risks

MDPI

Article

Modeling Under-Reporting in Cyber Incidents

Seema Sangari [1], Eric Dallal [2] and Michael Whitman [1,3,*]

[1] School of Data Science and Analytics, Kennesaw State University, 3391 Town Point Dr. NW, Kennesaw, GA 30144, USA
[2] Verisk Extreme Event Solutions, Lafayette City Center, 2 Ave de Lafayette 2nd Floor, Boston, MA 02111, USA
[3] Institute of Cybersecurity Workforce Development, Kennesaw State University, 3203 Campus Loop Road, Kennesaw, GA 30144, USA
* Correspondence: mwhitman@kennesaw.edu

Abstract: Under-reporting in cyber incidents is a well-established problem. Due to reputational risk and the consequent financial impact, a large proportion of incidents are never disclosed to the public, especially if they do not involve a breach of protected data. Generally, the problem of under-reporting is solved through a proportion-based approach, where the level of under-reporting in a data set is determined by comparison to data that is fully reported. In this work, cyber insurance claims data is used as the complete data set. Unlike most other work, however, our goal is to quantify under-reporting with respect to multiple dimensions: company revenue, industry, and incident categorization. The research shows that there is a dramatic difference in under-reporting—a factor of 100—as a function of these variables. Overall, it is estimated that only approximately 3% of all cyber incidents are accounted for in databases of publicly reported events. The output of this work is an under-reporting model that can be used to correct incident frequencies derived from data sets of publicly reported incidents. This diminishes the "barrier to entry" in the development of cyber risk models, making it accessible to researchers who may not have the resources to acquire closely guarded cyber insurance claims data.

Keywords: cyber insurance; cyber risk; under-reporting

Citation: Sangari, Seema, Eric Dallal, and Michael Whitman. 2022. Modeling Under-Reporting in Cyber Incidents. *Risks* 10: 200. https://doi.org/10.3390/risks10110200

Academic Editors: Gian Paolo Clemente, Nino Savelli and Diego Zappa

Received: 29 July 2022
Accepted: 9 October 2022
Published: 22 October 2022

1. Introduction

Under-reporting is the problem of event occurrences being reported only partially, especially where the mechanism/conditions under which reporting occurs is not representative. It is a common problem and is mostly studied in the medical field Hirvonen et al. (1997); Schuitemaker et al. (1997). The COVID pandemic provides a perfect example of this. Most people only get tested if they show symptoms or have been exposed to someone known to have been infected U.S. Centers for Disease Control and Prevention (2022). This results in an unrepresentative sample of the full population. Medical studies are often based on a sample of patients considering the cost associated and difficulty associated in obtaining complete census data Stratton (2021). Ideally, everyone should be tested for COVID regularly but this would be inordinately expensive. As a result, a less expensive small but accurate data set is a practical solution. The small data set would be obtained by appropriately sampling a portion of the population. This results in a sample that is unbiased by construction. The under-reporting correction factor is computed by comparing the small complete[1] but unbiased data set to the larger but biased data set.

Cyber research has always struggled with the established and growing problem of under-reporting of cyber incidents Touhill (2019). As a result, it is difficult to obtain a clear picture of the true rate of cyber incidents. Organizations are reluctant to report incidents when not required as doing so could directly impact their businesses by causing a loss of reputation or by deterring potential prospects. In addition, there is a belief that attackers will never be caught so that victims may consider incident reporting a waste of

time Cavusoglu et al. (2004); Fafinski and Minassian (2009); Goucher (2010); McMurdie (2016); McGuire (2013); Swinhoe (2019).

One of the prime cyber data providers, Advisen, gathers publicly reported information through news, multiple online data breach clearinghouses or state and federal governments and agencies using the freedom of information act Romanosky (2016). Highlighting the limitations of this cyber data, Romanosky (2016) stated that the cyber data is a collection of incidents which are detected and disclosed publicly. Such data neither includes undetected incidents nor incidents that are detected but not disclosed. Most of the state laws are in place to inform individuals when their personal identifiable information is compromised Romanosky et al. (2011). Certainly, such data does not capture the information cyber insurers have. The attacked organization might not report an incident but would definitely file a claim to cover the associated cost for any event with material losses. Palsson et al. (2020) also noted the problem of unreported events and focused on answering those questions that can potentially still be answered by the available data, such as estimating the cost of a cyber incident.

Cyber and Infrastructure Security Agency (2020) cited four key factors influencing under-reporting problem—Variation in under-reporting over time (monthly, quarterly, yearly), reporting requirements across different industries, analytical challenges to compute measures using historical data, and comparing the outcomes from different datasets. In this research, the under-reporting problem is addressed at the US level and compared between five industries as well as three different incident types.

Hence, correcting for under-reporting in data sets of publicly reported cyber incidents becomes necessary when building models from data that is typically included in these data sets, but not frequently included in claims data. This includes, e.g., the number of records lost in data compromise events. As will be shown, cyber events impacting small companies are more under-reported than those impacting large companies. Since the number of records compromised is typically higher for incidents impacting large companies, this skews the distributions that are directly constructed from data sets of publicly reported incidents. This is similar to the problem noted in the context of road accidents being under-reported Elvik and Mysen (1999).

Brookmeyer and Gail (1986) advised on exclusion of under-reported data. However, Wood et al. (2016) stated that this would lead to biased statistical models. Elvik and Mysen (1999) argued that under-reporting causes incomplete data sets and results in the analysis being biased towards the reported data only.

There are a number of studies found in the medical domain addressing under-reporting Hirvonen et al. (1997); Schuitemaker et al. (1997). The level of under-reporting in a data set is estimated by comparing with fully reported but less plentiful data Elvik and Mysen (1999); Wood et al. (2016), and more.

More than 85% of the literature applied a proportions approach that compares a smaller but unbiased set of records against the larger population data—the approach is easy to implement but obtaining a reliable data set, even if small, can be challenging.

Hirvonen et al. (1997) studied under-reporting and trends in dietary data to evaluate energy levels and realized that women and over-weight individuals often under-report. Lissner et al. (1989) performed a similar study but only on women data. They applied a multiple regression with various combination of body composition factors as independent predictors and computed a range for the under-reporting level of the mean daily weight change and standard error of mean (SEM^2). Again, the approach is simple but difficult to obtain the necessary data for studies involving individuals.

