Author/s:
Avanzi, B;Tan, X;Taylor, G;Wong, B

Title:
On the Evolution of Data Breach Reporting Patterns and Frequency in the United States: A Cross-State Analysis

Date:
2025

Citation:
Avanzi, B., Tan, X., Taylor, G. & Wong, B. (2025). On the Evolution of Data Breach Reporting Patterns and Frequency in the United States: A Cross-State Analysis. North American Actuarial Journal, 29 (4), pp.833-864. https://doi.org/10.1080/10920277.2025.2457491.

Persistent Link:
https://hdl.handle.net/11343/362822

North American Actuarial Journal

ISSN: 1092-0277 (Print) 2325-0453 (Online) Journal homepage: www.tandfonline.com/journals/uaaj20

On the Evolution of Data Breach Reporting Patterns and Frequency in the United States: A Cross-State Analysis

Benjamin Avanzi, Xingyun Tan, Greg Taylor & Bernard Wong

To cite this article: Benjamin Avanzi, Xingyun Tan, Greg Taylor & Bernard Wong (2025) On the Evolution of Data Breach Reporting Patterns and Frequency in the United States: A Cross-State Analysis, North American Actuarial Journal, 29:4, 833-864, DOI: 10.1080/10920277.2025.2457491

To link to this article: https://doi.org/10.1080/10920277.2025.2457491

North American Actuarial Journal, 29(4), 833–864, 2025

ISSN: 1092-0277 print / 2325-0453 online
DOI: 10.1080/10920277.2025.2457491

On the Evolution of Data Breach Reporting Patterns and Frequency in the United States: A Cross-State Analysis

Benjamin Avanzi,[1] Xingyun Tan,[1] Greg Taylor,[2] and Bernard Wong[2]
[1]*Centre for Actuarial Studies, Department of Economics, University of Melbourne, Victoria, Australia*
[2]*School of Risk and Actuarial Studies, UNSW Australia Business School, UNSW Sydney, New South Wales, Australia*

Understanding the emergence of data breaches is crucial for cyber insurance and risk management. However, analyses of data breach frequency trends in the current literature lead to contradictory conclusions. We put forward that those discrepancies may be (at least partially) due to inconsistent data collection standards, as well as reporting patterns, over time and space. We set out to carefully control both. In this article, we conduct a joint analysis of state attorneys general's publications on data breaches across eight states (namely, California, Delaware, Indiana, Maine, Montana, North Dakota, Oregon, and Washington), all of which are subject to established data collection standards; namely, state data breach (mandatory) notification laws. Thanks to our explicit recognition of these notification laws, we are capable of modeling frequency of breaches in a consistent and comparable way over time. Hence, we are able to isolate and capture the complexities of reporting patterns, adequately estimate incurred but not reported (IBNR) data breaches, and yield a highly reliable assessment of historical frequency trends in data breaches. Our analysis also provides a comprehensive comparison of data breach frequency across the eight U.S. states, extending knowledge on state-specific differences in cyber risk, which has not been extensively discussed in the current literature. We thus illustrate how each state's unique regulations, market dynamics, demographic profiles, and risk factors can significantly impact insurance products and pricing. Furthermore, we uncover novel features not previously discussed in the literature, such as differences in cyber risk frequency trends between large and small data breaches (i.e., breaches affecting more or fewer state residents), due to differences in state definitions of reportable data breaches. Overall, we find that the reporting delays are lengthening. We also elicit commonalities and heterogeneities in reporting patterns across states, severity levels, and time periods. After adequately estimating IBNRs, we find that frequency is relatively stable before 2020 and increasing after 2020. This is consistent across states. Implications of our findings for cyber insurance reserving, pricing, underwriting, and experience monitoring are discussed.

1. INTRODUCTION

1.1. Background

As the Internet and other digital networks are becoming increasingly vital to the functioning of the global economy, the threat posed by cybercriminals has risen in prominence (OECD 2020). In 2022, cybercrime was projected to result in an estimated economic loss of US$8 trillion in 2023, with the figure expected to rise to US$10.5 trillion annually by 2025 (Cybersecurity Ventures 2022).

In addition to adopting effective cyber hygiene practices, businesses are turning to cyber insurance policies for financial coverage and expert guidance in preventing and managing cyber incidents (Deloitte 2020). Cyber insurance direct written premiums in the United States in 2021 were around US$6.5 billion, an increase of 61% from 2020 (National Association of Insurance Commissioners 2022). The global estimated gross direct premiums written in 2022 reached approximately US$14 billion, with the United States contributing more than half of the total (Insurance Business 2023). A survey conducted by Marsh and Microsoft found that 61% of organizations purchase some type of cyber insurance (Marsh 2022). A major component of cyber risk is data breaches, which are the focus of the article. According to National Association of Attorneys General

Address correspondence to Xingyun Tan, Centre for Actuarial Studies, Department of Economics, The University of Melbourne, Level 3, FBE Building, 111 Barry Street, Parkville, VIC 3010, Australia. E-mail: xingyunt@student.unimelb.edu.au

(2024), "A data breach is the illegal and unauthorized access to personal information that jeopardizes its security, confidentiality, or integrity."

Generally speaking, insurers operating across several states in the United States must accurately account for differences across those jurisdictions, such as regulation, market dynamics, and generally any risk factor that can significantly impact insurance products and pricing. Examples of insurance types that vary significantly across states include health insurance, workers' compensation, auto insurance, and homeowners' and renters' insurance (American Academy of Actuaries 2009; National Academy of Social Insurance 2022; Insurance Information Institute 2023; Allchoice Insurance 2024). Cyber insurance is no different (see, e.g., Chen et al. 2023; Cho, Eling, and Jung 2024).

1.2. Data Breach Reporting Patterns and Frequency

To more accurately price cyber insurance, one needs to understand the statistical properties of various types of cyber incidents and model their frequency and/or severity. Unfortunately, we believe that the current literature's understanding of the evolution of data breaches is lacking. Though there are several rigorous analyses of frequency trends, their conclusions do not agree (see, e.g., Maillart and Sornette 2010; Edwards, Hofmeyr, and Forrest 2016; Romanosky 2016; Xu et al. 2018; Jung 2021; Wheatley, Hofmann, and Sornette 2021; Eling, Ibragimov, and Ning 2023). Where could those discrepancies come from?

It is important to note that conclusions on cyber frequency trends are generally based on three main cyber databases: Data Breach Chronology provided by Privacy Rights Clearinghouse[1] (Privacy Rights Clearinghouse 2024), Cyber Loss Data by Advisen (Advisen 2024), and SAS OpRisk Global Data by SAS (SAS 2024). For the rest of this article, we will refer to these datasets as follows: the PRC dataset, the Advisen dataset, and the SAS dataset. Unfortunately, the exact data collection standards of these datasets are unknown, because they are secondary data sources that collect data from multiple sources (e.g., media, state attorneys general, company websites). Because all three datasets are subject to unknown data collection standards, it is challenging to judge the reliability of any particular conclusion.

Some research suggests an increase in cyber event counts over time (e.g., Eling, Ibragimov, and Ning 2023). Without questioning the worth of those studies (whose main focus is generally elsewhere), it is unclear whether the observed increase is due to an actual increase of risk frequency or not. In particular, factors that enhance data collection capacities over time could drive increases in event counts, even if the level of risks remains unchanged. They include the following:

1. The introduction of various reporting mandates at different times may have led to sudden increases in the number of events reported. Businesses were forced to report more incidents, subsequently inflating the dataset (Jung 2021; Aldasoro et al. 2022). Both the PRC and the Advisen datasets collect data from data breach notification laws of various states, which are introduced at different points in time (Advisen 2024; Privacy Rights Clearinghouse 2024).
2. Increasing media attention in cybersecurity (Harry and Gallagher 2023) may have led to an increase in the number of events collected by the dataset from media sources. The Advisen the SAS datasets source part of their data from the media (Wei, Li, and Zhu 2018; Malavasi et al. 2022).
3. The increasing number of data sources used by data maintainers over time may have led to increases in the number of events collected by the dataset over time (Palsson, Gudmundsson, and Shetty 2020). This issue is likely to be present in all three datasets (the PRC, the Advisen, and the SAS datasets).

Overall, to reliably assess frequency trends of cyber risks, we should analyze data that follow established and consistent data collection standards over time and space. This approach serves two critical purposes: firstly, it allows us to precisely delineate the scope of our conclusions; secondly, it helps mitigate the influence of the above biases on event counts, thereby enabling a more nuanced understanding of the evolution of cyber risks.

1.3. State Attorney General Data Breach Information

In contrast to the three databases mentioned in the previous section, the publications of data breaches by state attorneys general are not influenced by the issue of enhanced data collection capacities. This allows for a more reliable assessment of frequency trends in cyber risks or cyber insurance risks. Details can be found in Maryland Attorney General (2023); Montana Department of Justice (2023); Oklahoma Office of Management & Enterprise Services (2023); California Attorney

[1]Here we are referring to the public data: Data Breach Chronology Archive (2005–2018).

General (2024); Delaware Attorney General (2024); Hawaii Department of Commerce and Consumer Affairs (2024); Indiana Attorney General (2024); Iowa Attorney General (2024); Maine Attorney General (2024); New Hampshire Department of Justice (2024); New Jersey Cybersecurity & Communications Intergration Cell (2024); North Dakota Attorney General (2024); Oregon Department of Justice (2024); Texas Attorney General (2024); Vermont Attorney General (2024); Washington State Office of the Attorney General (2024); Wisconsin Department of Agriculture, Trade and Consumer Protection (2024) and are developed in Section 2. Salient properties of this set of data include the following:

1. State attorneys general's data are collected under state data breach notification laws in the United States (National Conference of State Legislatures 2024). They are less affected by the issue of increasing reporting incentives, because reporting standards have remained consistent over time at the state level. For example, since January 1, 2012, California has required notification of the California attorney general regarding breaches that affect more than 500 California residents.
2. The data sources consist of data breach reports submitted by business entities, rather than being derived from media sources. Consequently, they remain unaffected by the bias associated with escalating media attention.
3. The only data sources are data breach reports submitted by business entities, thereby insulating them from the bias stemming from the introduction of new data sources.

Crucially, state attorneys general's data provide information on reporting delays (i.e., the lag between the date of breach occurrence and the date of breach notification), which allows for the analysis of the number of incurred but not reported (IBNR) data breaches. Furthermore, in state attorneys general's data, each business affected by cyber events is considered as a separate event, which aligns with the perspective of insurers.

In summary, datasets like the Advisen dataset offer detailed information on individual cyber events, adding depth with a greater variety of variables. However, their coverage (how many actual events are included) remains limited and somewhat ambiguously defined. In contrast, data from state attorneys general provide less detail on individual events but are reliable and offer clearly defined coverage. This important observation enables the generation of results in this article that are distinct and original compared to the existing literature, because we are the first to extract and analyze the state attorneys general's data breach data to our stated purposes.

Though commercial datasets like Advisen offer greater depth, the reliability and consistency of government-sourced data, derived from legal mandates and rigorous collection standards, make it a uniquely valuable resource for cyber risk analysis.

1.4. Statement of Contributions

In this article, we provide a rigorous examination of data breach reporting patterns and frequency trends in the information published by state attorneys general, by applying generalized additive models (GAMs) to overdispersed Poisson (ODP) observations based on run-off triangles (e.g., Taylor 2012). By "development profile/pattern" or "reporting pattern" we refer to the pattern or trend over time of event notifications following the occurrence of a cyber event.

We first generate a quarterly estimation of the number of IBNR data breaches. Beyond providing useful insights, this step is essential for reliably assessing the frequency trends of data breaches within the jurisdictions included in our set of data. It is also a contribution in itself. Though Eling, Ibragimov, and Ning (2023) accounted for reporting delay and provide estimates of the number of cyber IBNRs using the Advisen dataset (which already presents potential issues owing to its nature as explained above), their estimation is opaque. The frequency model utilized in Eling, Ibragimov, and Ning (2023) includes some parameters to allow for changes in reporting delay; unfortunately, the algebraic form of this quantity is not visible. Therefore, it is unclear whether and how changes in reporting delays (over time) are accounted for in their study. In contrast, we uncover specific characteristics of the IBNR data breach counts. By recognizing changes of these characteristics over time, we can generate an accurate estimation of the number of IBNR breaches for different quarters of origin.

Once IBNRs are reliably determined, the emergence of total data breaches (whether reported or not) can be analyzed in detail. Because our modeling attends to the fine details of the data, we are able to extract features from the frequency data that are material but have not been mentioned in the prior literature. An understanding of these features, such as the commonalities among states and severity levels measured by the number of affected state residents, is of considerable value to pricing, reserving, and general appreciation of the evolution of cyber experience. Our analysis may serve as an example of the kind of analysis that could be performed on cyber data elsewhere.

In addition, our modeling jointly analyzes the state attorneys general's publications of data breaches across the eight states that explicitly include reporting delay information (i.e., California, Delaware, Indiana, Maine, Montana, North Dakota,

Oregon, and Washington). This yields the most comprehensive comparison of cyber risk frequency across states in the literature to the best of our knowledge. Though the current academic literature analyzes cyber events within the United States at a broad, aggregate level (Edwards, Hofmeyr, and Forrest 2016; Romanosky 2016; Wheatley, Maillart, and Sornette 2016; Eling and Loperfido 2017; Eling and Jung 2018; Xu et al. 2018; Eling and Wirfs 2019; Strupczewski 2019; Kesan and Zhang 2020; Palsson, Gudmundsson, and Shetty 2020; Poyraz et al. 2020; Bessy-Roland, Boumezoued, and Hillairet 2021; Farkas, Lopez, and Thomas 2021; Jung 2021; Sun, Xu, and Zhao 2021; Wheatley, Hofmann, and Sornette 2021; Eling et al. 2022; Liu, Li, and Daly 2022; Malavasi et al. 2022; Li and Mamon 2023; Lu, Zhang, and Zhu 2023; Shevchenko et al. 2023), it lacks a nuanced examination of these events at a more granular, state level. Although some academic research and industry reports explore global variations in cyber risks (e.g., Deloitte 2023; Bruce et al. 2024; Verizon 2024), there is a notable absence of insights into state-level differences within the United States.

