '*Raw Data' Is an Oxymoron*. Cambridge, Massachusetts ; London, England: The MIT Press.

Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., Di Stefano, R., Gil, Y., Groth, P., Hedstrom, M., Hogg, D. W., Kashyap, V., Mahabal, A., Siemiginowska, A., & Slavkovic, A. (2014). Ten simple rules for the care and feeding of scientific data. *PLoS computational biology*, *10*(4), e1003542. https://doi.org/10.1371/journal.pcbi.1003542

Grispos, G. (2016) *On the enhancement of data quality in security incident response investigations*. PhD thesis, University of Glasgow.

Grolemund, G., & Wickham, H. (2014). A cognitive interpretation of data analysis. *International Statistical Review, 82*(2), 184–204. https://doi.org/10.1111/insr.12028

Huvila, I., Andersson, L., Friberg, Z., Liu, Y.-H., & Sköld, O. (2025). *Paradata: Documenting Data Creation, Curation and Use*. Cambridge: Cambridge University Press.

Ignatovski M. (2022). Healthcare Breaches During COVID-19: The Effect of the Healthcare Entity Type on the Number of Impacted Individuals. *Perspectives in health information management*, *19*(4), 1c.

Jeong, C., Lee, S. & Lim, J. (2019). Information Security Breaches and IT Security Investments: Impacts on Competitors. *Information & Management*. 56(5), 681-695. https://doi.org/10.1016/j.im.2018.11.003.

Joseph, K. S., Lisonkova, S., Boutin, A., Muraca, G. M., Razaz, N., John, S., Sabr, Y., Chan, W. S., Mehrabadi, A., Brandt, J. S., Schisterman, E. F., & Ananth, C. V. (2024). Maternal mortality in the United States: Are the high and rising rates due to changes in obstetrical factors, maternal medical conditions, or maternal mortality surveillance? *American journal of obstetrics and gynecology*, *230*(4), 440.e1–440.e13. https://doi.org/10.1016/j.ajog.2023.12.038

Kitchin, R. (2021). *Data Lives: How Data Are Made and Shape Our World* (1st ed.). Bristol University Press. https://doi.org/10.2307/j.ctv1c9hmnq

Kitchin, R., & Stehle, S. (2021). Can Smart City Data be Used to Create New Official Statistics? *Journal of Official Statistics, 37*(1), 121-147. https://doi.org/10.2478/jos-2021-0006

Leonelli S. (2015). What Counts as Scientific Data? A Relational Framework. *Philosophy of science*, *82*(5), 810–821. https://doi.org/10.1086/684083

Leonelli, S. (2016). *Data-Centric Biology: A Philosophical Study*. London: University of Chicago Press.

Leonelli, S. (2020). Learning from Data Journeys. In: Leonelli, S., Tempini, N. (eds) *Data Journeys in the Sciences*. Springer, Cham. https://doi.org/10.1007/978-3-030-37177-7_1

Lin, H., Akbaba, D., Meyer, M. & Lex, A. (2023) Data hunches: Incorporating personal knowledge into visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 29(1), pp. 504–514. https://doi.org/10.1109/tvcg.2022.3209451

McLeod, A., & Dolezel, D. (2018). Cyber-analytics: Modeling factors associated with healthcare data breaches. *Decision Support Systems* (108), pp. 57-68.

Mejias, U. A. & Couldry, N. (2019). Datafication. Internet Policy Review, 8(4). https://doi.org/10.14763/2019.4.1428

Morgeson, F. P., Mitchell, T. R., & Liu, D. (2015). Event Systems Theory: An Event-oriented Approach to the Organizational Sciences. Academy of Management Review, 40 (4), 515-537. https://doi.org/10.5465/amr.2012.0099

Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q., Dugan, V. C. & Erickson, Thomas (2019). How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery*, New York, NY, USA, Paper 126, 1–15. https://doi.org/10.1145/3290605.3300356

Raghupathi, W., Raghupathi, V., & Saharia, A. (2023). Analyzing Health Data Breaches: A Visual Analytics Approach. *AppliedMath*, *3*(1), 175-199. https://doi.org/10.3390/appliedmath3010011

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Tanweer, A., Gade, E. K., Krafft, P. M., & Dreier, S. (2021). Why the Data Revolution Needs Qualitative Thinking. *Harvard Data Science Review*, *3*(3). https://doi.org/10.1162/99608f92.eee0b0da

UNECE (2014). A Suggested Framework for the Quality of Big Data. United Nations Economic Commission for Europe.

Vetro, A., Canova, L., Torchiano, M., Minotas, C., Iemma, R. & Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*. 33. https://doi.org/10.1016/j.giq.2016.02.001.

Vis, F. (2013). A critical reflection on Big Data: Considering APIs, researchers and tools as data makers. *First Monday*, *18*(10). https://doi.org/10.5210/fm.v18i10.4878

Voermans, N., & Lelli, F. (2024). A Dataset Containing S&P500 Information Security Breaches and Related Financial Firm Performances. Preprints. https://doi.org/10.20944/preprints202406.0975.v1

Wickham, H. 2014. "Tidy data," *Journal of Statistical Software* (59:10), pp. 1 - 23. doi:http://dx.doi.org/10.18637/jss.v059.i10

Wikina S. B. (2014). What caused the breach? An examination of use of information technology and health data breaches. *Perspectives in health information management*, *11*(Fall), 1h. https://pmc.ncbi.nlm.nih.gov/articles/PMC4272442/

Wing, J. M. (2019). The Data Life Cycle. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.e26845b4