Hazell and Shakir (2006) collected 37 different studies on adverse drug reactions. They estimated an under-reporting level as the median of the inter-quartile range. This is the most simplistic and rapid approach but it is difficult to obtain the research with under-reporting estimates.

Krantz et al. (2020) and Krantz and Rao (2020) studied COVID-19 data before first peak and proposed a new method with harmonic analysis and wavelets to compute the level of

under-reporting. This approach develops complete data from the incomplete partial data but involves complex mathematical models and is computationally intensive.

The proposed method in this work models the under-reporting correction factor as a function of population characteristics. The study shows that there are extremely large differences in correction factors observed as a function of these variables.

How to Use This Work

This work presents parameters of a model of under-reporting. The frequency of cyber incidents of different types changes rapidly as attacker tactics evolve and the overall number of bad actors increase. However, the level of under-reporting of these cyber events is expected to change more slowly, as this would be primarily a consequence of legal changes. In the US, applicable laws are typically at the state level, making large changes in the level of under-reporting at the national level less likely. Therefore, the model of under-reporting presented here should have continued value for longer than a model of event frequency that would directly provide event frequency as a function of event type and company characteristics.

The model of under-reporting presented here is constructed by joining a number of proprietary data sets (see Section 2). All the constituent data sets are commercially available, with the exception of the claims and policy data which are proprietary and thus unavailable. It is the intent of this project for the results of this work to be used in conjunction with commercially available historical incident data sets and firmographic data sets in order to build unbiased cyber models *without requiring access to claims data*. Providing this model of under-reporting to the academic community should therefore help lower the barrier to entry in the development of cyber models by eliminating the need to acquire claims and policy data.

2. Data

Two proprietary data sets are used for this study: claim-exposure data, which is small but unbiased, and historical incident-IED[3] data, which is large but biased. The proprietary claim-exposure data is a collection of more than 30,000 US cyber insurance policies under-written by multiple insurers and any claim information where there exist claims against those policies. The data set includes a policy ID, start and end dates of the policy, information about the insured company such as employee counts, geographic location, industry, and revenue, and information about any associated claims, including an ID, date, amount, an incident description, and an incident type/categorization (extracted from the incident description).

The proprietary historical incident-IED data set used consists of a collection of more than 100,000 publicly reported historical incidents in the US over a period of decades. Although the data contains fields that are missing for many or even most records, none of these fields were required for this research and hence no filtering of rows or data imputation was done to the data set. The incidents in this data set were gathered via numerous collection methods, including scraping of technology and news websites, Securities Exchange Commission (SEC) filings, and other sources. An aggregated data set was constructed by combining a historical incident data set (Advisen) with a proprietary firmographic data set (IED) of companies that includes, among other things, their name, location, industry, and revenue.

From this point onwards, all provided measures and statistics are based on the US subset of incidents.

Limitations: There might be cyber incidents which are not reported to insurers because the losses associated with them are less than the deductible. This research does not address this. The incident types examined are covered under any insurance policy. Incidents of other types may have different under-reporting levels.

The number of companies in the firmographic data set was validated against US census data and is well aligned in the number of companies by industry and size for

companies with at least 15 employees. Below this value (which accounts for the majority of US companies), the firmographic data set contains a much larger number of companies than indicated in US census data. This could have the impact of overestimating the level of under-reporting for small companies—those whose revenue is below approximately one million USD.

3. Methodology

The proposed approach aims to correct for under-reporting in cyber incidents in more than one dimension–revenue, event type and industry. A model of event frequency as a function of company revenue, industry, and incident type is obtained for both the claims-exposure data set and the historical incident-IED data set. An under-reporting factor is computed as a function of these variables by taking the ratio of these.

Due to insufficient data when examining combinations of revenue, industry, and incident type, separability of the models is assumed. That is, the incident frequency for a combination of variables is expressed as a product of functions of a single variable each Weirich (2015). First, under-reporting corrections are computed as a function of revenue as shown in Equation (1). Assuming the revenue corrections are correct, the under-reporting corrections for revenue given incident type are computed as a function of revenue and incident type as shown in Equation (2). Similarly, the under-reporting corrections for revenue given industry are computed as a function of revenue and industry as shown in Equation (3). Extending further, the under-reporting corrections for revenue given incident type and industry can be computed as a function of the revenue, r, incident type, t, and industry, i, as shown in Equations (2) and (3), assuming correctness of the preceding under-reporting model for revenue, $UR(r)$ (Equation (1)).

$$\text{Function of revenue, } r\text{: } UR_r(r) \tag{1}$$

$$\text{Function of revenue, } r\text{, and incident type, } t\text{: } UR_{r,t}(r,t) = UR_r(r) \times UR_t(t) \tag{2}$$

$$\text{Function of revenue,} r \text{ and industry, } i\text{: } UR_{r,i}(r,i) = UR_r(r) \times UR_i(i) \tag{3}$$

where UR refers to under-reporting model.

A model of under-reporting as a function of revenue is constructed first as this was found to be the variable with which there is the most variation in under-reporting.

3.1. Revenue Based Corrections

The factor $UR_r(r)$ is computed as the proportion of event frequency as a function of revenue, r from claim-exposure data, $freq_{CE,r}(r)$, and historical incident-IED data, $freq_{Inc-IED,r}(r)$, as shown in Equation (4).

$$UR_r(r) = \frac{freq_{CE,r}(r)}{freq_{Inc-IED,r}(r)} \tag{4}$$

For claim-exposure data, the raw revenue frequency, $freq_{CE,raw,r}(r)$, is computed as the ratio of the number of claims and the sum of policy years of the policies under-written for companies with given revenue r, as shown in Equation (5). The policy year refers to the time period, in years, the policy is written for.

$$freq_{CE,raw,r}(r) = \frac{Claims(r)}{\sum_{p \in P_r} Policy\ Years(p)} \tag{5}$$

where P_r refers to policies written for companies with revenue r.

Similarly, for historical incident-IED data, the raw event frequency as a function of revenue, $freq_{Inc-IED,raw,r}(r)$, is computed as the ratio of the number of incidents and the number of companies with given revenue r, as shown in Equation (6).

$$freq_{Inc-IED,raw,r}(r) = \frac{Incidents(r)}{N(r)} \tag{6}$$

where $N(r)$ is the number of organizations with revenue, r.