Moreover, we further enrich our analysis with a comparative analysis of data breaches across different severity levels and states. This approach yields valuable insights into cyber risk frequency variations across severity levels. Indeed, a proper understanding of the disparities in trends between large and small data breaches is crucial for insurers to effectively manage risks, allocate resources, and optimize their operations to remain competitive in the insurance market.

The following findings underscore important trends in data breach reporting and frequency that have critical implications for the cyber insurance industry. Our findings reveal shifting reporting patterns and lengthening delays for data breaches across states and severities, challenging the constant delay assumption of the basic chain-ladder method. Relying on the chain-ladder method may result in underestimating IBNRs and breach frequency, suggesting that cyber insurers should test its validity and be prepared to adopt an approach that more accurately reflects the data.

Another important observation is that the frequency of data breaches remained relatively stable prior to 2020 but showed an upward trend post-2020 across severity levels and states. This lends support to the hypothesis that the frequency of cyber events changed after the onset of COVID-19 (see, e.g., Cyber Insurance Academy 2021; U.S. Government Accountability Office 2021).

Finally, we summarize our results in an extensive and holistic discussion of their implications on the pricing, reserving, underwriting, capital needs, and experience monitoring of cyber insurance.

1.5. Structure of the Article

The article is organized as follows. Section 2 distinguishes state attorneys general's datasets from the PRC dataset, discusses key data analysis considerations for cyber data, and emphasizes the role of government data in reliable cyber research and analysis. Section 3 outlines the selection and processing of state attorneys general's data before they are modeled. Section 4 describes the construction of the model used in this article. Section 5 provides details of the model output, including the discovery of data features not mentioned in prior literature that could provide valuable insights into cyber risks. Section 6 highlights key research findings of this article for academic researchers and cyber insurers, including a detailed discussion of the insurance implications of the main model output found in Section 5. Section 7 concludes.

2. STATE ATTORNEYS GENERAL'S PUBLICATIONS OF DATA BREACHES

This section contains three parts. First, we introduce the datasets of state attorneys general in Section 2.1. Because they have not yet been sufficiently acknowledged by the literature, we offer some background for cyber risk researchers to understand the differences among the datasets of individual state attorneys general. State attorneys general's publications should serve as important data sources in the current climate of scarce public data on cyber events, because they constitute one of the primary sources of major existing datasets and provide valuable information.

Second, in Section 2.2, we compare state attorneys general's datasets with the widely used PRC dataset to help researchers better utilize both sources for insights into data breach risks. State attorneys general's data offer several advantages over the PRC dataset: they provide more timely and detailed information on breaches (see Section 2.2.1) and better capture interdependence among businesses using common services (see Section 2.2.2). This is crucial for understanding the implications of dependent cyber policies, a key challenge for market growth. Additionally, with a more constant reporting propensity, the frequency trend of reported breaches more accurately reflects actual occurrences (see Section 2.2.3). The reporting propensity refers to the ratio of breaches reported to those actually occurring.

Third, we emphasize that government data remain a critical resource owing to their unique strengths, including reliability and consistency derived from legal mandates and rigorous collection standards. This section highlights these advantages, noting how government data have been successfully used in fields such as public health to demonstrate its importance, with

similar approaches applicable to cybersecurity. Data from state attorneys general, in particular, provide accuracy, impartiality, and timeliness, making them essential for effective risk assessment, accurate pricing, threat analysis, and preparation for emerging data breach risks. Though commercial datasets like Advisen can offer additional depth, the foundational consistency of government data supports our analysis.

2.1. Introduction of Public Datasets Underrepresented in the Current Literature

In this article, we use data breaches published by individual state attorneys general in the United States to investigate changes in data breach reporting patterns and frequency trends. These data breaches are subject to state reporting guidelines, and they are publicly accessible on the websites of state attorneys general (see the references that follow).

As of June 2023, 17 state attorneys general publicly publish data breaches that they collect. See Maryland Attorney General (2023); Montana Department of Justice (2023); Oklahoma Office of Management & Enterprise Services (2023); California Attorney General (2024); Delaware Attorney General (2024); Hawaii Department of Commerce and Consumer Affairs (2024); Indiana Attorney General (2024); Iowa Attorney General (2024); Maine Attorney General (2024); New Hampshire Department of Justice (2024); New Jersey Cybersecurity & Communications Intergration Cell (2024); North Dakota Attorney General (2024); Oregon Department of Justice (2024); Texas Attorney General (2024); Vermont Attorney General (2024); Washington State Office of the Attorney General (2024); Wisconsin Department of Agriculture, Trade and Consumer Protection (2024).

2.1.1. Inconsistent Data Breach Notification Laws across States and over Time in the United States

At this time, the protection of private information in the United States is provided by a mixture of sector-specific federal statutes (i.e., covering financial services, health care, telecommunication, and education) and state laws, which differ in their scope and jurisdiction (International Comparative Legal Guides 2024). The National Conference of State Legislatures (NCSL) publishes state data breach notification laws in the United States (National Conference of State Legislatures 2024). As of June 2023, data breach notification laws have been implemented in all 50 states, the District of Columbia, Guam, Puerto Rico, and the Virgin Islands.

States in the United States imposed their own data breach notification laws at different times, each of which protects the privacy of its residents (National Conference of State Legislatures 2024). For instance, California has had a data breach notification law in effect since 2002, Mississippi since 2010, and South Dakota and Alabama since 2018, the last two states to enact such legislation.

Though most state data breach notification laws have similar elements, there are still variations. Key disparities include definition of what constitutes personally identifiable information, whom entities must notify, the number of affected state residents above which notification to the state attorney general becomes mandatory, and when the notification must be made once an obligation is triggered. The content of the breach notice, whether the state publishes breach data publicly, and any exemptions from reporting also vary across states (Privacy Rights Clearinghouse 2023). For example, Table 1 shows the varying definitions of reportable data breaches to state attorneys general in the United States in 2021 (International Association of Privacy Professionals 2021). The same breach might require notification of multiple state attorneys general, if it affects residents from multiple states.

Additionally, state laws are amended on a regular basis. Common trends include expanding the number of data items that constitute personally identifiable information, reducing reporting time frame, and requiring notification of the state attorney

TABLE 1
Definitions of Reportable Data Breaches in the United States in 2021

Notification to state attorney general	Number of states[a]
No obligations	17
Yes	14
Yes if more than 250 state residents	4
Yes if more than 500 state residents	8
Yes if more than 1000 state residents	7
Others	4

Note: [a]Including 50 states, the District of Columbia, Guam, Puerto Rico, and the Virgin Islands

general. According to Maine Legislature (2019), an additional requirement of the state attorney general effective from September 19, 2019, is to report breaches no later than 30 days after their discovery. Table 5 presents the dates when some states' statutes began to require notification of the attorney general.

2.1.2. Differing Fields Contained in the Datasets of Individual State Attorneys General

Tables 2 and 3 distinguish the information provided by individual state attorneys general on data breaches, as of June 2023. Attorneys general of New Hampshire, New Jersey, Vermont, and Wisconsin also publish breach information, but the breaches are presented only by individual notice letters. The second column of Table 2 lists the notification requirement of the state attorney general. The fourth and fifth columns of Table 3 present the requirements of maximum notification time frames from the discovery of a breach. The last column of Table 3 references state statutes.

2.2. Advantages of Datasets from State Attorneys General over the PRC Database on Frequency Modeling

The main dataset used in data breach frequency and severity modeling is the PRC dataset. The PRC dataset obtains most of its data from state attorneys general and the U.S. Department of Health and Human Services; the former are the data sources of this article and the latter collects breaches of protected health information under a federal regulation (i.e., The Health Insurance Portability and Accountability Act of 1996), which is also publicly available (U.S. Department of Health and Human Services 2024).

Data from state attorneys general are more suitable for the purpose of this study than the PRC dataset due to the considerations in the following section.

2.2.1. Additional Information Provided by State Attorneys General

Dates of occurrence As major sources of the PRC dataset, data breaches publicized by state attorneys general contain the information necessary for assessing reporting delay—date of breach incidence—that is absent from the PRC dataset. Shown in Table 2, among all states that publicly publish data breaches, eight explicitly present information regarding dates of occurrence (i.e., California, Delaware, Indiana, Maine, Montana, North Dakota, Oregon, and Washington). Maine provides two additional dates, including date of discovery and date of consumer notification.

Data breach notification letters Some state attorneys general provide access to data breach notification letters that are submitted by organizations as part of regulatory requirements. These letters typically contain descriptions of the breaches that have occurred.

Up-to-date breaches State attorneys general are updating their own database daily to include the newest reported data breaches, making the examination of breaches that occurred in 2020 and 2021 possible. However, the PRC dataset contains few data breaches that are reported after 2018 and none after 2019 and thus it is difficult to see the most recent changes, including those affected by the COVID-19 pandemic. Wheatley, Hofmann, and Sornette (2021) found that the number of data breaches in 2018–2019 present in the PRC dataset is far less than those in previous years and suspect incomplete data. Li and Mamon (2023) suggested that the PRC dataset is only reliable until 2017.

2.2.2. Richer Description of Data Breaches

Two different event definitions could be used One of the peculiarities of cyber risks is that event definition can be complicated by interdependencies among certain security incidents (Wheatley, Hofmann, and Sornette 2021), namely, third-party cyber events. A **third-party data breach** refers to a data breach that occurs at a service provider, vendor, or other third-party organization that has access to other companies' data (Prevalent 2024). It can be considered as (1) a single event that occurred at the third-party provider or one of its affected client firms or (2) a series of correlated events at the provider and all of its affected client firms. The choice of event definition will impact the derived frequency trend, because the latter will result in a greater number of data breaches.

A practical example of a third-party data breach is the SolarWinds breach, which occurred in 2020 when attackers infiltrated the company's software supply chain, injecting malicious code into a software update (Oladimeji and Kerner 2023). SolarWinds is a major IT management company that provides network monitoring tools to thousands of clients. This update was distributed to thousands of customers, including government agencies and private companies, leading to a widespread compromise. This incident illustrates how the failure or compromise of a third-party provider can be considered either a single event at the provider (i.e., SolarWinds) or multiple correlated events across all affected client organizations (i.e., public and private organizations). Depending on how the event is defined, it could be recorded as a single breach at the provider or as multiple incidents across affected organizations, which would affect the overall breach count.

TABLE 2
Summary of Datasets from State Attorneys General—1

State	Notification to state attorney general	Breach notice	Organization Name	Start of breach date	End of breach date	Reported date	Number of persons affected (state)	Number of persons affected (total)
California	Yes if more than 500 California residents	Y	Y	Y	Y	Y		
Delaware	Yes if 500 Delaware residents	Y	Y	Y	Y	Y	Y	
Hawaii	Yes if 1000 Hawaii residents	Y	Y	Y	Y	Y	Y	
Indiana	Yes		Y	Y		Y	Y	Y
Iowa	Yes if 500 Iowa residents	Y	Y			Y		
Maine	Yes	Y after 2020	Y	Y	Y	Y	Y	Y after 2018
Maryland	Yes	Y	Y			Y	Y	
Massachusetts	Yes		Y			Y	Y	
Montana	Yes	Y	Y	Y	Y	Y	Y	
North Dakota	Yes if 250 North Dakota residents	Y	Y	Y	Y	Y	Y	
Oregon	Yes if 250 Oregon residents		Y	Y	Y	Y		
Texas	Yes if 250 Texas residents		Y			Y	Y	
Washington	Yes if 500 Washington residents	Y	Y	Y	Y	Y	Y	

TABLE 3
Summary of Datasets from State Attorneys General—2

State	Cause of breach	Information breached	Time frame of notification to individuals (law enforcement exceptions)	Time frame of notification to state attorney general	Breach notification statutes
California			As expediently as possible (AEAP) and "without unreasonable delay" [a]		Cal. Civ. Code 1798.82 et seq.
Delaware			"Without unreasonable delay but no later than 60 days after determination of the breach of security" [b]	"No later than the time when notice is provided to the resident" [b]	Del. Code Ann. tit. 6 § 12B-101 et seq.
Hawaii	Y		"Without unreasonable delay" [c]	"Without unreasonable delay" [c]	Haw. Rev. Stat. § 487N-1 et seq.
Indiana			"Without unreasonable delay" [d]	"Without unreasonable delay" [d]	Ind. Code § 24-4.9-1-1 et seq.
Iowa			AEAP and "without unreasonable delay" [e]	"Within 5 business days after giving notice of the breach of security to consumers" [e]	Iowa Code § 715C.1 – 2
Maine		Y	AEAP and "without unreasonable delay, no more than 30 days after awareness of a breach of security and identification of its scope" [f]	"Without unreasonable delay" [f]	Me. Rev. Stat. tit. 10 § 1346 et seq.
Maryland	Y	Y	"As soon as reasonably practicable but no later than 45 days after the business discovers or is notified of the breach of the security of a system" [g]	"Prior to notification to individuals" [g]	Md. Code Com. Law § 14-3504 et seq.
Massachusetts		Y	"As soon as practicable and without unreasonable delay" [h]	"Without unreasonable delay" [h]	Mass. Gen. Laws 93H § 1 et seq.
Montana			AEAP and "Without unreasonable delay" [i]	"Simultaneous with notification to individual" [i]	Mont. Code § 30-14-1701 et seq.
North Dakota			AEAP and "without unreasonable delay" [j]	"Without unreasonable delay" [j]	N.D. Cent. Code § 51-30-01 et seq.
Oregon			AEAP and "without unreasonable delay, but no later than 45 days after discovering or receiving notification of the breach of security" [k]	"Without unreasonable delay, but no later than 45 days after …" [k]	Or. Rev. Stat. §§ 646A.600 - 646A.604

(*Continued*)

TABLE 3
(*Continued*).

State	Cause of breach	Information breached	Time frame of notification to individuals (law enforcement exceptions)	Time frame of notification to state attorney general	Breach notification statutes
Texas		Y	"Without unreasonable delay and in each case no later than the 60th day after the date on which the person determines that the breach occurred" [l]	"No later than the 60th day after … " [l]	Tex. Bus. & Com. Code § 521.053
Washington		Y	AEAP and "without unreasonable delay, and no more than 30 calendar days after the breach was discovered" [m]	"No more than 30 days after the breach was discovered" [m]	Wash. Rev. Code § 19.255.010 et seq.