The computed raw event frequencies are then smoothed over a rolling window of size 'd' with revenue taken on a $\log_{10}$ scale. Smoothing is required since most of the raw frequency values are either zero or one as they are based on single policies or companies. This smoothed event frequency, $freq_{Smooth,r}$, is computed as an average of the raw frequencies in the revenue window $(\log_{10} r - d, \log_{10} r + d)$, as shown in Equation (7).

$$freq_{\cdot,Smooth,r}(r) = \frac{\sum_{p \in P} Weight(p) freq_{\cdot,raw,r}(r)}{\sum_{p \in P} Weight(p)} \tag{7}$$

$$\text{where } Weight(p) = \begin{cases} 1 - \frac{|\log_{10} r - \log_{10} p_r|}{d} & \log_{10} p_r \in (\log_{10} r - d, \log_{10} r + d) \\ 0 & \text{else} \end{cases}$$

$$p_r = \text{revenue for policy } p$$

The value of 'd' is chosen to be 1.5 because smaller values of 'd' were showing noisy, non-monotonic behavior whereas larger values resulted in a flattened curve at extreme revenue values.

Considering the trend in event frequencies as a function of revenue shown in Figure 1, an exponential function is fitted to the trend from the claim-exposure data (Figure 1a) and a polynomial function, with event frequency on a $\log_{10}$ scale, is fitted to the trend from the historical incident-IED data (Figure 1b). In both functions, revenue is taken on a $\log_{10}$ scale. This is done due to heavy concentration of companies with smaller revenues, especially in the historical incident-IED data. For the historical incident-IED data, the event frequency is also on a $\log_{10}$ scale as the (uncorrected) event frequencies derived from this data set are vanishingly small for smaller companies. In both cases, the models ensure positive values for event frequency.

The exponential function is defined as the power function of the form shown in Equation (8).

$$Y_{Exp}(x) = ae^{bx} \tag{8}$$

where a and b are the fitted parameters

The parameters a and b computed using RMSE approach, as shown in Equation (9).

$$\underset{a,b}{\text{Minimize}} \quad \sqrt{\sum_{r \in R} \left(freq_{Smooth,r}(r) - Y_{Exp}(r)\right)^2} \tag{9}$$

where R is the set of revenues over all policies. To find the optimal degree for the polynomial used to fit the historical incident-IED data, a polynomial is fitted for multiple degrees on a training set from the historical incident-IED data and tested on the remaining data. Let the polynomial of degree n be denoted by $Y_{Poly}(x;a)$ (as shown in Equation (10)).

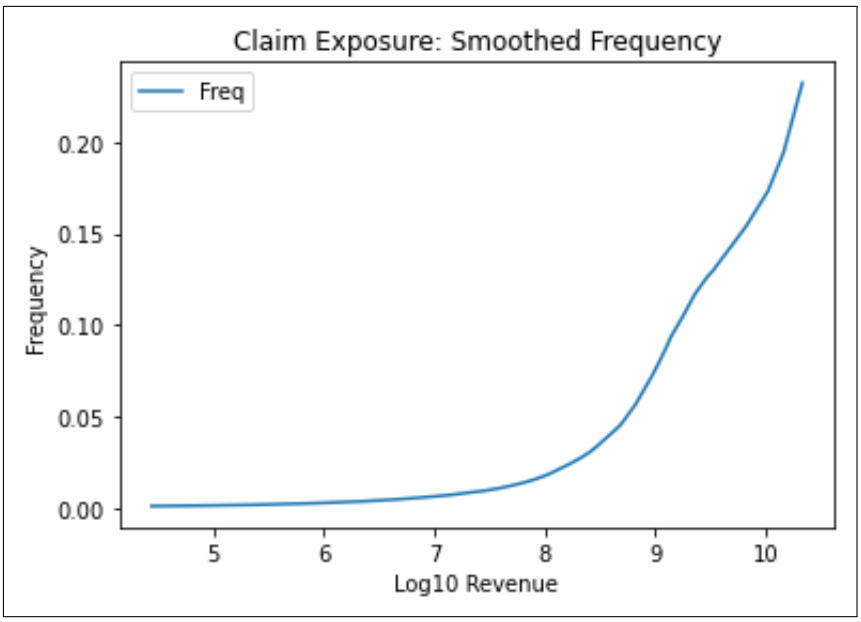
$$Y_{Poly}(x;a) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \tag{10}$$

where a = $(a_0, a_1, \ldots, a_n)$ is a vector of fitted parameters

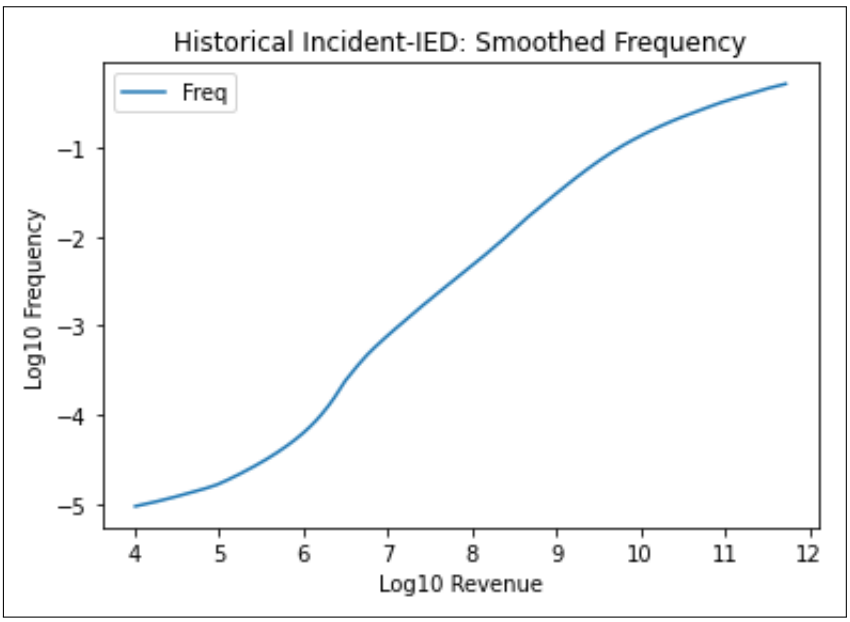
The optimal polynomial degree is four, derived by minimizing the root mean square error (RMSE), as shown in Equation (11).

$$\underset{n\in 0,1,\ldots}{\text{Minimize}} \quad \underset{a\in\mathbb{R}^{n+1}}{\text{Minimize}} \quad \sqrt{\sum_{r\in R}\left(freq_{Smooth,r}(r) - Y_{Poly}(r;a)\right)^2} \tag{11}$$

where n refers to the range of degree to be evaluated for the polynomial and R is the set of revenues over all policies (for the claim exposure dataset) or companies (for the historical incident-IED dataset).