Sources: [a]California Legislature (2024); [b]Delaware Legislature (2024); [c]Hawaii Legislature (2024); [d]Indiana Legislature (2022); [e]Iowa Legislature (2024); [f]Maine Legislature (2024); [g]Maryland Legislature (2024); [h]Massachusetts Legislature (2024); [i]Montana Legislature (2024); [j]North Dakota Legislature (2024); [k]Oregon Legislature (2024); [l]Texas Legislature (2024); [m]Washington Legislature (2024)

In the case of a third-party breach, the PRC dataset follows the former event definition (Benaroch 2021) and datasets of state attorneys general follow the latter. For example, In 2017, Sabre, a travel company, experienced a data breach in its Hospitality Solutions system that affected its business partners who used its central reservations booking engine (Fortra 2021). In the PRC dataset, this data breach is recorded as a single event under Sabre. However, state attorneys general's datasets classify it as multiple breaches involving various organizations, including Sabre. This is because the affected companies that outsourced services to Sabre were obligated to report the breach to relevant state attorneys general individually.

Which event definition should be used to count the number of events According to Wheatley, Hofmann, and Sornette (2021), taking into account all affected organizations in the case of a third-party cyber event provides a richer description of the event. Also, from both an economic and a cyber insurer's point of view, a third-party data breach should be regarded as a series of events rather than a single occurrence, which will be explained below.

To avoid underestimating the risk of a data breach, a third-party data breach should be viewed as a series of events that occurred at both the third-party provider and all of its affected client firms. It should not be considered as a single event that occurred at the third-party provider or one of its affected client firms. If not, we would underestimate the total number of businesses that are affected by such a breach, resulting in an overall underestimation of the frequency rate. Second, we would underestimate the economic impact of the third-party data breach, by failing to account for the impact on each affected business. Third, we would underestimate the dependencies across organizations because the data do not capture the dependence resulting from the utilization of common services and providers.

From the point of view of cyber insurance pricing, considering all affected businesses in the case of a third-party data breach has significant value in understanding the consequences of dependent cyber policies, which is the major impediment to the market's expansion. A cyber liability policy covers financial loss in the event of a data breach, irrespective of who was accountable for the loss of data (Woodruff Sawyer 2020). Therefore, when the insurance company insures a third-party provider and its client firms, a breach that occurred at the provider may result in multiple claims to the insurer, rather than a single claim, which can be an aggregation problem.

When the data of an organization that has purchased cyber insurance is compromised within a third party's system, the insurer will incur two kinds of costs. First, the organization is responsible for the costs related to the data breach, including regulatory compliance, potential litigation, and related costs. When the contract with the third-party vendor is not enough to cover these costs, the insurer is responsible for the rest (Woodruff Sawyer 2020). Second, the insurer may take the lead in executing the organization's contractual rights with the service provider directly accountable for the breach. This is known as subrogation (Woodruff Sawyer 2020).

Third-party data breaches require particular attention, because the cloud has become a ubiquitous part of corporate IT networks, and most breaches are found to be caused by third-party vendors. In 2022, 94% of businesses used cloud services in some capacity to hold and process their data (Flexera 2022). Ponemon Institute (2022) found that among 1162 cybersecurity experts surveyed, a majority of the them, accounting for 59%, acknowledged that their organizations had encountered a data breach originating from third parties.

In addition, the current literature has not investigated trends in data breach frequency that attempt to take into account all affected client firms of the third-party vendor in the case of a third-party data breach. Therefore, we use datasets of state attorneys general in our frequency analysis.

Which event definition should be used to assess frequency trends When a third-party data breach is regarded as a single event, referred to as Definition 1, changes in the frequency trend reflect shifts in the occurrences of primary events (e.g., the initial breaches at third-party service providers). This definition focuses solely on the breaches occurring at the source, with any observed fluctuations in the frequency trend directly indicating changes in the rate of primary events.

In contrast, when a third-party data breach is regarded as a series of events, referred to as Definition 2, changes in the frequency trend may reflect not only the occurrences of primary events but also the frequency of secondary events triggered by the primary breach. In this scenario, the breach frequency may increase owing to a ripple effect, where the primary breach impacts multiple client organizations. As a result, fluctuations in the frequency trend could stem from either changes in the number of initial breaches or the extent to which those breaches propagate across affected clients.

There are four possible scenarios for changes in frequency trends, each reflecting variations in primary and secondary events.

1. Scenario 1: Increase in Definition 1, constant/decrease in Definition 2
   - Primary events increase but fewer secondary entities are affected. Frequency rises owing to more primary breaches, even though each breach impacts fewer clients.
2. Scenario 2: Increase in Definition 1, increase in Definition 2
   - Both primary events and secondary entities increase, leading to a compounded rise in frequency, as breaches become more frequent and widespread.
3. Scenario 3: Constant/decrease in Definition 1, increase in Definition 2
   - Primary events remain constant or decline, whereas more secondary entities are affected. Frequency increases despite steady or fewer primary breaches, as each breach impacts more clients.
4. Scenario 4: Constant/decrease in Definition 1, constant/decrease in Definition 2
   - Both primary events and secondary entities remain stable or decrease, resulting in a constant or declining frequency trend, with fewer widespread breaches.

By considering both definitions and identifying the scenario at play, insurers can better capture the complete picture of data breach frequency, including both primary events and their ripple effects on associated organizations. Frequency increases may indicate growth in primary breaches, their spread across client organizations, or both. This broader perspective helps insurers assess whether primary breaches are rising and whether their systemic impact on associated businesses is expanding.

The approach to pricing, reserving, and risk assessment will differ for each scenario. Increases in primary events require insurers to address the increasing exposure and susceptibility of organizations to cyber risk, whereas increases in cascading effects (secondary events) call for a focused evaluation of supply chain risks and interdependencies among organizations. Recognizing these changing trends enables insurers to fine-tune their models for pricing, reserving, and risk assessment, ensuring that they are prepared for the increasing interconnectedness and complexity of cyber risks.

2.2.3. A More Constant Reporting Propensity

Some assumption about the reporting propensity is required to assess risks over time The ultimate goal of modeling the frequency and severity of cyber risks in the context of cyber insurance is to estimate the loss distribution of insurance claims resulting from cyber incidents. Trend analysis of events can shed light on the risks associated with insurance claims over time. However, the limitation of any datasets that collect real events is that they can only record those that are publicly disclosed; not all events that have occurred are known.

Nonetheless, we can learn about the trend of actual events by analyzing reported ones, as long as we can make a valid assumption about the reporting propensity (see the definition in Section 2). For example, if the reporting propensity remains

constant over occurrence periods, counts of reported events will vary proportionately with those incurred, and the former will validly reflect any trend in the latter.

The data breach reporting propensity in the United States may have shifted over time, complicating efforts to identify actual frequency trends with reported incidents. From this perspective, individual state attorneys general's datasets are more appropriate for frequency analysis because they provide a more reliable basis for assuming a constant reporting propensity compared to the PRC dataset. Detailed discussion is presented next.

A likely change in the reporting propensity in the United States The reporting propensity of data breaches is heavily influenced by legal requirements, because organizations are reluctant to disclose their security incidents unless necessary, in the fear that they could tarnish the reputation of the brand and instill mistrust among customers. As a result, notification laws play a crucial role in the disclosure of security incidents, and the enactment and amendment of such mandates could have a substantial effect on the events that come to light. For example, in Australia, the total number of data breach notifications increased by 712% under the mandatory reporting scheme (i.e., the Notifiable Data Breaches scheme) compared to the previous year under the voluntary scheme (Office of the Australian Information Commissioner 2019).

The propensity to report data breaches with varying severities in the United States may have changed over time as a result of differing state laws governing data breach notification obligations and the evolution of such requirements over time within individual states. We have identified two key disparities in state laws that materially affect the reporting propensity.

First, over time, data breach notification regulations have evolved to require breached organizations to notify not only affected customers but also government bodies, which may have resulted in a greater number of breaches being made public following this change. For example, since January 1, 2012, California has required notification of the California attorney general regarding breaches that affect more than 500 California residents. Considering that a sizable percentage of the PRC dataset comes from state attorneys general, including the California attorney general, this requirement may have led to more breaches being collected by the California attorney general and subsequently by the PRC dataset. In addition, notification of the respective state attorney general regarding certain breaches is enacted at different times across different states' statutes (see Table 5). For example, Washington State added this requirement at a much later date than California, on July 24, 2015. Such variation in reporting requirements over time across different states could potentially lead to changes in the reporting propensity of data breaches in general at the national level.

Second, the varying definitions of reportable data breaches across states may have also resulted in differing reporting propensity of data breaches with different severities. The number of affected state residents above which notification to the respective state attorney general becomes mandatory varies by state (see Table 1). For example, the Washington attorney general requires the reporting of data breaches that affect more than 500 Washington residents, whereas the Indiana attorney general requires the reporting of all breaches that affect Indiana residents (see Table 2). Because organizations are generally reluctant to disclose data breaches unless they are legally required to do so, the breaches reported to the Washington attorney general will be larger in size than those reported to the Indiana attorney general. This is reflected in the actual data: 7% of data breaches published by the Indiana attorney general affect more than 500 state residents, compared to 90% of data breaches published by the Washington attorney general.

The reporting propensity of the PRC dataset The PRC dataset relies heavily on information provided by individual state attorneys general. Hence, any change in the propensity to report to individual state attorneys general would likely cause a shift in the reporting of the PRC dataset. It is possible that the reporting propensity of individual state attorneys general may have varied throughout the PRC dataset's duration (2005–2018) owing to differences in when states implemented the reporting requirement to their respective attorneys general. This variability could result in the reporting propensity of the PRC dataset being less constant, and we could not make a valid assumption about its reporting propensity over time. For example, it became mandatory for organizations to notify the California attorney general of data breaches affecting more than 500 California residents at the beginning of 2012. As a result, more such breaches might have been captured by the California attorney general and subsequently by the PRC dataset. The California attorney general alone is the source of almost 10% of breaches in the PRC dataset, and thus the change in the reporting propensity of California can materially affect that of the PRC dataset.

The PRC dataset may not be suitable for performing national trend analysis due to variations in the definitions of reportable data breaches among its data sources, which may have resulted in differing reporting propensity for breaches with varying severities. Because various state attorneys general, one of the primary data sources for the PRC dataset, require the reporting of breaches with different severities, it can be challenging to derive meaningful insights from aggregated analysis across all states.

Additionally, as noted by Li and Mamon (2023), concerns about data collection reliability have led to suspicions that the PRC dataset's reporting propensity may have been subject to changes. In particular, there has been a noticeable decline in the number of incidents reported for nonmedical institutions after 2012, which seems inconsistent with the growing prevalence of e-commerce and increasing awareness of cyber risks.

Thus, we cannot reasonably conclude that the reporting propensity of data breaches in the PRC dataset is constant. Though the PRC dataset is useful for tracking the earliest and biggest data breaches, the evolution of data breach risks can be difficult to unravel, given that we cannot make a reliable assumption about its reporting propensity over time.

The reporting propensity of datasets provided by state attorneys general The datasets from individual state attorneys general, when analyzed individually, are likely to be subject to a more constant reporting propensity than the PRC dataset. First, these are data breaches that were reported following mandatory notification of individual state attorneys general (i.e., after the major shift in the reporting propensity within individual states). Second, the definition of reportable data breaches remains consistent over time in each state. Although the reporting propensity in any state could still change due to changes in legal environments (for instance, stronger penalties could increase the reporting propensity), we could assume with greater confidence that the reporting propensity of individual state attorneys general's datasets is constant than we could for the PRC dataset.

2.3. The role of government data in cyber research and analysis

Using government data in cybersecurity is both practical and necessary, offering a solid foundation for addressing cyber threats in today's data-driven world. This section highlights its proven value across fields like public health, environmental science, and social sciences, underscoring its reliability and potential as a key asset for analyzing threats and mitigating risks.

Government data stand out for their reliability, consistency, and authority (Washington 2014). Collected through documented procedures and legal mandates, government data ensure quality, transparency, and sustainability, making government data a stable, dependable resource. These standards provide a strong basis for big data algorithms, predictive models, and trend analyses.

Government data have been a crucial benchmark across sectors. For example, public health data from agencies like the Centers for Disease Control and Prevention have long been used to model disease trends, aiding in crisis preparedness (Washington 2014), and postal codes, initially for mail delivery, now support demographic analysis and socioeconomic clustering. These uses show government data's foundational role in large-scale, reliable analytics, demonstrating their relevance across disciplines.

The impact of government data has grown with the introduction of government open data initiatives, which has broadened its accessibility and applications. Researchers across fields, including medicine, environmental science, and social sciences, rely on government open data for diverse purposes: as data sources in statistical models, benchmarks for validating existing datasets, and context for comprehensive studies (Yan and Weber 2018). This broad usage highlights their versatility, reinforcing their value as a credible, standardized resource for specialized and interdisciplinary research.

In cyber insurance, government data offer substantial benefits, providing accurate, unbiased, and timely information essential for effective risk assessment and pricing. For instance, breach data from U.S. state attorneys general, collected under mandatory notification laws, ensure transparency, accuracy, and impartiality (U.S. General Services Administration 2018; USAFacts 2023). This consistent, dependable view of breaches is ideal for trend analysis and cross-validation with other data sources. With breach details typically available online within days of submission, cybersecurity professionals gain timely insights into emerging threats, a crucial advantage in a rapidly evolving cyber landscape to help mitigate financial losses.

Leveraging government data provides cybersecurity professionals with a trusted resource for precise threat analysis and effective risk mitigation. The established value of government data across fields like public health and environmental science demonstrates their adaptability, making government data an essential asset for addressing today's complex cyber challenges.

3. DATA SELECTION AND PROCESSING

In this section, we cover the specific details of data selection, aggregation, and processing related to the state attorneys general's data. This includes the identification of data segments for analysis, the definition of reporting delay, the selection of time periods for investigation, the choice of frequency aggregation, and any required data cleaning. These measures are necessary to eliminate potential biases and make the conclusions of this article more relevant to cyber insurers. A summary of the data manipulations can be found in the Online Appendix A.

3.1. Differentiating Data Breaches by State and Severity

In the previous section, we saw that to control for the reporting propensity, we should analyze datasets from individual state attorneys general separately. When comparing across states, we should compare breaches under the same definition, because different states may not share the same definition of reportable breaches.

Therefore, we categorize data breaches by state and definition. All cases of comparison are shown in Table 4; 15 data segments are investigated, consisting of eight states and four severities. First, we study reporting delays of all eight states that provide information on both reported date/date of notification and date of breach (occurrence date). We exclude breaches with an unknown date of occurrence from the analysis. Second, we differentiate among data breaches with various severities (i.e., the number of state residents affected), because these states collect data breaches that are subject to different severities.

3.2. Definition of Reporting Delay

The reporting delay consists of the time lag between breach occurrence and discovery and the time lag between discovery and notification of relevant parties. Both lags are of significant interest, but we cannot study them separately because only Maine has 3 years of breaches with dates of discovery. Therefore, we study the lag between breach occurrence and reporting. We consider the occurrence date to be the earliest possible date when the breach might have occurred, because this determines coverage or not for most insurance contracts. The date of notification is also defined as the earliest date, because this approximates when insurers receive claims. Therefore, when multiple dates are present, only the earliest date is retained.

We assume that the similarity in urgency requirements between insurers and state attorneys general ensures that the reporting delay to state attorneys general should reasonably approximate the delay to insurers. Legally mandated deadlines, such as requirements to report as expediently as possible or Maine's 30-day post-discovery rule, often align with insurance policies mandating prompt notification. This approximation captures the minimum lag period, though some variability may arise owing to internal reviews or legal consultations before notifying insurers.

3.3. Selection of Time Periods for Investigation

We analyze the time periods for which complete and unbiased data are available (see Table 5) to ensure that we do not underestimate the number of breaches that occurred in earlier years. First, we exclude breaches that occurred before notification of the state attorney general was made mandatory. If we included them, we may have significantly underestimated the number of breaches that occurred prior to the mandatory notification requirement, and we may have wrongly identified an increasing frequency trend (see Section 2.2.3).

Second, we also exclude breaches that occurred after the mandatory notification requirement but before the earliest reported date of all breaches in the database. For example, the mandatory notification requirement of the North Dakota attorney general was effective from April 13, 2015. However, all breaches in the database were reported after January 2, 2019. If we included breaches that occurred prior to 2019, we would again be at risk of underestimating the number of data breaches.

3.4. Choice of Frequency Aggregation

A critical decision that has to be made is around frequency aggregation: annually, quarterly, or monthly? This depends on how often cyber insurers should monitor the reporting delay. Because the United States holds the largest market size of cyber insurance, it might be worth looking at the regulations of the property and casualty insurance industry in the United States, which cyber insurance falls under.

Regulation of the insurance industry in the United States is mainly executed by the respective states, with state insurance regulators being members of the National Association of Insurance Commissioners (NAIC). Regulatory filings of insurance companies consist of those required by the NAIC, which are identical for all, and those required by the state where the insurers

TABLE 4
Comparison by State and Severity

Number of state residents affected	State
0–249	Indiana, Montana, Maine
250–499	Indiana, Montana, Maine, North Dakota
$\geq$250	Oregon
$\geq$500	Indiana, Montana, Maine, North Dakota, Washington, Delaware, California

TABLE 5
Eight States with Recorded Dates of Breach Occurrence

State	Notification to state attorney general (effective date)	Earliest reported date	Period of analysis (date of occurrence)	Accident quarters
California (CA)	January 1, 2012	January 20, 2012	January 1, 2012, to December 31, 2021	2012Q1–2021Q4
Delaware (DE)	April 14, 2018	April 11, 2018	April 1, 2018, to December 31, 2021	2018Q2–2021Q4
Indiana (IN)	2006 (exact date unknown)	January 2, 2014 (only a few in 2013)	January 1, 2014, to July 31, 2021	2014Q1–2021Q2
Maine (ME)	2005 (exact date unknown)	January 2, 2013 (only a few before then)	January 1, 2013, to July 31, 2020	2013Q1–2020Q2
Montana (MT)	October 1, 2015	October 1, 2015 (only a few before then)	October 1, 2015, to December 31, 2021	2015Q4–2021Q4
North Dakota (ND)	April 13, 2015	January 2, 2019	January 1, 2019, to December 31, 2021	2019Q1–2021Q4
Oregon (OR)	January 1, 2016	January 14, 2016 (only two before then)	January 1, 2016, to December 31, 2021	2016Q1–2021Q4
Washington (WA)	July 24, 2015	August 11, 2015	October 1, 2015, to December 31, 2021	2015Q4–2021Q4

are admitted to do business, to the NAIC Financial Data Repository (National Association of Insurance Commissioners 2024a). Quarterly financial statements are one of the sets of financial statements that need to be completed in accordance with the NAIC (National Association of Insurance Commissioners 2024b). Given the need for cyber insurers to quantify their liabilities quarterly, monitoring reporting delay on a quarterly basis is necessary.

3.5. Necessary Data Cleaning

Two features of data breaches disclosed by state attorneys general are worth noting: recording errors occur periodically, and sometimes officers record supplementary breach notices as separate entries to update the information contained in the original breach notice.

To fix the first, we retrieve correct dates from breach notices for breaches with a negative delay (i.e., occurrence dates later than discovery/notification or discovery dates later than notification). If the breach notice does not contain the correct dates or is unavailable, we remove the breach. In the cases of errors that are not obvious (e.g., notification lag of 1 day), they have been accepted as correct because there is no obvious means of filtering them.

Second, we should handle notices/entries related to the same breach with care to avoid double-counting. After transferring the updated information from the supplementary notices to the original notice, we delete entries related to supplementary notices. In addition, if a breach only contains dates when supplementary notices are submitted but not when the original notice is submitted (one breach in Maine met this criterion), we exclude it from the analysis to avoid overestimating the reporting delay. This is because supplementary notices are submitted after the original notice, resulting in a longer delay between the occurrence date and the reported date for supplementary notices compared to original notices.

4. A MODEL OF DATA BREACH REPORTING PATTERNS AND FREQUENCY

Quarterly data breach development patterns have received little attention in the cyber risk literature, with inadequate consideration of changing reporting delays when estimating IBNR breaches (e.g., Kapoor and Nazareth 2013; Wheatley, Hofmann, and Sornette 2021; Sangari, Dallal, and Whitman 2022; Eling, Ibragimov, and Ning 2023).

In this article, we address these gaps by exploring changes in quarterly data breach development patterns and reporting delays over time, enabling a more accurate estimation of IBNRs and frequency of data breaches. To achieve this, we apply parameter reduction techniques to an ODP cross-classified model, resulting in a GAM. Where appropriate, categorical variables

in the generalized linear model (GLM) are replaced with semiparametric forms, such as splines, to enhance parameter efficiency (Taylor and McGuire 2016). We first formulate a GAM for each state and severity and then fuse these various models into one by taking advantage of the commonalities across states and severities.

We choose GAMs owing to their flexibility and long-standing use in actuarial research, particularly in claims reserving; see a review of claims reserving models in Chang, Gao, and Shi (2023). Compared to traditional chain-ladder models and their GLM representations, which require many parameters and struggle with extrapolating tail factors, GAMs incorporate nonparametric smoothing to capture nonlinear trends more effectively. This allows for smoother modeling of accident and development period effects, reducing parameter complexity. Additionally, GAMs nest GLMs as a special case, offering a broader and more adaptable framework for modeling claims frequency, while preserving the core structure of traditional models. This makes GAMs a well-established and improved alternative in actuarial literature.

We begin with a brief description of the ODP cross-classified model (see Section 4.1). In Section 4.2, we provide an overview of the prototype model that serves as the foundation for this study. Next, we present our proposed model equation outlining the components of the GAM, accompanied by relevant examples.

The data used in the model are run-off triangles, which will be explained shortly and can be found in Online Appendix B. The GAM for all states and severities is provided in Online Appendix C. Model diagnostics is provided in Online Appendix D.

4.1. Preliminary: ODP Cross-Classified Model

The chain-ladder method and its extensions based on run-off triangles are widely used in IBNR reserve estimation (Kremer 1982; Mack 1993, 1994; Verrall 1994, 2000; England and Verrall 1998, 2001, 2002; Renshaw and Verrall 1998; Pinheiro, Andrade e Silva, and de Lourdes Centeno 2003; Antonio and Beirlant 2008; Wüthrich and Merz 2008; Taylor 2012; Grize 2015; Costa, Pizzinga, and Atherino 2016; Peremans et al. 2017; Shi 2017; Sriram and Shi 2021). They estimate IBNRs by completing the lower triangle of a run-off triangle using information from the upper triangle, which represents the experience to date.

A run-off triangle is a matrix of numbers that shows claim observations for each period, such as the number of claims filed, the amount paid out in claims, and the average cost of claims. The rows of the matrix represent the accident periods and the columns depict the development periods. Accident periods are the periods in which the claims occurred, and development periods are the periods in which the claims were reported, developed, and ultimately closed. The triangle has a third orientation, the diagonal, which is also known as the calendar period. Each diagonal represents claim experience during a particular calendar period.

The ODP cross-classified model assumes that the claim observations C_{ij} in row i and column j, the incremental reported claim counts in our case, are distributed as independent ODP random variables. Mean and variance are as follows:

$$\mathrm{E}[C_{ij}] = \mu_{ij} = x_i y_j \quad \text{and} \quad \mathrm{Var}[C_{ij}] = \phi x_i y_j,$$

where

$$\sum_{j=1}^{n} y_j = 1.$$

Here, x_i is the expected ultimate claims up to the latest development period n observed in the triangle for accident period i, and y_j is the proportion of ultimate claims to emerge in development period j. Overdispersion is introduced through the parameter ϕ, which is unknown and estimated from the data.

The maximum likelihood estimators are equivalent to the conventional chain-ladder estimators (Renshaw and Verrall 1998; England and Verrall 2002; Taylor and McGuire 2016).

This can be recognized as a GLM, with a log link

$$\ln(\mu_{ij}) = c + \alpha_i + \beta_j, \tag{1}$$

where c is the intercept, α_i is a parameter for each accident period i, β_j is a parameter for each development period j, and, conventionally, the corner constraint is imposed, such that $\alpha_1 = \beta_1 = 0$.

4.2. Model Construction

4.2.1. Prototype Model

The reporting pattern of data breaches is modeled by modifying the Hoerl curve (England and Verrall 2002), also known as the Gamma curve, which is the most popular parametric form to describe development patterns in general insurance. The Hoerl curve is produced by substituting the column parameters β_j in Equation (1) with

$$\ln(\mu_{ij}) = c \; + \; \alpha_i + \beta_i \cdot \ln(j) + \gamma_i \cdot j. \tag{2}$$

The development time j is considered as a continuous variable, which allows extrapolation outside the observed range of development times. The development pattern adheres to a predetermined parametric structure, and it is allowed to be different for each accident period. The Hoerl curve has a general shape that resembles the typical development pattern of incremental claims, with a steep increase to a peak followed by an asymptotically exponential decline.

4.2.2. Proposed Model Equation

The structure of our model combines several components to capture different aspects of the data. Beginning with the ODP cross-classified model, we first simplify the development and accident period effects and then incorporate calendar period effects. Next, we explore interactions between accident and development periods (i.e., shifts in the development pattern across accident periods), followed by the treatment of exceptional observations.

The general prototype equation for the model is as follows:

$$\ln(\mu_{ij}) = c + \sum_{k=1}^{K} \beta_k \cdot f_k(j) + \sum_{m=1}^{M} \alpha_m \cdot g_m(i) + \sum_{n=1}^{N} \phi_n \cdot 1_{\{c \in A_n\}} + \sum_{p=1}^{P} \theta_p \cdot f_p(j) \cdot g_p(i) + \sum_{q=1}^{Q} \psi_q \cdot 1_{A_q}(i,j) \tag{3}$$

The different terms in the model represent the following:

- **Development profile** $f_k(j)$: Common forms include $\ln(j+1)$, j, and $j \cdot \ln(j+1)$, which modify the Hoerl curve.
- **Accident period trend** $g_m(i)$: This could be linear as i, be quadratic as i^2, or modeled using linear splines such as $(i-\kappa)_+$, where κ is a knot point and $(x)_+ = \max(0,x)$.
- **Calendar period effects** $1_{\{c \in A_n\}}$: These use indicator functions for specific calendar periods A_n, where $c = i + j - 1$, modeling the effects that apply only during those periods.
- **Interaction terms** $f_p(j) \cdot g_p(i)$: Examples include $\ln(j+1) \cdot (i-\kappa)_+$ to capture interactions between development and accident periods, introducing a change in the development profile at $i = \kappa$.
- **Exceptional observations** $1_{A_q}(i,j)$: These indicator functions are used for specific cells A_q to handle outliers or exceptional cases.

5. ANALYSIS OF MODEL OUTPUT

In this section, we present the output of the GAM model built for the 15 data segments (i.e., eight states and four severities in Table 4). First, from Section 5.1 to 5.5, we present the most important results and their interpretation. Second, we summarize residual model effects that consist of interesting model features without a clear interpretation (see Section 5.6). Finally, we present the limitations of our analysis in Section 5.7.

Our analysis focuses on quarterly patterns of reported breaches over time, which we evaluate using run-off triangles. For brevity, we refer to accident quarter, development quarter, and calendar quarter as AQ, DQ, and CQ, respectively.

Data segments are represented by abbreviations such as IN(0–249), where the state is abbreviated (see Table 5) and the number inside the parentheses corresponds to the range of individuals affected by the breaches being investigated. As an example, the data segment abbreviation IN(0–249) refers to breaches that affect between 0 and 249 Indiana residents. Henceforth, larger breaches are referred to as those that affect more than 500 state residents, and smaller breaches are referred to as those that affect between 0 and 249 state residents. In addition, the numbering of AQs follows YYYYQQ format, provided in the last column of Table 5.

The analysis focuses on the four largest data segments IN(0–249), MT(0–249), ME(0–249), and CA(>499), with WA(>499) and OR(>249) following behind. The remaining data segments deserve less emphasis owing to their low average

number of notifications per cell in their respective quarterly run-off triangle, which is <5 (see Online Appendix B). This creates challenges in deriving meaningful insights from the data.

IBNR breaches are projected by extrapolating historical development profiles to the future to complete the lower triangle; the number of ultimate breaches incurred is calculated by adding actual counts in the upper triangle to the forecasts of IBNRs in the lower, which then reveals frequency trends.

5.1. General Development Profile

5.1.1. Cyber Insurance Is a Short-Tailed Business

The data breach notification component of cyber insurance is a short-tailed business. At least 80% of breaches are reported within a year of occurrence and 90% within a year and a half.