(**a**) Claim Exposure



(**b**) Historical incident-IED

Figure 1. Smoothed Frequency Plots.

The under-reporting correction factors are computed as the ratio of the frequencies computed from the claim-exposure and historical incident-IED models as a function of revenue, as shown in Equation (12).

$$UR_r(r) = \frac{Y_{Exp}(r)}{10^{Y_{Poly}(r)}} \tag{12}$$

where Y_{Exp} is the exponential model used for the claim-exposure data, and Y_{Poly} is the polynomial model used for the historical incident-IED data with frequency on a $\log_{10}$ scale.

3.2. Revenue and Incident Type Corrections

Incident type factors are determined by comparing the event frequency as a function of revenue and incident type to the overall event frequency as a function of revenue (irrespective of any incident type). By the separability assumption, event frequency as a function of revenue and incident type is equal to the corresponding incident type factor multiplied by the revenue-only event frequency.

Since the policies are not specifically under-written for a subset of the incident types of interest, all policies are taken into consideration when computing event frequency w.r.t. revenue and incident type. For claim-exposure data, the raw event frequency as a function of revenue and some incident type t can be computed as shown in Equation (13).

$$freq_{CE,raw,r,t}(r,t) = \frac{Claims(r,t)}{\sum_{p \in P_r} Policy\ Years(p)} \tag{13}$$

where $Claims(r,t)$ refers to the number of claims with incident type t for insureds with revenue r. Similarly, the raw event frequency with respect to revenue and incident type is computed from the historical incident-IED data as shown in Equation (14).

$$freq_{Inc-IED,raw,r,t}(r,t) = \frac{Incidents(r,t)}{N(r)} \tag{14}$$

where $Incidents(r,t)$ refers to number of incidents with revenue r and incident type t and N(r) refers to the number of companies with revenue r. Again, these raw frequencies, $freq_{CE,raw}$ and $freq_{Inc-IED,raw}$, are further smoothed over a rolling window of size d to compute smooth frequencies, $freq_{CE,Smooth}$ and $freq_{Inc-IED,Smooth}$, respectively, as discussed earlier.

Incident type factors, $f_{\cdot,t}(t)$, are then computed under the assumption that the models of frequency as a function of revenue are correct. The frequency scalar for each given incident type, $f_{\cdot,t}(t)$, are computed such that the models of frequency given incident type and revenue can be determined by scaling the overall (i.e., non event type specific) models of frequency given revenue, $freq_{\cdot,Fitted,r}$, by incident type factor, $f_{\cdot,t}(t)$. This is done for both the claim-policy and historical incident-IED data sets, as shown in Equations (15) and (16).

$$freq_{CE,Smooth,r,t}(r,t) \approx f_{CE,t}(t) \times freq_{CE,Fitted,r}(r) \tag{15}$$

$$freq_{Inc-IED,Smooth,r,t}(r,t) \approx f_{Inc-IED,t}(t) \times freq_{Inc-IED,Fitted,r}(r) \tag{16}$$

where $f_{CE,t}(t)$ and $f_{Inc-IED,t}(t)$ are incident type frequency scalars computed by a curve fitting approach that minimizes the sum of squared differences between smoothed frequencies, $freq_{\cdot,Smooth,r,t}(r,t)$, and incident type frequency scalar times model fitted frequencies, $f_{\cdot,t}(t) \times freq_{\cdot,Fitted,r}(r)$, where the sum is taken over the set of revenue values, as shown in Equation (17).

$$\underset{f_{\cdot,t}(t)}{\text{Minimize}} \quad \sum_{r \in R} \left(freq_{\cdot,Smooth,r,t}(r,t) - f_{\cdot,t}(t) \times freq_{\cdot,Fitted,r}(r) \right)^2 \tag{17}$$

where R is the set of policies for claim exposure data and revenue for historical incident data.

Accordingly, the under-reporting correction factor for a given incident type t is computed as a function of both revenue and incident type, as shown in Equations (18) and (19).

$$UR_{r,t}(r,t) = UR_r(r) \times UR_t(t) \approx \frac{f_{CE,t}(t) \times freq_{CE,Fitted,r}(r)}{f_{Inc-IED,t}(t) \times freq_{Inc-IED,Fitted,r}(r)} \tag{18}$$

$$\approx \frac{f_{CE,t}(t)}{f_{Inc-IED,t}(t)} \times UR_r(r) \tag{19}$$

3.3. Revenue and Industry Corrections

Under-reporting corrections for revenue and industry are computed in almost the same way as for revenue and incident type. The only difference is that–unlike incident type–industry is a property of the company itself. Therefore, when computing event frequencies, both the set of claims/historical incidents and the set of policies/companies are filtered to only those for companies of the industry under consideration.

From claim exposure data, the revenue frequency for industry i can be computed as shown in Equation (20).

$$freq_{CE,raw,r,i}(r,i) = \frac{Claims(r,i)}{\sum_{p \in P_{r,i}} Policy\ Years(p)} \tag{20}$$

where $Claims(r,i)$ refers to number of claims with revenue r, and industry i.

From historical incident-IED data, the revenue frequency for industry i can be computed as shown in Equation (21).

$$freq_{Inc-IED,raw,r,i}(r,i) = \frac{Incidents(r,i)}{N_{r,i}} \tag{21}$$

where $Incidents(r,i)$ refers to number of incidents occurred in the industry i and $N_{r,i}$ refers to number of companies with revenue r in the industry i.

4. Results

In this section, under-reporting corrections are presented as a function of revenue, revenue and incident type, and revenue and industry.

4.1. Under-Reporting Factors: Revenue

Figure 2 shows how under-reporting varies as a function of revenue (defined in Equation (12)). These results show that the under-reporting factor is highest for low revenue companies—most events are not reported—and approaches one for high revenue companies—most events are reported.

4.2. Under-Reporting Factors: Revenue and Incident Type

Based on the availability of claim-exposure data, three incident types are investigated: Hacking (HACK), Social Engineering (SOC) and Ransomware (RAN). Although the list of incident types is not exhaustive, these incident types account for the majority of incidents in both the claim-exposure and the historical incident data sets. HACK and SOC are different types of data compromise incidents. An incident which begins with HACK or SOC but ultimately leads to ransom is classified as RAN.