5.1.2. Larger Breaches Have Longer Delay between Occurrence and Notification than Smaller Breaches

On average, 80% of larger breaches are reported within a year of occurrence and 90% within a year and a half (see Panel B of Figure 1a). Ninety percent of smaller breaches are disclosed within a year of occurrence, and almost all breaches are reported within a year and a half (see Panel B of Figure 1b).

5.2. Accident Period Trend

5.2.1. The Inclusion of IBNRs Reveals Escalating Frequency

The inclusion of projected IBNRs leads to a notable rise in claim frequency across all states and severities beyond 2020, as opposed to the decrease in frequency that would be seen if actual counts were the only consideration. See Figure 2a for CA(>499) and Figure 2b for IN(0–249) and all other major cases in Online Appendix E.

5.2.2. CA, IN, MT, and ME Exhibit Highly Similar Frequency Trends

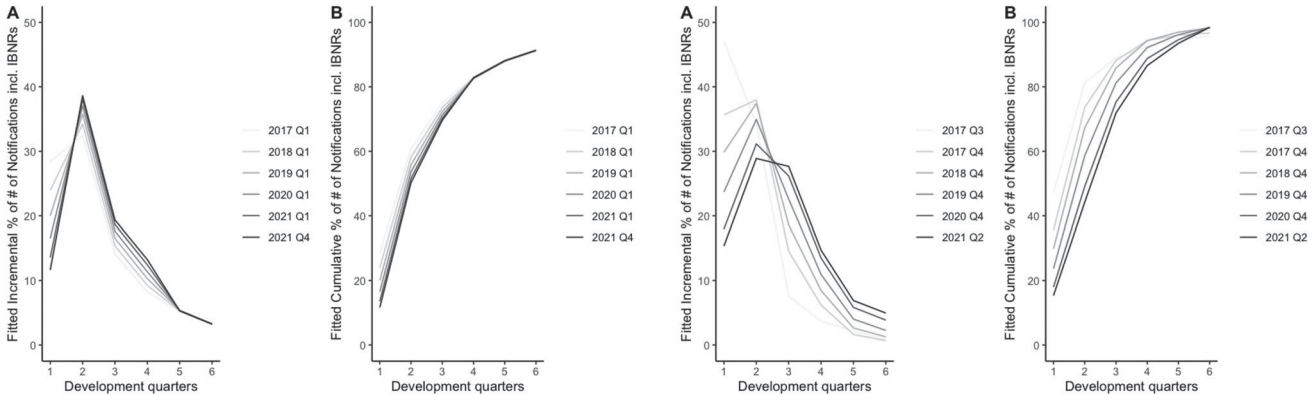
The growth of quarterly ultimate breaches incurred between 2016Q1 and 2021Q4 is similar across all four states when compared to the average quarterly number of breaches between 2015Q4 and 2016Q3. See Figure 3.

Starting from 2014Q2, the growth rates of four states often share the same sign for the same period. Furthermore, for the periods with the highest or lowest growth rates, except for 2018Q1, the growth rates consistently maintain both sign and magnitude across states. See Figure 4.

5.2.3. 2020Q1 Marks a Break Point in Frequency Trends for All States

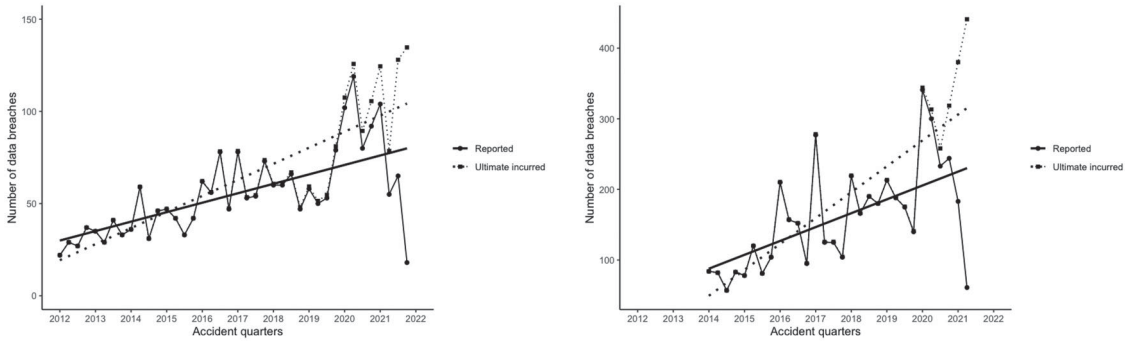
The full model is given in Equation (3), and the terms associated with the α parameters (i.e., frequency trends) are expanded in Equation (4).

$$\ln(\mu_{ij}) = \ \ldots + \ \alpha_1 \cdot i + \alpha_2 \cdot i^2 + \alpha_3 \cdot 1_{i \geq 2020Q1}(i) \tag{4}$$



(a) AQ 2017Q1–AQ 2021Q4, CA(>499) (b) AQ 2014Q1–AQ 2021Q2, IN (0–249)

FIGURE 1. Development Pattern Trends.



(a) AQ 2012Q1–AQ 2021Q4, CA(>499)

(b) AQ 2014Q1–AQ 2021Q2, IN (0–249)

FIGURE 2. Reported versus Ultimate Incurred Breaches.

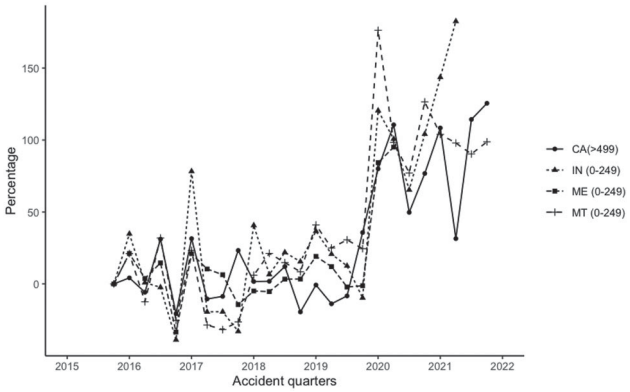


FIGURE 3. Ultimate Breaches Incurred by AQ, Expressed as a Percentage of the Average over AQs 2015Q4 to 2016Q3.

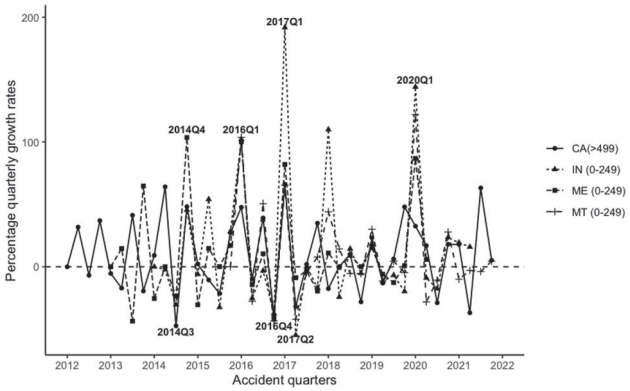


FIGURE 4. Quarterly Growth Rates of Ultimate Breaches Incurred (AQ 2012Q1–2021Q4).

$$+\alpha_4 \cdot 1_{2020Q1}(i) + \alpha_5 \cdot 1_{2020Q2}(i) + \alpha_6 \cdot \max(0, i - 2020Q2) \ + ...$$

α_3 is statistically significant in most states and severities, meaning that 2020Q1 is the break point in their frequency trends. In the following we will discuss the differences in the frequency trends across individual cases.

- **CA(>499), IN(0–249):** α_4 and α_5 are insignificant, all else significant in CA(>499). α_4 is insignificant in IN(0–249). There is a sudden increase in the quarterly number of breaches in 2020Q1, followed by a continuous upward trend.
- **MT(0–249), ME(0–249), WA(>499), OR(>250), IN(>499), MT(250–499), ME(250–499), ND(250–499):** α_4, α_5, α_6 are insignificant. Following the spike in 2020Q1, the quarterly number of breaches stabilizes at a relatively higher level.

- **ND(>499), IN(250–499):** α_5, α_6 are insignificant. After the sharp increase in 2020Q1, the quarterly count of breaches declines to a level below that of 2020Q1 but stays elevated compared with pre-2020Q1 periods.

5.2.4. Smaller Breaches Show Faster Growth and Greater Volatility Compared to Larger Breaches

IN(0–249) exhibits the highest and lowest growth rates among the four major data segments (see Figure 4).

To ensure meaningful comparisons among states with varying data availability, we calculate the mean and standard deviation of percentage quarterly growth rates based on the largest common time periods observed in two or more states. The resulting values are presented in Table 6. On average, IN(0–249) has the highest and most volatile growth rates, followed by ME(0–249) and MT(0–249) and then CA(>499).

5.3. Interactions between Development Periods and Accident Periods

5.3.1. All States Experience Shifts in Their Reporting Patterns, and the Reporting Patterns Vary Based on Breach Severity

Equations (5) and (6) show a subset of terms associated with the β and θ parameters in Equation (3) (i.e., reporting patterns across accident periods in different states).

$$\begin{aligned}\ln(\mu_{ij}) = \ &\ldots + \ \beta_1 \cdot \ln(j+1) + \beta_2 \cdot \ln(j+1) \cdot 1_{[1,4]}(j) \\ &+ \theta_1 \cdot \ln(j+1) \cdot 1_{[1,4]}(j) \cdot \max(0, i_1 - i) \ + \ldots\end{aligned} \tag{5}$$

$$\ln(\mu_{ij}) = \ \ldots + \ \beta_1 \cdot j + \beta_2 \cdot \max(0, j-6) \tag{6}$$

$$+\beta_3 \cdot 1_{\{1,2,5\}}(j) + \beta_4 \cdot 1_{\{2,5\}}(j) + \beta_5 \cdot 1_{\{5\}}(j) \tag{7}$$

$$+\theta_1 \cdot \max(0, j-6) \cdot \max(0, i - i_2) \ + \ldots$$

For larger breaches (see Equation (5)), in the case of CA(>499), the number of notifications declines as a power function of DQ, with different exponents before and after DQ 4. In addition, the before–DQ 4 exponent varies with AQ, indicating that the short-term notification profile has varied with AQ. See Figure 1a for the trends in the notification profile between 2017Q1 and 2021Q4. For detailed descriptions of the changes in trends across all AQs, see Online Appendix F.1.

For smaller breaches (see Equation (6)), IN(0–249), MT(0–249), and ME(0–249) share another type of reporting pattern. The number of notifications decreases exponentially with a change in decay factor at DQ 6 and corrections at DQs 1, 2, and 5. The decay factors used to model the notification profile also vary with AQ. See Figure 1b for the trend in the notification profile of IN(0–249). See Online Appendix F.2 to F.4 for descriptions of the changes in trends of IN(0–249), MT(0–249), and ME(0–249).

5.3.2. Around 2017, All States Observe a Shift in Their Reporting Patterns

After some point in 2017, all states observe a different trend in their reporting patterns. One of the two change points in the reporting pattern of CA(>499) is 2017Q1 (see Online Appendix F.1). IN(0–249), MT(0–249), and ME(0–249) experience relatively constant reporting patterns before and including 2017Q3 and shift away from them since 2017Q4 (see Online Appendix F.2 to F.4).

TABLE 6
Mean and Standard Deviation of Percentage Quarterly Growth Rates

	CA(>499)	ME(0–249)	IN(0–249)	MT(0–249)
2013Q2–2014Q1	3.47 (28.40)	2.51 (47.97)	NA	NA
2014Q2–2015Q4	8.90 (39.66)	11.62 (44.28)	8.09 (35.03)	NA
2016Q1–2020Q2	10.41 (30.41)	11.89 (38.27)	20.36 (68.48)	14.97 (46.64)
2020Q3–2021Q2	−7.46 (29.56)	NA	10.29 (18.86)	1.07 (18.24)
2021Q3–2021Q4	34.12 (40.85)	NA	NA	0.29 (5.87)

5.3.3. The Reporting Delay Has Been Getting Longer in Most States

Figures 5a and 5b and Online Appendix G compare the fitted average delay and its trend (i.e., assume no calendar period effects and exceptional observations) for the four major cases. For periods that have been assigned zero weight (see notes at the bottom of Online Table P of Online Appendix C), the figure compares the actual delay observed in the data and the trend.

The average delays of CA(>499), IN(0–249), and ME(0–249) have lengthened to different extents after 2017 compared to prior periods. The average delay of MT(0–249) has decreased slightly but not significantly compared to previous periods.

CA(>499) As shown in Figure 5a, the average time to report data breaches is 2.8 quarters in AQ 2012Q1, increases to 3.2 quarters in AQ 2014Q3, decreases slightly to 3.1 quarters in AQ 2017Q1, and increases again to 3.4 quarters in AQ 2021Q4. It takes 2 quarters to reach 65% of reporting in AQ 2012Q1 but takes 3 quarters in AQ 2021Q4 (see Online Appendix F.1).

IN(0–249) As shown in Figure 5b, the average time to report data breaches is 2.2 quarters between AQ 2014Q1 and AQ 2017Q3 and increases to 2.9 quarters in AQ 2021Q2. It takes 2 quarters to reach 80% of reporting between AQ 2014Q1 and AQ 2017Q3 but takes 4 quarters in AQ 2021Q2 (see Online Appendix F.2).

5.3.4. The Reporting of Data Breaches Shifts Away from the First Quarter of Occurrence across All States

The percentage of breaches that are reported within the quarter of their occurrence has decreased across all states, regardless of whether the average delay has improved or gotten worse (see Online Appendix F). On average, it has decreased from more than 30% in 2017 to less than 15% in 2021.

For example, MT(0–249) is the only state that has seen a slight improvement in the average reporting delay after 2017 (see Figure 6b). However, just like the other states with deteriorating reporting delay, the percentage of notifications received in the quarter of breach occurrence decreased from 35% in 2017Q3 to 12% in 2021Q4, as shown in Figure 6a.