Table 1 shows the computed under-reporting factors for three incident types, and Figure 3 shows these factors as a function of both revenue and event type. Although HACK and SOC have a comparatively lower under-reporting factor when compared to RAN, the correction factor for SOC is more than double that of HACK.

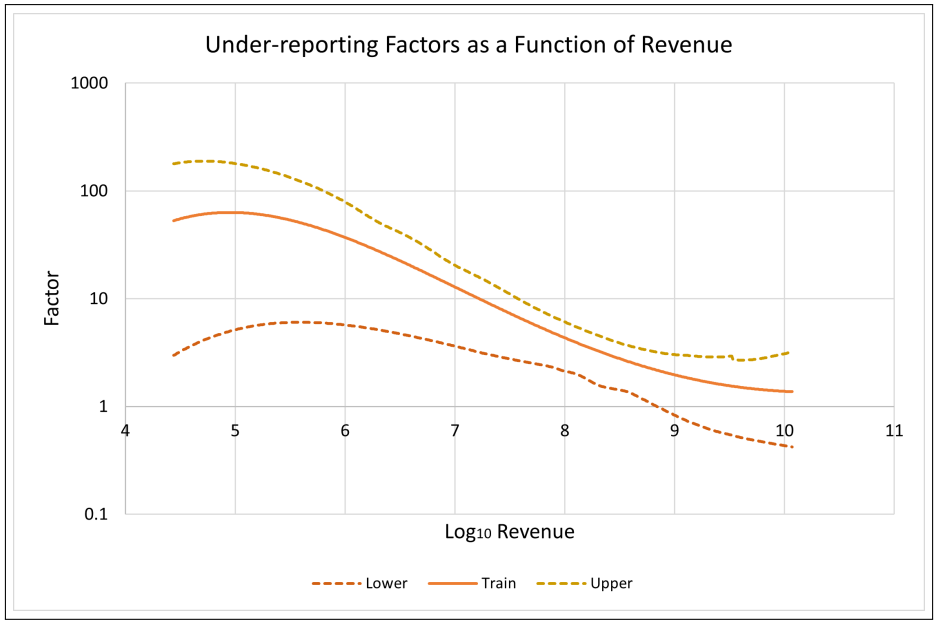


Figure 2. Under-reporting Factors as a function of Revenue.

Table 1. Under-reporting Factors: Incident Type.

Incident Type	Factor
HACK	0.7607
SOC	1.5926
RAN	5.2864

A likely explanation for the much higher factor for RAN is that there are reporting requirements for data compromise incidents such as HACK and SOC whereas RAN does not have such requirements. The results emphasize the need for under-reporting factors to be computed individually by incident type.

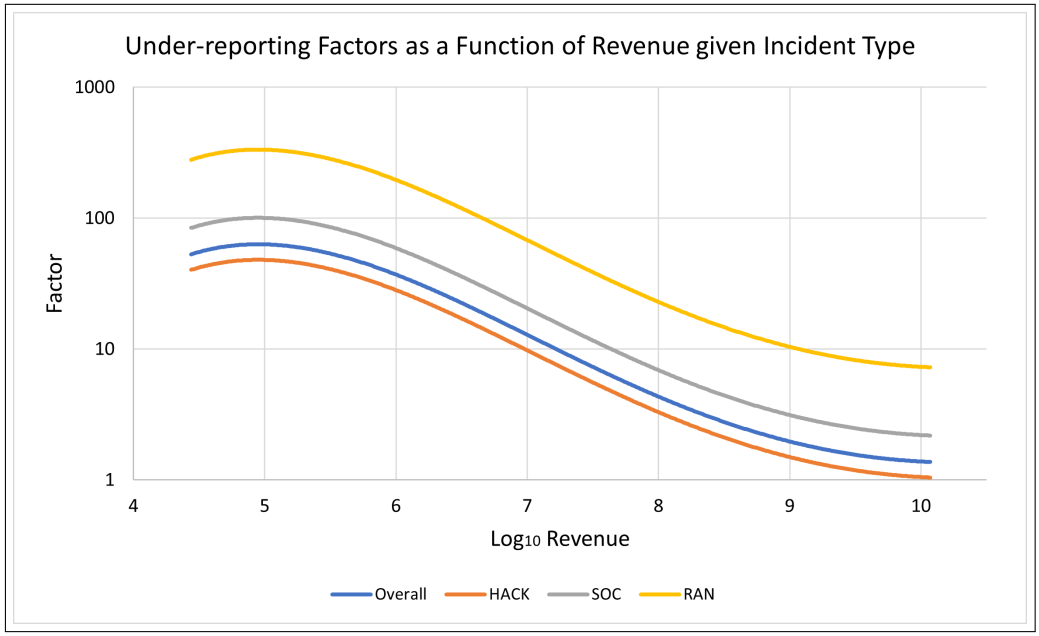


Figure 3. Under-reporting Factors as function of Revenue and Incident Type.

4.3. Under-Reporting Factors: Revenue and Industry

Based on the availability of claim-exposure data, five industries are investigated: Retail Trade (RT), Manufacturing (MFG), Finance and Insurance (FnI), Professional, Scientific, and Technical Services (PSTS), and Wholesale Trade (WT). Table 2 shows the under-reporting factors for five different industries. The RT and MFG industries have under-reporting factors less than one whereas FnI, PSTS and WT have more than one. The WT industry has the largest under-reporting factor at 4.2.

Table 2. Under-reporting Factors: Industry.

Industry	Factor
RT	0.0931
MFG	0.8577
FnI	1.4919
PSTS	1.1985
WT	4.2024

Figure 4 shows how under-reporting factors vary as a function of revenue for the five examined industries in comparison to the overall (i.e., industry-independent) under-reporting curve. It should be noted that the under-reporting factor for RT is below one at revenues above approximately ten million; this indicates that the separability assumption may not be adequate for this industry. Again, these results show that there are substantial differences in under-reporting factors for different industries.