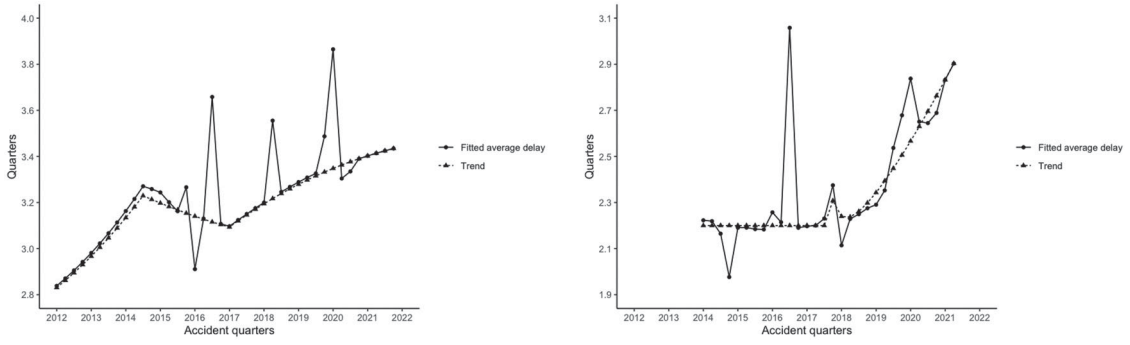
5.4. Exceptional Observations

5.4.1. Some Periods of Erratically Long Reporting Delay Are Shared across All States

During three periods (AQ 2016Q1, AQ 2016Q3, and AQ 2020Q1), all states have experienced significant delays in reporting breaches. These are isolated changes that are significantly off-trend. In addition to the four major cases, we also observe them in in WA(>499), OR(>249), and IN(>499).

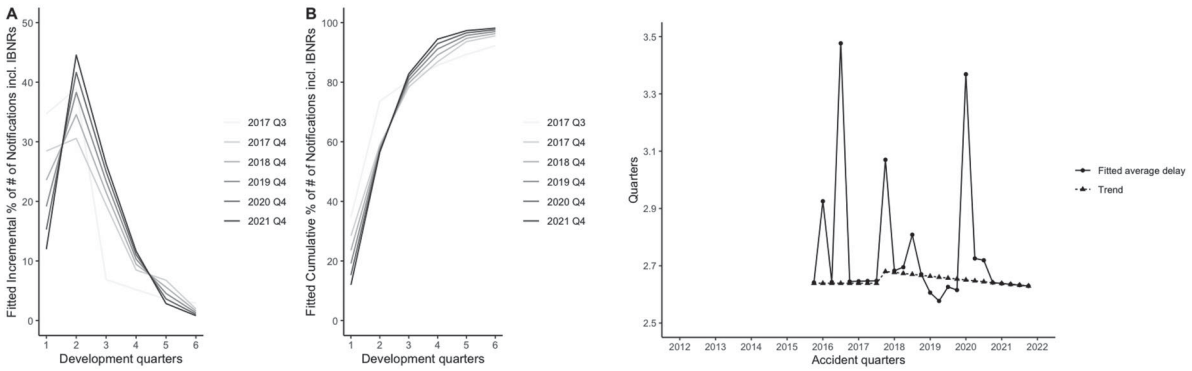
We see the most drastic change in the average delay of breaches that occurred for each of these periods in MT(0–249) (see Table 7). Table 7 shares the same statistics with Figures 5a and 5b and Online Appendix G. See Section 5.3.3 for how we compute these statistics. Although there is no evidence of sustained change of the reporting pattern in MT(0–249), MT(0–249) is subject to the greatest volatility in the average reporting delay across accident periods (see Online Appendix F.3).

In AQ 2016Q3 and AQ 2020Q1, all states are subject to extended reporting delays. For example, in AQ 2020Q1, the average delay in MT(0–249) increases to 3.4 quarters, which would otherwise be 2.6 quarters. IN(0–249) is up by 0.3 quarters to 2.9, from 2.6 quarters. For larger breaches, in CA(>499), the average delay increases to 3.9 quarters, which would have been 3.4 quarters. The delay experience in AQ 2020Q1 is exceptional compared to other AQs for IN(0–249), MT(0–249), CA(>499), and WA(>499); to avoid the distorting effects of such experience, we assign zero weight to it in the modeling.



(a) CA(>499) (b) IN(0–249)

FIGURE 5. Fitted Average Delay and Its Trend.



(a) Development Pattern Trend (b) Fitted Average Delay and Its Trend

FIGURE 6. AQ 2015Q4–AQ 2021Q4, MT(0–249).

TABLE 7
Average Delay in Quarters, before and after Accounting for Exceptions

	2016Q1	2016Q3	2020Q1
CA(>499)	3.1 → 2.9	3.1 → 3.7	3.4 → 3.9
IN(0–249)	2.2 → 2.3	2.2 → 3.1	2.6 → 2.9
MT(0–249)	2.6 → 2.9	2.6 → 3.5	2.6 → 3.4
ME(0–249)	2.4 → 2.5	2.4 → 3.2	NA

In AQ 2016Q1, almost all states are subject to longer-than-usual reporting delays, except that CA(>499) has shorter-than-usual delay. CA(>499) experiences extra-long average delay in AQ 2015Q4. It is also worth noting that WA(>499) and OR(>249) also experienced longer delay (see Online Appendix B). We assign zero weight to the entire AQ 2016Q1 in the modeling of these two states.

CA(>499) and IN(0–249), both of which experience nontrivial lengthening reporting delays after 2017, also see longer-than-usual delays in AQ 2019Q4. See Figures 5a and 5b.

5.4.2. An Increase in Notifications at DQ 4 or DQ 5 Is Noted during Periods of Extended Reporting Delay

During periods of extended reporting delay, most of the time we observe a spike in notifications at DQ 4 or DQ 5, which is at least higher than the number of breaches reported at DQ 3.

Online Appendix B shows the actual quarterly triangles in all data segments. Among breaches that occurred in 2016Q3, the number of notifications at DQ 5, in all four major cases and also IN(>499), is equivalent to or even higher than any other DQs in the same AQ. Among breaches that occurred in 2015Q4 of CA(>499), the number of notifications at DQ 4 is close to the highest. Among breaches that occurred in 2016Q1, IN(0–249), MT(0–249), ME(0–249), and IN(>499) see a spike of notifications at DQ 5.

5.5. Calendar Period Effects

5.5.1. Opposite Calendar Period Effects in Adjacent Periods

A higher/lower total number of notifications received by the state attorney general in a CQ is compensated by a lower/higher number in the next CQ, observed in IN(0–249) and ME(0–249). This often results in a period of longer-than-usual delay followed by shorter-than-usual delay, when speaking to accident periods, and vice versa.

IN(0–249) The number of notifications received in CQ 2017Q4 is 24% lower than what is modeled in the absence of CQ effects, whereas the number of notifications in CQ 2018Q1 is 31% higher. Consequently, breaches that occurred in 2017Q4 have a longer-than-usual delay, followed by 2018Q1 with a shorter-than-usual delay (see Figure 5b).

ME(0–249) The number of notifications received in CQ 2018Q2 is 27% higher than what is modeled in the absence of CQ effects, whereas the number of notifications in CQ 2018Q3 is 21% lower. The number of notifications received in CQ 2018Q4 is 29% lower than what is modeled in the absence of CQ effects, whereas the number of notifications in CQ 2019Q1 is 41% higher. As a result, breaches that occurred in 2018Q1 have a shorter-than-usual delay, followed by 2018Q3 with a longer-than-usual delay and 2019Q1 with a shorter-than-usual delay.

5.6. Residual Model Effects

Item 1: For the 11 less numerous cases (e.g., IN(>499)), we find some similarities between their development patterns. The general structure of their development patterns is represented by Equation (8), and interactions are not explicitly shown to maintain focus.

$$\begin{aligned}\ln(u_{ij}) = \ldots &+ \beta_1 \cdot j + \beta_2 \cdot \max(0, j-6) \\ &+ \beta_3 \cdot 1_{\{1,2,5\}}(j) + \beta_4 \cdot 1_{\{2,5\}}(j) + \beta_5 \cdot 1_{\{5\}}(j) \quad + \ldots\end{aligned} \tag{8}$$

The development pattern amounts to an exponential curve with a change in decay factor at DQ 6 and corrections at DQs 1, 2, and 5. Some might share a common set of coefficients, whereas others might have identical coefficients for the exponential curves (i.e., β_1, β_2) or the corrections (i.e., $\beta_3 \ldots \beta_5$). See the former presented by Terms 8–10 in Table P of Online Appendix C, and the latter presented by Terms 1–7.

When two data segments share the same β_3, it indicates that the DQ coefficient at $j = 1$ in the two segments may be significantly different, but the difference between that coefficient and the coefficients of their respective exponential curves may not be. A common β_4 suggests that the difference between the coefficients at $j = 2$ and $j = 1$ is identical, and a common β_5 suggests that the difference between the coefficients at $j = 5$ and $j = 2$ is identical.

Item 2: To characterize accident period effects, we have trend terms and indicator functions to correct for certain anomalies. For example, in Equation (9), there is a quadratic trend and two indicator functions to address poor fit of two periods.

In some instances, we find that the deviation from the general accident period trend across periods is similar in a single data segment (i.e., similar α_3 and α_4). See Terms 19–22.

Sometimes, different data segments may show a similar deviation from their respective accident period trends in the same period or different periods. This means some segments might have a common α_3 or the α_3 in one segment is similar to the α_4 in another segment. See Terms 11–13. However, it is worth noting that these data segments may not necessarily have the same accident period trend.

$$\ln(u_{ij}) = \ldots + \alpha_1 \cdot i + \alpha_2 \cdot i^2 + \alpha_3 \cdot 1_{i_1}(i) + \alpha_4 \cdot 1_{i_2}(i) \quad + \ldots \tag{9}$$

Item 3: We observe similar calendar period effects across different periods in a single data segment and in the same period across adjacent data segments. In Equation (10), ϕ_1 and ϕ_2 are similar in WA(>499) for CQs 2017Q2 and 2018Q4. A common ϕ_2 is shared across WA(>499) and OR(>249) for CQ 2018Q4, although they do not have the same accident period trend and development pattern. See Term 14.

$$\ln(u_{ij}) = \ldots + \phi_1 \cdot 1_{c_1}(c) + \phi_2 \cdot 1_{c_2}(c) \quad + \ldots \tag{10}$$

Item 4: When correcting for exceptional cells, we find that sometimes the correction is similar across multiple exceptional cells in a single data segment (i.e., similar ψ_1 and ψ_2 in Equation (11)). Sometimes the correction for the same exceptional cell is identical across multiple data segments, indicated by a common ψ_1. For example, IN(>499) has the same correction for (DQ 5, AQ 2016Q1), (DQ 5, AQ 2016Q3), and (DQ 5, AQ 2017Q4); the correction for (DQ 5, AQ 2016Q3) is similar among IN(>499), WA(>499), and OR(>249). See Terms 23 and 24.

$$\ln(u_{ij}) = \ldots + \psi_1 \cdot 1_{i_1}(i) \cdot 1_{j_1}(j) + \psi_2 \cdot 1_{i_2}(i) \cdot 1_{j_2}(j) \quad + \ldots \tag{11}$$

5.7. Limitations

Our analysis presents some limitations. First, despite the fact that we attempted to employ an event definition that takes into account both the third-party provider and all of its affected client firms in the event of a third-party data breach, we only included businesses that met state notification requirements owing to data limitations. Second, the conclusions on data breach notifications and frequency trends were drawn based on a subset of state attorneys general's data that explicitly provides dates of occurrence and dates of notification. Data breaches in other states may display distinct features. Third, we did not incorporate a national analysis owing to the lack of data. When more data are available, it would be interesting to perform multistate analysis on breaches with the same definition, after filtering out common breaches across states. Fourth, it is unclear whether the lengthening delay between occurrence and reporting is the result of a lengthening delay between occurrence and discovery, between discovery and reporting, or both. This ambiguity arises because most states do not explicitly provide dates of discovery. By extracting discovery dates from breach notices, it may be possible to conduct additional research into the causes of lengthening delay. Fifth, it is unknown whether similarities among states in breach experience are the result of a high proportion of third-party data breaches or of breaches affecting residents of multiple states. This could be investigated further by identifying third-party breaches via breach notice letters and identifying breaches that affect residents of multiple states by identifying common breaches across states.

Additionally, in this study, we focus exclusively on data breaches and do not address other prominent cyber risks such as business interruptions, cyber extortion, or ransomware. Practitioners looking for insights into these types of risk will need to consult other sources. It is important to recognize that these risks are becoming major concerns for industries, as evidenced by several high-profile incidents in recent years. For example, ransomware attacks have surged in both frequency and severity, with double and triple extortion methods becoming more common, where attackers not only encrypt systems but also steal sensitive data to leverage extortion demands across business partners and clients (Allianz 2022, 2023). Business interruptions caused by cyber incidents, such as IT system failures or supply chain disruptions, are now the leading source of cyber insurance claims globally, reflecting the critical nature of digital infrastructures to business continuity. These emerging trends highlight the evolving nature of cyber threats, making it essential for organizations to expand their risk management strategies beyond data breaches and for insurers to assess and continuously monitor these broader risks.

6. INSIGHTS AND IMPLICATIONS

This section highlights the main research findings and their implications for pricing, reserving, underwriting, and capital needs in cyber insurance. In Section 6.1, we outline key data considerations to more accurately assess the frequency of both historical and recent data breach risks in the United States, accounting for certain peculiarities in event data analysis, as discussed in Section 2. In addition, we explore how variations in state-level reporting practices impact the ability to derive national-level insights and propose a unified approach to data breach reporting as a potential solution. We also advocate for expanding the scope of reporting to include broader cyber risks, such as business interruptions.

Next, we analyze the delay between the occurrence of a data breach and the reporting to state regulators and examine the frequency of data breaches. Our model uncovers new insights, such as cross-state similarities in breach severity, which are critical for pricing, reserving, and understanding the evolution of cyber incidents. Section 6.2 presents key results and their insurance implications, summarized in Table 8.

6.1. Data Analysis Considerations

6.1.1. A Different Event Definition to Assess the Actual Impact of Data Breaches

Third-party breaches can be viewed as either a single event at the provider level or a series of correlated events across both the provider and affected clients, with each definition shaping a different frequency count, trend, and perceived impact. Defining these breaches as a series of events prevents underestimating frequency, economic impact, and dependencies across organizations. For insurers, adopting this definition helps avoid underestimating total claims and portfolio dependencies, resulting in more accurate pricing. Insurers should consider trends in both primary (provider-level) and secondary (client-level) events. Analyzing these separately clarifies whether increases in frequency stem from more primary breaches or greater ripple effects on client organizations, supporting precise pricing, reserving, and risk assessment. More details can be found in Section 2.2.2.