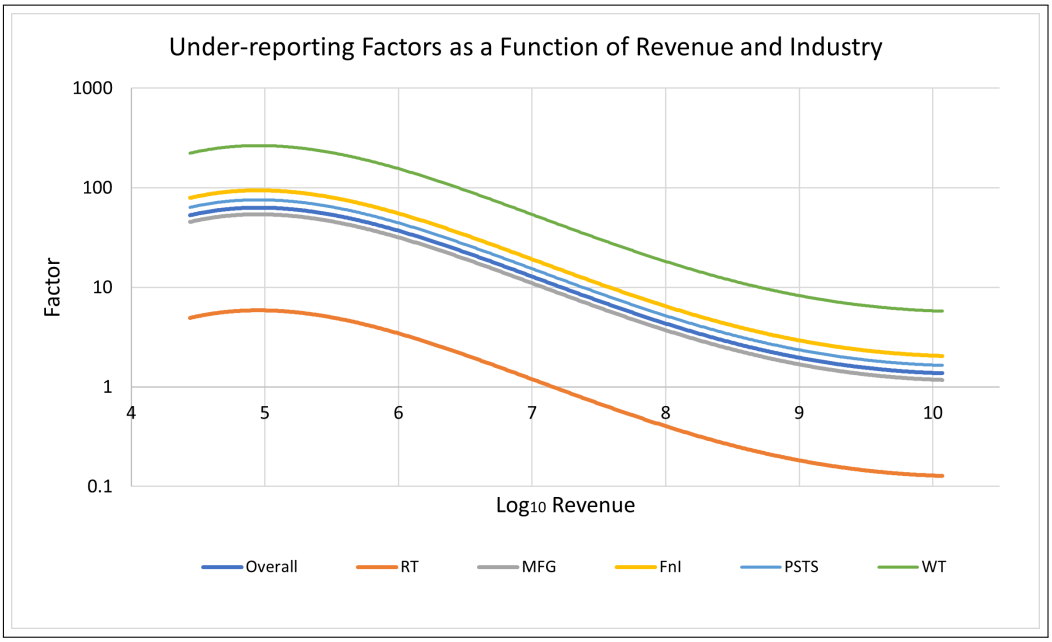


Figure 4. Under-reporting Factors as function of Revenue and Industry.

5. Validation

To validate the model, both the claim-exposure and the historical incident-IED data sets are split into two-thirds for training and one-third for test, and an under-reporting model is built from each of 100 bootstrapped samples of the training data. This produces 95% confidence intervals for the model, against which a model built from the test data is compared. The train-test split of the claim-exposure data used stratified sampling to keep the proportion of policies with claims approximately the same in both the train and test sets. The train-test split was not balanced with respect to any other variables.

With 100 samples each from the independent claim-exposure and historical incident-IED data sets, 10,000 under-reporting factors are computed by comparing all claim-exposure samples with each historical incident-IED sample. The 95% confidence intervals for the under-reporting factors are computed separately for $UR_r(r)$, $UR_t(t)$, and $UR_i(i)$ to validate against the corresponding factors obtained from the test data. Figure 5 shows the validation of the under-reporting curve for revenue with 95% confidence interval bands. The plot indicates that there is higher level of under-reporting for organizations with lower revenue when compared to ones with higher revenue. The under-reporting factors could be more than 100 for low revenue organizations but are found to be lower than one for organizations with revenues above 100 million in the test data. This is likely a statistical anomaly, as the confidence intervals widen for larger revenues, and the under-reporting factor computed at these revenues is at the lower 95% confidence interval.

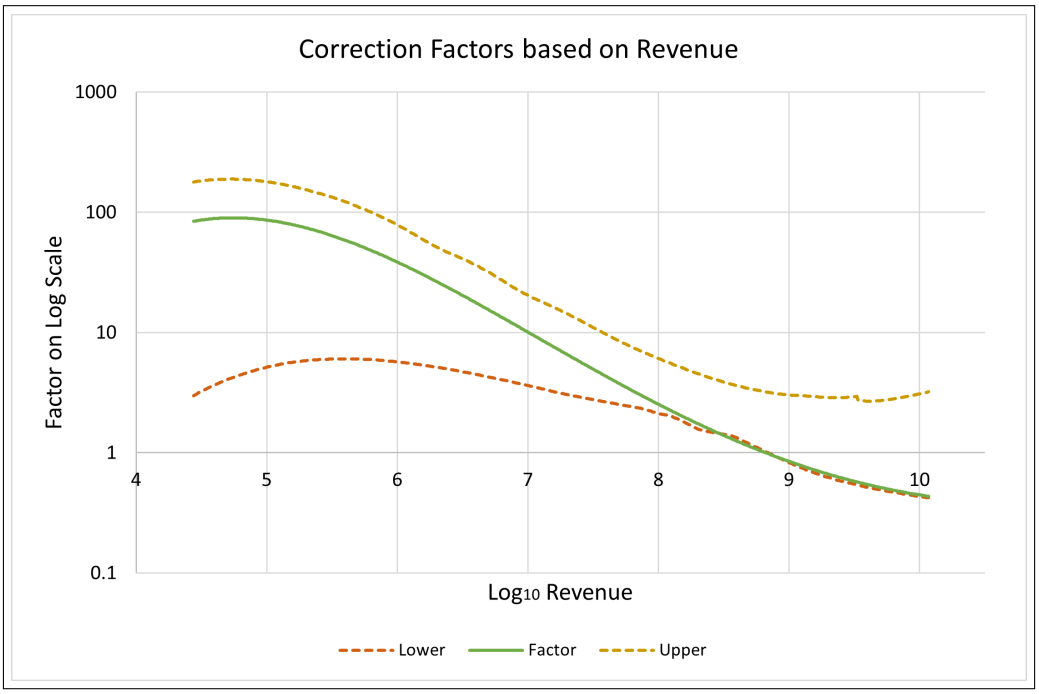


Figure 5. Under-reporting Factors: Revenue.

Figure 6 shows under-reporting factors for three incident types. The under-reporting factor for RAN could range from 1.1 to 5.5. The upper estimate of the HACK under-reporting factor is less than the lower estimate for RAN, and the upper estimate of the SOC under-reporting factor is around 2.2 times the lower estimate for RAN. There is a much larger range of uncertainty in the SOC under-reporting estimate.

Figure 7 shows that there is a contrast observed between the retail and wholesale trade industries where RT has the lowest under-reporting factor and WT has the highest. A potential explanation for this is that retail trade is a business-to-consumer (B2C) industry and would therefore have large quantities of PII[4], placing greater legal requirements on them in the event of a data breach. Conversely, wholesale trade is a business-to-business (B2B) industry. The latter point, however, should in principle apply to manufacturing as well, which has one of the lower under-reporting factors.