6.1.2. Frequency Trend Uncovered by Controlling for Changes in the Reporting Propensity

To estimate the true frequency trend of data breaches, we account for changes in reporting propensity over time, influenced by differing state laws and evolving notification requirements. These variations in state laws have likely affected the reporting

TABLE 8
Summary of Results and Insights

Results	Insights
Result 1: Lengthening delay between data breach occurrence and reporting after 2017 in California, Indiana, and Maine	Insight 1.1: Longer time to identify a breach → increased cost of data breaches For policies provided on a discovery basis: • Insight 1.2: Cyber insurers less certain of the actual level of risk assumed • Insight 1.3: A greater assessment of historical attack probability of the insured needed when underwritting Insight 1.4: More efforts toward forecasting IBNR data breach claims
Result 2: Longer delay between occurrence and notification observed in larger breaches than smaller breaches	Insight 2.1: Differentiating between breaches with varying severities in loss reserving
Result 3: Shifting reporting patterns of data breaches with various severities	Insight 3.1: At risk of underestimating IBNRs using the basic chain-ladder method of reserving
Result 4: Occasional extremely lengthy reporting delay	Insight 4.1: For policies provided on a discovery basis, more capital required owing to greater variations of eligible data breach claims (Results 3 and 4) Insight 4.2: Cyber insurers should not assume a favorable position too early
Result 5: States sharing experience in breach frequency trends, the timing of change in reporting patterns, and trends in the average delay between occurrence and reporting	Insight 5.1: Three potential causes • *Reason 1*: data breaches impacting residents of multiple states • *Reason 2*: third-party data breaches • *Reason 3*: attackers being more active during certain periods or becoming more successful in launching widespread attacks If *Reason 2* and *Reason 3* explain → positive dependencies among organizations: • Insight 5.2: Reduced diversification benefits from the portfolio perspective and accumulation risks → higher aggregate premiums and reserves than assuming independence • Insight 5.3: Expect actual frequency experience at the national level to vary widely across quarters

propensities of breaches across different time periods and severities (see Section 2.2.3 for more details). To address this, our analysis focuses on breaches governed by specific state laws and selected time periods.

Key factors affecting reporting propensity include the following:

1. Expanded notification requirements: The inclusion of government bodies like state attorneys general has likely increased publicized breaches. Therefore, we analyze breaches reported to individual state attorneys general that occurred after the respective mandatory notification laws were enacted.
2. Varying state definitions of a reportable breach: Different states have different thresholds for reporting breaches, particularly in terms of the number of affected state residents. This may result in variations in reporting propensity across states, especially for breaches of varying severities. To ensure consistency when comparing data across states, we focus on breaches that are subject to the same reporting criteria.

6.1.3. Advocating for a Unified Approach to Cyber Event Reporting: Reducing Inconsistencies and Enhancing National Insights

The significant variation in state data breach notification laws makes it difficult to derive national-level insights from breach data. Some states do not publish breach data at all, and even among those that do, the definition of reportable breaches varies significantly. These differences make it challenging to aggregate breach data across states, because breaches of different sizes cannot be directly compared or combined. Moreover, organizations that experience breaches affecting residents in multiple states must report separately to the attorneys general of each affected state, further complicating the process of identifying and linking the same breach across different states, which is needed for national aggregation. This fragmentation prevents the development of a unified national perspective on data breaches and adds considerable complexity to the analysis.

To overcome these challenges, we advocate for a unified approach to data breach reporting. A standardized framework across all states would harmonize reporting requirements, simplifying the aggregation of data and facilitating the generation of national insights. Such an approach would also enable more accurate and comprehensive data analysis, supporting better risk management and informed policy decisions at both the state and national levels.

As the landscape of cyber incidents evolves beyond traditional data breaches——with the rise of ransomware attacks, business interruptions, and other forms of cybercrime——we propose expanding the scope of reporting regulations. A unified reporting system should cover a broader spectrum of cyber events to ensure that the full range of cyber risks is being monitored and understood. This would provide regulators and businesses with deeper insights into emerging threats, enhancing the ability to manage and mitigate cyber risks.

6.2. Key Insights from Numerical Analysis and Their Implications

6.2.1. Lengthening Delay between Data Breach Occurrence and Reporting

Result 1: The average delay of data breaches between the first possible date of breach occurrence and the date reported to government bodies lengthened to different extents after 2017 compared to prior periods in California, Indiana, and Maine. For example, for breaches that affect more than 500 California residents, it takes two quarters to reach a reporting rate of 65% for breaches that occurred in 2012Q1, but it takes three quarters for breaches that occurred in 2021Q4.

Insight 1.1: If the lengthening delay between occurrence and reporting is the result of a lengthening delay between discovery and reporting, cyber attacks that result in data breaches are becoming more effective and costly.

In cyber security terminology, the time to detect a breach is referred to as dwell time, the period of time between a cyberattacker's entry and removal from a system (Security Boulevard 2022). Cybercriminals are surreptitious and persistent, frequently operating invisibly on a network for weeks or even months at a time. Long dwell times provide cybercriminals more time to understand the network architecture, the access they have through stolen credentials, and the location of sensitive data (SecurityBrief Australia 2021). Consequently, they are given more opportunity to access personally identifiable information, compromise financial accounts, and insert malicious malware through newer and more sophisticated attacks.

Longer intruder dwell times continue to be associated with greater potential impact of a data breach, a finding noted by IBM since 2016 (IBM 2022). In 2022, the average cost of a data breach with an average time to identify and contain of fewer than 200 days was US$3.74 million, and the cost of breaches with an average time of more than 200 days was US$4.86 million. For breaches with a shorter than 200-day life span, this difference provides an average cost reduction of US$1.12 million, or 26.5%.

Insight 1.2: Cyber insurers are less certain of the risk assumed at the underwriting stage, in terms of whether the potential insured has already been compromised.

A cyber insurance policy can be written on an **occurrence** basis or on a **discovery** basis. On an occurrence basis, the policy will cover insured events that occur during the policy period. On a discovery basis, the policy will cover events that are discovered during the policy period.

The increase in dwell times suggests that enterprises are having a more difficult time spotting threats within their increasingly complex and hybrid networks (Microsoft 2020). For covers that are provided on a discovery basis, such as those on the market, this means that when insurers write the policy, there is greater uncertainty regarding whether or not attackers have already lingered on the network. If this is the case, the organization's likelihood of being attacked increases, thereby increasing the insurer's potential liability.

Insight 1.3: For policies that are provided on a discovery basis, cyber insurers should more thoroughly assess the historical attack probability of the insured.

For policies on a discovery basis, longer dwell time means that cyber insurers might be liable for hacks that are dated back longer if they could not be detected until after policy commencement. For example, the insurer is responsible for a data breach that occurred a year ago, provided the breach is discovered after the policy takes effect. Therefore, at the underwriting stage, cyber insurers should conduct a more comprehensive assessment of a potential policyholder's historical security position to ensure that the premiums charged reflect the underwritten risk. For example, an organization that implemented multifactor authentication 1 month ago should be rated differently than one that did so 1 year ago, owing to the increased likelihood of incurring an attack that remains undetected.

Insight 1.4: Cyber insurers should direct more efforts toward forecasting the financial coverage of IBNR data breach claims.

A greater proportion of data breach claims is expected to be reported in later quarters. Because cyber insurers are required to report their financial condition on a quarterly basis, the need to estimate IBNR data breach claims has increased.

6.2.2. The Reporting Patterns of Data Breaches Shift over Time and Differ by Size of the Breach

Result 2: Larger breaches have longer delay between occurrence and notification than smaller breaches. This is in line with the findings of the IBM *Cost of a Data Breach Report 2022* (IBM 2022).

Insight 2.1: Loss reserving should differentiate among breaches with varying severities.

Because data breach reporting patterns significantly differ by severity of the breach, the severity of breaches should be factored into reserving.

Result 3: The reporting patterns of data breaches with various severities have shifted over time. The percentage of breaches reported in each of development quarters 1 to 6 varies by time periods. For example, an important finding is that the percentage of breaches that are reported within the quarter of their occurrence has decreased from more than 30% in 2017 to less than 15% in 2021 across all states.

Insight 3.1: The basic chain-ladder method of reserving should not be used to predict IBNR breaches, because it may lead to underestimation.

There is substantial evidence of shifting reporting patterns and lengthening delays in data breach reporting across all investigated states and severities, which contradicts the assumption of the basic chain-ladder method of reserving. Using the basic chain-ladder method may lead to inaccurate forecasting and underestimation of IBNRs and breach frequency.

Result 4: Sometimes the reporting delay is unusually lengthy. In such instances, we observe a spike in the reporting of breaches in the fourth and fifth quarters after their occurrence, which can represent as much as 25% of the total number of breaches in the relevant accident quarter.

Insight 4.1: Cyber insurers need more capital for policies that are provided on a discovery basis owing to increased loss reserve uncertainty resulting from greater variations of reporting delay.

Due to increased variations of reporting delay across periods, suggested by **Result 3** and **4**, the number of eligible data breach claims may deviate from the estimated number to a greater extent across different accident periods for policies on a discovery basis. This increases the need for capital, which would occur naturally in any system where capital requirements depend on estimated forecast uncertainty (e.g., the United States, Europe, Australia).

Insight 4.2: Cyber insurers may have an estimation of the number of data breaches that will occur in each time period. If the number received is lower than expected in early development quarters, cyber insurers should not assume a favorable position, because there is a possibility that the number of reported data breaches will spike in later development quarters.

Occasionally, a significant proportion of data breaches that are attributed to the same accident period might be reported in later development quarters. Consequently, even if insurers observe lower-than-usual number of claims in the first three development quarters, they are still uncertain about the ultimate liability and have to wait until later.

6.2.3. Shared Data Breach Experience among States

Result 5: States observe similarities in breach frequency trends, the timing of change in reporting patterns, and trends in the average delay between occurrence and reporting. Different states follow highly similar frequency trends, relatively stationary up to accident quarter 2020Q1 and increasing subsequently. The historical change in reporting pattern commences at a similar point in time across states, around 2017. All states have shown no sign of decreasing average delay between occurrence and reporting, and most have experienced lengthening delay.

Our findings share similarities with the trends observed in Eling, Ibragimov, and Ning (2023), particularly in the post-2020 period, where both studies report an increase in frequency. However, whereas Eling, Ibragimov, and Ning (2023) identified

increasing cyber risk trends prior to 2020, we observe relatively stable frequency trends of data breaches before the first quarter of 2020 across the states in our analysis. Eling, Ibragimov, and Ning (2023) also provided a thorough examination of prior literature on the heterogeneity of frequency trends across datasets, including findings from Maillart and Sornette (2010); Edwards, Hofmeyr, and Forrest (2016); Romanosky (2016); Wheatley, Hofmann, and Sornette (2021); Jung (2021). For brevity, we direct interested readers to their paper, which discusses these prior studies in detail.

Insight 5.1: There are three potential causes of commonalities among states, and none of them are encouraging.

Reason 1: One possible reason might be that a significant proportion of breaches are those that impact residents of multiple states. State attorneys general are concerned about breaches that affect residents of their state. As a result, if a breach affects people in more than one state, the affected organization must submit a separate report to each state attorney general of the affected states regarding this breach. When this kind of breach accounts for a sizable fraction of the total, there is considerable overlap of breaches across states. Then, we would anticipate similarities among states. This means that a substantial percentage of breaches incur the cost of complying with regulations in multiple jurisdictions, which is more expensive than complying with regulations in a single jurisdiction.

Reason 2: Another reason might be that a sizable portion of breaches are third-party breaches. In the event of a third-party data breach, as defined in Section 2.2.2, both the third-party vendor and its affected client firms are obligated to separately report such a breach to the attorneys general of the affected states, in the name of the affected organization. Because these reports are considered breaches in the same occurrence period, we can expect positive state correlations, which lead to shared experience across states. A third-party data breach creates an aggregation problem for the insurer, because it will lead to multiple claims at the same time, when more than one insured is affected by such a breach, whether the vendor or its affected client firms.

Reason 3: If we cannot attribute the similarities among states to either of the first two, the only explanation is that attackers were more active during certain periods than others to launch attacks regardless of geographic location. Alternatively, attackers have become more successful in launching attacks that could compromise a variety of businesses. For instance, attackers can successfully compromise multiple organizations at the same time by exploiting a common vulnerability shared by them. Because these breaches occur at roughly the same time, we can also anticipate positive state correlations. Consequently, positive interdependencies exist among organizations, which is undesirable for cyber insurers.

Regardless of the possible causes mentioned above, the fact that states exhibit similar characteristics in breach experience could have far-reaching implications for cyber insurers, as discussed above. Furthermore, if *Reason 2* and *Reason 3* explain most of the shared experience, there are positive dependencies among organizations, which brings about two further implications.

Insight 5.2: The aggregate premiums and reserves for a cyber portfolio should be higher than those that assume independence of policies.

Positive dependencies across organizations reduce diversification benefits over insureds from the portfolio perspective, compared to more independent policies. As a result, cyber insurers should collect more premiums and need more reserves.

Whether the dependencies stem from the monoculture of software, hardware, or outsourcing services, there is a nontrivial possibility of a single catastrophic event that could affect the majority of the portfolio, which would lead to a significant loss for insurers.

Insight 5.3: When compared to more independent policies, cyber insurance is expected to have greater variations in actual experience across quarters.

The number of breaches may fluctuate widely from quarter to quarter at the national level, as states experience the ups and downs of breach frequency together.

For pricing purposes, insurers should estimate the frequency rate. The frequency rate refers to the number of breaches relative to the exposure, which in this case is the number of businesses subject to data breach notification laws. For example, in California, the exposure consists of businesses that hold information on more than 500 California residents, as required by state law. Because these legal thresholds remain consistent, we assume the exposure remains relatively stable over time. As a result, the frequency trends in the number of data breaches should align with the trends in frequency rate.

Insurers may need to further adjust the trends in frequency rate implied by this study. Specifically, insurers should account for firm-specific characteristics across the cyber portfolio, such as company revenue, industry sector, volume of sensitive data held, and the number of third-party providers the company uses, when estimating the risk of policyholders affected by data breaches, including third-party breaches. These adjustments will help insurers more accurately predict claims and determine appropriate premium levels.