The under-reporting factor for the FnI industry based on test data is found to be above the 95 percentile range, whereas the under-reporting factor for the WT industry based on test data is found to be below the 95 percentile range. This could be a consequence of the split in data between the train and test sets, which used sampling that was stratified on claims but not on revenue, incident type or industry.

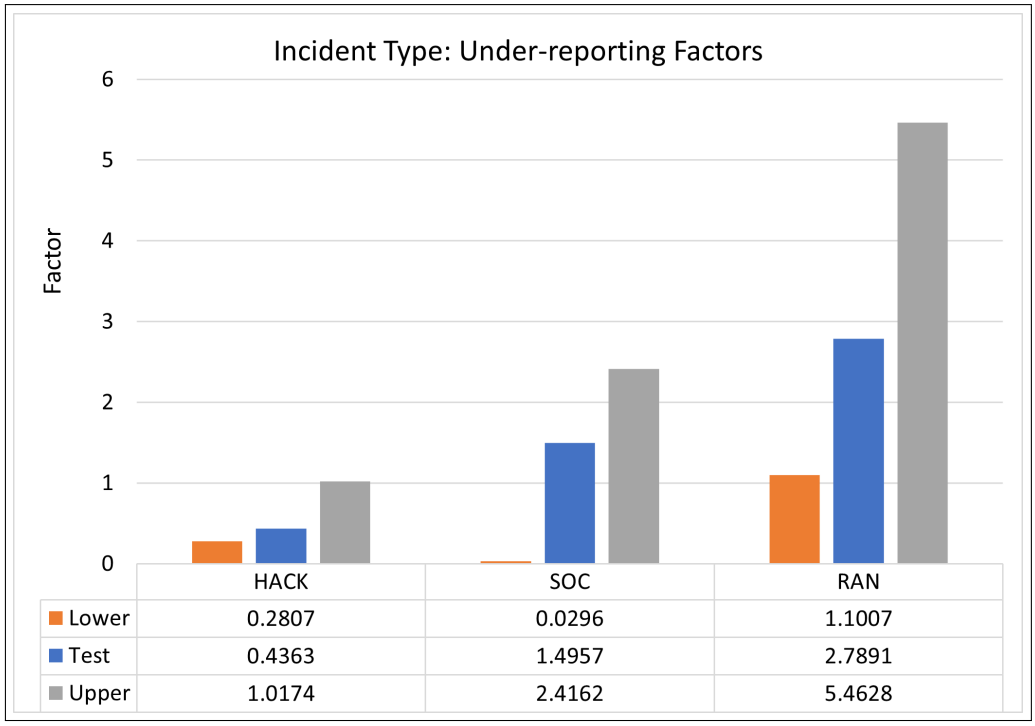


Figure 6. Under-reporting Factors: Incident Type.

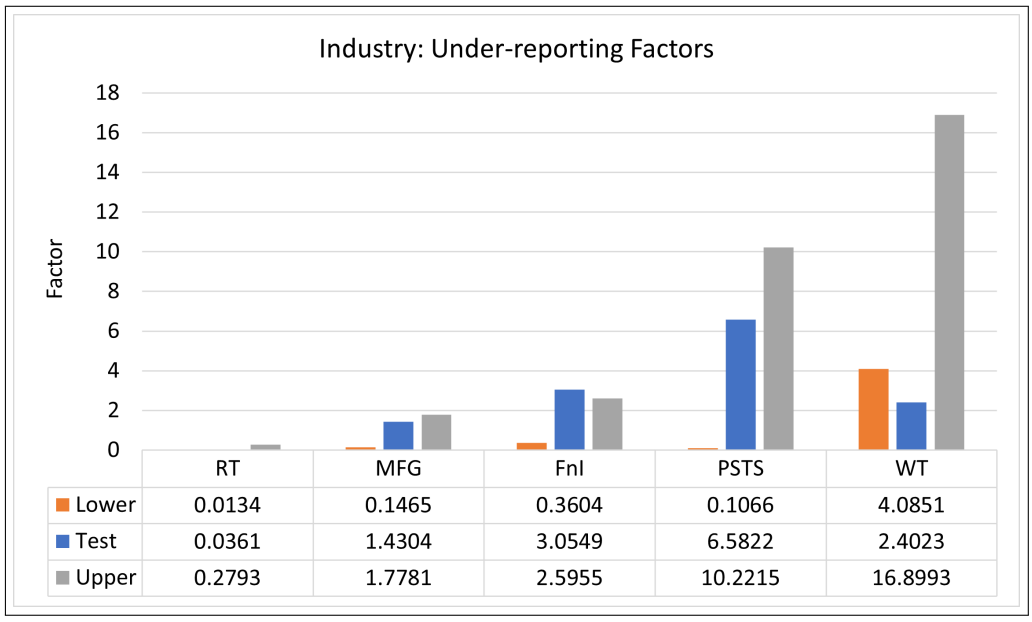


Figure 7. Under-reporting Factors: Industry.

6. Conclusions

This research proposed a method to quantify the extent of under-reporting in data sets of public cyber incidents. It quantified the under-reporting factor with respect to revenue, as well as some incident types and industries. It found significant differences in under-reporting with respect to each of these variables, but especially with respect to revenue, where lower revenue companies were found to have approximately 100 times the under-reporting of larger revenue companies. With a model of under-reporting as a function of revenue as a baseline, under-reporting factors were computed for three incident types: hacking (HACK), social engineering (SOC) and ransomware (RAN). Again, this research showed large differences in these factors, with RAN having a significantly higher level of under-reporting compared to either HACK or SOC. Thirdly, under-reporting factors

were computed for five industries–Retail Trade (RT), Manufacturing (MFG), Finance and Insurance (FnI), Professional Scientific Technical Services (PSTS) and Wholesale Trade (WT)–again under the assumption that the model of under-reporting with respect to revenue is correct. The research showed that the WT, FnI, and PSTS industries have greater levels of under-reporting whereas the MFG and especially the RT industries have lower levels. The research indicates the necessity not only to correct for under-reporting in data sets of publicly reported cyber incidents, but to take multiple variables in consideration when doing so.

Using the under-reporting model and the historical incident-IED data, an estimate of the overall proportion of all incidents that are accounted for by databases of publicly reported events can be determined. By scaling the number of reported events at each revenue by the reciprocal of the under-reporting factor for that revenue and summing over all revenues, an estimate for the true number of cyber incidents can be arrived at. Taking the ratio of the number of reported incidents and this value gives an estimate of 3% for the proportion of events that are accounted for in databases of publicly reported events. This is not far from the level of under-reporting for the smallest companies (since the vast majority of companies have revenues below 1M USD), but not quite as dramatic since smaller companies nevertheless suffer fewer cyber incidents when compared to larger ones.

Author Contributions: Conceptualization, E.D.; methodology, S.S. and E.D.; software, S.S.; validation, E.D.; formal analysis, S.S.; investigation, S.S. and E.D.; data curation, E.D. and S.S.; writing–original draft preparation, S.S. and E.D.; writing–review and editing, E.D. and M.W.; visualization, S.S.; supervision, E.D. and M.W.; project administration, E.D. and M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: This work was produced using three types of third party sets: historical incident data sets, firmographic data sets, and sets of claims and policy data. All are proprietary and cannot be shared by the authors. The first two types of data sets are commercially available from numerous sources. The last type is not commercially available.

Acknowledgments: The study is conducted with Verisk Extreme Events Solutions using their proprietary cyber data.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

1 Data are considered complete where all incidents are reported for a particular population.

2 $SEM = \frac{\sigma}{\sqrt{n}}$, where n is sample size.

3 IED stands for Industry Exposure Database.

4 Personal Identifiable Information.

References

Brookmeyer, Ron, and Mitchell H. Gail. 1986. Minimum Size of the Acquired Immunodeficiency Syndrome (Aids) Epidemic in the United States. *The Lancet* 328: 1320–22. [CrossRef]

Cavusoglu, Huseyin, Birendra Mishra, and Srinivasan Raghunathan. 2004. The effect of internet security breach announcements on market value: Capital market reactions for breached firms and internet security developers. *International Journal of Electronic Commerce* 9: 70–104. [CrossRef]

Cyber and Infrastructure Security Agency. 2020. *Cost of a Cyber Incident: Systematic Review and Cross-Validation*; Technical Report. Arlington: CISA.

Elvik, Rune, and Anne Borger Mysen. 1999. Incomplete accident reporting: Meta-analysis of studies made in 13 countries. *Transportation Research Record* 1665: 133–40. [CrossRef]

Fafinski, Stefan, and Neshan Minassian. 2009. *UK Cybercrime Report 2009*. London: Garlik.

Goucher, Wendy. 2010. Being a cybercrime victim. *Computer Fraud and Security* 2010: 16–18. [CrossRef]

Hazell, Lorna, and Saad A.W. Shakir. 2006. Under-reporting of adverse drug reactions: A systematic review. *Drug Safety* 29: 385–96. [CrossRef] [PubMed]

Hirvonen, Tero, Satu Männistö, Eva Brita Roos, and Pirjo Pietinen. 1997. Increasing prevalence of underreporting does not necessarily distort dietary surveys. *European Journal of Clinical Nutrition* 51: 297–301. [CrossRef] [PubMed]

ISACA. 2019. New Study Reveals Cybercrime May Be Widely Underreported Even When Laws Mandate Disclosure. *ISACA Press Release*. 1–2. Available online: https://www.isaca.org/why-isaca/about-us/newsroom/press-releases/2019/new-study-reveals-cybercrime-may-be-widely-underreported-even-when-laws-mandate-disclosure (accessed on 21 September 2021)

Krantz, Steven G., and Arni S. R. Srinivasa Rao. 2020. Level of underreporting including underdiagnosis before the first peak of COVID-19 in various countries: Preliminary retrospective results based on wavelets and deterministic modeling. *Infection Control and Hospital Epidemiology* 41: 857–59. [CrossRef] [PubMed]

Krantz, Steven G., Peter Polyakov, and Arni S.R.Srinivasa Rao. 2020. True epidemic growth construction through harmonic analysis. *Journal of Theoretical Biology* 494: 110243. [CrossRef]

Lissner, Lauren, Jean-Pierre Habicht, Barbara J. Strupp, David A. Levitsky, Jere D. Haas, and Daphne A. Roe. 1989. Body composition and energy intake: Do overweight women overeat and underreport? *American Journal of Clinical Nutrition* 49: 320–25. [CrossRef]

McGuire, Mike, and Samantha Dowling. 2013. *Cyber Crime: A Review of the Evidence*. Technical Report. London: Home Office, vol. 75, pp. 1–29.

McMurdie, Charlie. 2016. The cybercrime landscape and our policing response. *Journal of Cyber Policy* 1: 85–93. [CrossRef]

Palsson, Kjartan, Steinn Gudmundsson, and Sachin Shetty. 2020. Analysis of the impact of cyber events for cyber insurance. *Geneva Papers on Risk and Insurance: Issues and Practice* 45: 564–79. [CrossRef]

Romanosky, Sasha. 2016. Examining the costs and causes of cyber incidents. *Journal of Cybersecurity* 2: 121–35. [CrossRef]

Romanosky, Sasha, Rahul Telang, and Alessandro Acquisti. 2011. Do data breach disclosure laws reduce identity theft? *Journal of Policy Analysis and Management* 30: 256–86. [CrossRef]

Schuitemaker, Nico, Jos Van Roosmalen, Guus Dekker, Pieter Van Dongen, Herman Van Geijn, and Jack Bennebroek Gravenhorst. 1997. Underreporting of maternal mortality in the Netherlands. *Obstetrics and Gynecology* 90: 78–82. [CrossRef]

Stratton, Samuel J. 2021. Population Research: Convenience Sampling Strategies. *Prehospital and Disaster Medicine* 36: 373–74. [CrossRef] [PubMed]

Swinhoe, Dan. 2019. Why Businesses Do Not Report Cybercrimes to Law Enforcement. Available online: https://www.csoonline.com/article/3398700/why-businesses-don-t-report-cybercrimes-to-law-enforcement.html (accessed on 22 September 2021).

U.S. Centers for Disease Control and Prevention. 2022. COVID-19 Quarantine and Isolation. Available online: https://www.cdc.gov/coronavirus/2019-ncov/your-health/isolation.html (accessed on 19 July 2022).

Weirich, Paul. 2015. Separability. In *Models of Decision Making*, 1st ed. Cambridge: Cambridge University Press, chp. 1, pp. 23–50.

Wood, Jonathan S., Eric T. Donnell, and Christopher J. Fariss. 2016. A method to account for and estimate underreporting in crash frequency research. *Accident Analysis and Prevention* 95: 57–66. [CrossRef] [PubMed]