7. CONCLUSION

This article sheds new light on data breach frequency and reporting patterns, by utilizing an underrecognized set of public data provided by U.S. state attorneys general. This set of data was collected based on mandatory state data breach notification laws and provides valuable descriptions of data breaches.

Some of our major takeaways include the following. First, the average reporting delay of data breaches has lengthened after 2017. In light of this finding, cyber insurers may expect a higher cost of data breaches and should direct more effort toward forecasting the financial coverage of IBNR data breach claims. The underwriting of policies on a discovery basis should incorporate a greater assessment of historical attack probability of the insured. Second, the reporting profile of events varies significantly across different accident periods and breach sizes. This means that any loss reserving would require more sophistication than a vanilla chain-ladder technique and would benefit from differentiating across severity levels. More capital may be required for policies provided on a discovery basis. Third, the frequency of data breaches remains relatively stable before 2020 but shows an upward trend across severity levels and states after 2020. This supports the hypothesis that the frequency of cyber events was affected by the pandemic (see, e.g., Cyber Insurance Academy 2021; U.S. Government Accountability Office 2021). Fourth, states share similar experience in breach frequency trends, the timing of change in reporting patterns, and trends in the average delay between occurrence and reporting. Even though such similarities are arguably due to third-party breaches or other common vulnerabilities shared by organizations, this suggests that diversification benefits across states may be less than otherwise anticipated.

Compared to the existing literature, our data and frequency analysis present significant benefits, which are all of particular importance to cyber insurers. The consistency and completeness of the data collection in our set of data allow us to isolate frequency trends from reporting delays, other unrelated trends (such as media attention), as well as differing reporting propensities over time (such as breach notification legislation; see Section 6.1.2). Furthermore, our definition of "event" aligns with what is typical in an insurance portfolio (see Section 6.1.1).

Eventually, our analysis and its methodology can help cyber insurers project IBNR reserves and gain a deeper insight into data breach claim frequency trends. It offers a fresh perspective on the actual magnitude of cyber insurance risks. Our extensive discussion on the implications of our findings is useful for cyber insurance pricing, reserving, underwriting, capital needs, and experience monitoring. Furthermore, commonalities and differences across eight different U.S. states were highlighted and discussed.

Although this study focuses solely on data breaches, it highlights key areas in cyber insurance that may have been previously overlooked but are equally relevant to other types of cyber events. A notable insight from our analysis is that though practitioners often rely on the chain-ladder method to estimate IBNR claims, this approach may underestimate the number of IBNR breaches and, consequently, breach frequency. Cyber insurers should test the validity of the chain-ladder method for reserving events beyond data breaches. If reporting delays or patterns are shifting, insurers may need to be prepared to replace it with an approach that better represents the data and to account for potential implications on cyber insurance.

Though this study focuses on data breaches, it is important to recognize that cyber insurance covers a broader spectrum of risks, such as business interruption, ransomware, and system outages. These risks are not always triggered by data breaches alone and require more comprehensive models and considerations for pricing and reserving. Future research should incorporate these additional risk types to better capture the evolving dynamics of cyber insurance. Specifically, acknowledging the importance of business interruption, particularly in the context of operational disruptions, is crucial to avoid overreliance on data breach–centric projections in the cyber insurance market.

DISCLOSURE STATEMENT

No potential competing interest was reported by the authors.

FUNDING

BA and BW acknowledge support under the Australian Research Council's Discovery Project (DP200101859) funding scheme. The views expressed herein are those of the authors and are not necessarily those of the supporting organizations.

SUPPLEMENTAL MATERIAL

Supplemental material for this article can be accessed on the publisher's website at Online Appendix A.

ORCID

Benjamin Avanzi http://orcid.org/0000-0002-5424-4292
Xingyun Tan http://orcid.org/0000-0002-9397-7659
Greg Taylor http://orcid.org/0000-0002-1439-769X
Bernard Wong http://orcid.org/0000-0002-7124-5342

DATA AVAILABILITY STATEMENT

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

The code used in the analyses, as well as the set of data, is available at https://github.com/agi-lab/data-breaches-reporting.

REFERENCES

Advisen. 2024. Cyber loss data. https://www.advisenltd.com/data/cyber-loss-data/.

Aldasoro, I., L. Gambacorta, P. Giudici, and T. Leach. 2022. The drivers of cyber risk. *Journal of Financial Stability* 60:100989. doi: 10.1016/j.jfs.2022.100989

Allchoice Insurance. 2024. How much is auto insurance? https://allchoiceinsurance.com/auto-insurance-education/how-much-is-auto-insurance/.

Allianz. 2022. Allianz risk barometer 2022: Cyber perils outrank COVID-19 and broken supply chains as top global business risk. https://www.allianz.com/en/press/news/studies/220118_Allianz-Risk-Barometer-2022.html.

Allianz. 2023. Allianz risk barometer 2023. https://www.allianz.com.au/about-us/media-hub/allianz-risk-barometer-2023.html.

American Academy of Actuaries. 2009. Critical issues in health reform: State-level impacts. https://www.actuary.org/sites/default/files/files/publications/state_level_nov2009.pdf.

Antonio, K., and J. Beirlant. 2008. Issues in claims reserving and credibility: A semiparametric approach with mixed models. *Journal of Risk and Insurance* 75:643–76. doi: 10.1111/j.1539-6975.2008.00278.x

Aon. 2021. 2021 Global risk management survey. https://www.aon.com/getmedia/4495a5fa-fef0-4459-8f68-0ab05a5cba25/Aon-2021-GRMS-Asia-Pacific-Region-Report-EN.pdf.aspx.

Benaroch, M. 2021. Third-party induced cyber incidents—Much ado about nothing? *Journal of Cybersecurity* 7:tyab020. doi: 10.1093/cybsec/tyab020

Bessy-Roland, Y., A. Boumezoued, and C. Hillairet. 2021. Multivariate Hawkes process for cyber insurance. *Annals of Actuarial Science* 15:14–39. doi: 10.1017/S1748499520000093

Bruce, M., J. Lusthaus, R. Kashyap, N. Phair, and F. Varese. 2024. Mapping the global geography of cybercrime with the world cybercrime index. *PLOS One* 19. doi: 10.1371/journal.pone.0297312

California Attorney General. 2024. Search data security breaches. https://oag.ca.gov/privacy/databreach/list.

California Legislature. 2024. California civil code § 1798.82. https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?sectionNum=1798.82.&lawCode=CIV.

Chang, L., G. Gao, and Y. Shi. 2023. Claims reserving with a robust generalized additive model. *North American Actuarial Journal* 28(4):840–60.

Chen, S., M. Hao, F. Ding, D. Jiang, J. Dong, S. Zhang, Q. Guo, and C. Gao. 2023. Exploring the global geography of cybercrime and its driving forces. *Humanities and Social Sciences Communications* 10:1–10. doi: 10.1057/s41599-023-01560-x

Cho, J., M. Eling, and K. Jung. 2024. Spatial cyber loss clusters at county level and socioeconomic determinants of cyber risks. *North American Actuarial Journal* 1–45. doi: 10.1080/10920277.2024.2408263

Costa, L., A. Pizzinga, and R. Atherino. 2016. Modeling and predicting IBNR reserve: Extended chain ladder and heteroscedastic regression analysis. *Journal of Applied Statistics* 43:847–70. doi: 10.1080/02664763.2015.1079305

Cyber Insurance Academy. 2021. Cyber insurance market 2022: What you need to know. https://www.cyberinsuranceacademy.com/knowledge-hub/guide/what-you-need-to-know-about-the-cyber-insurance-market-before-2022/.

Cybersecurity Ventures. 2022. Cybercrime to cost the world 8 trillion annually in 2023. https://cybersecurityventures.com/cybercrime-to-cost-the-world-8-trillion-annually-in-2023/.

Delaware Attorney General. 2024. Data security breach database. https://attorneygeneral.delaware.gov/fraud/cpu/securitybreachnotification/database/.

Delaware Legislature. 2024. Delaware code, title 6, chapter 12b. https://delcode.delaware.gov/title6/c012b/index.html.

Deloitte. 2020. Unlocking the value of cyber insurance. https://www2.deloitte.com/content/dam/Deloitte/in/Documents/risk/in-ra-CyberThoughtpaper-noexp.pdf.

Deloitte. 2023. Cybersecurity threats and incidents differ by region. https://www.deloitte.com/global/en/services/risk-advisory/perspectives/cybersecurity-threats-and-incidents-differ-by-region.html.

Edwards, B., S. Hofmeyr, and S. Forrest. 2016. Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity* 2:3–14. doi: 10.1093/cybsec/tyw003

Eling, M., R. Ibragimov, and D. Ning. 2023. Time dynamics of cyber risk. SSRN:4497621.

Eling, M., and K. Jung. 2018. Copula approaches for modeling cross-sectional dependence of data breach losses. *Insurance: Mathematics and Economics* 82:167–80. doi: 10.1016/j.insmatheco.2018.07.003

Eling, M., K. Jung, and J. Shim. 2022. Unraveling heterogeneity in cyber risks using quantile regressions. *Insurance: Mathematics and Economics* 104:222–42. doi: 10.1016/j.insmatheco.2022.03.001

Eling, M., and N. Loperfido. 2017. Data breaches: Goodness of fit, pricing, and risk measurement. *Insurance: Mathematics and Economics* 75:126–36. doi: 10.1016/j.insmatheco.2017.05.008

Eling, M., and J. Wirfs. 2019. What are the actual costs of cyber risk events? *European Journal of Operational Research* 272:1109–19. doi: 10.1016/j.ejor.2018.07.021

England, P. D., and R. J. Verrall. 1998. *Standard errors of prediction in claims reserving: A comparison of methods*. London: Institute of Actuaries.

England, P. D., and R. J. Verrall. 2001. A flexible framework for stochastic claims reserving. In *Proceedings of the Casualty Actuarial Society*, 1–38. Arlington, VA :Casualty Actuarial Society (CAS).

England, P. D., and R. J. Verrall. 2002. Stochastic claims reserving in general insurance. *British Actuarial Journal* 8(3):443–544.

Farkas, S., O. Lopez, and M. Thomas. 2021. Cyber claim analysis using generalized Pareto regression trees with applications to insurance. *Insurance: Mathematics and Economics* 98:92–105. doi: 10.1016/j.insmatheco.2021.02.009

Flexera. 2022. 2022 State of the cloud report by Flexera. https://www.flexera.com/about-us/press-center/2022-state-of-the-cloud-report-by-flexera.

Fortra. 2021. Sabre agrees to $2.4m settlement following 2017 data breach. https://www.digitalguardian.com/blog/sabre-agrees-24m-settlement-following-2017-data-breach.

Grize, Y. L. 2015. Applications of statistics in the field of general insurance: An overview. *International Statistical Review* 83:135–59. doi: 10.1111/insr.12066

Harry, C., and N. W. Gallagher. 2023. Categorizing cyber effects. In *The Elgar companion to digital transformation, artificial intelligence and innovation in the economy, society and democracy*, ed. E. G. Carayannis, E. Grigoroudis, D. F. J. Campbell, and S. K. Katsikas, 7–31. Northampton, MA: Edward Elgar.

Hawaii Department of Commerce and Consumer Affairs. 2024. Security breach notices. https://cca.hawaii.gov/ocp/notices/security-breach/.

Hawaii Legislature. 2024. Hawaii revised statutes, chapter 487n. https://www.capitol.hawaii.gov/hrscurrent/Vol11_Ch0476-0490/HRS0487N/HRS_0487N-.htm.

IBM. 2022. Cost of a data breach report 2022. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.key4biz.it/wp-content/uploads/2022/07/Cost-of-a-Data-Breach-Full-Report-2022.pdf.

Indiana Attorney General. 2024. Security breaches. https://www.in.gov/attorneygeneral/consumer-protection-division/id-theft-prevention/security-breaches/.

Indiana Legislature. 2022. Indiana code, title 24, article 4.9. https://iga.in.gov/laws/2022/ic/titles/24#24-4.9.

Insurance Business. 2023. Cyber insurance market edges towards US$14bn. https://www.insurancebusinessmag.com/au/news/cyber/cyber-insurance-market-edges-towards-us14bn-450004.aspx.

Insurance Information Institute. 2023. Average premiums for homeowners and renters insurance by state, 2021. https://www.iii.org/fact-statistic/facts-statistics-homeowners-and-renters-insurance#Average%20Premiums%20for%20Homeowners%20and%20Renters%20Insurance%20by%20State,%202021%20(1).

International Association of Privacy Professionals. 2021. State data breach notification chart. https://iapp.org/resources/article/state-data-breach-notification-chart/.

International Comparative Legal Guides. 2024. Data protection laws and regulations USA 2023–2024. https://iclg.com/practice-areas/data-protection-laws-and-regulations/usa.

Iowa Attorney General. 2024. Security breach notifications. https://www.iowaattorneygeneral.gov/for-consumers/security-breach-notifications.

Iowa Legislature. 2024. Iowa code, chapter 715c. https://www.legis.iowa.gov/docs/code/715C.pdf.

Jung, K. 2021. Extreme data breach losses: An alternative approach to estimating probable maximum loss for data breach risk. *North American Actuarial Journal* 25:580–603. doi: 10.1080/10920277.2021.1919145

Kapoor, A., and D. L. Nazareth. 2013. Medical data breaches: What the reported data illustrates, and implications for transitioning to electronic medical records. *Journal of Applied Security Research* 8:61–79. doi: 10.1080/19361610.2013.738397

Kesan, J. P., and L. Zhang. 2020. Analysis of cyber incident categories based on losses. *ACM Transactions on Management Information Systems* 11:1–28. doi: 10.1145/3418288

Kremer, E. 1982. IBNR-claims and the two-way model of ANOVA. *Scandinavian Actuarial Journal* 1982:47–55. doi: 10.1080/03461238.1982.10405432

Li, Y., and R. Mamon. 2023. Modelling health-data breaches with application to cyber insurance. *Computers & Security* 124:102963. doi: 10.1016/j.cose.2022.102963

Liu, J., J. Li, and K. Daly. 2022. Bayesian vine copulas for modelling dependence in data breach losses. *Annals of Actuarial Science* 16(2):401–24.

Lu, Y., J. Zhang, and W. Zhu. 2023. Cyber risk modeling: A discrete multivariate count process approach. *Scandinavian Actuarial Journal* 2024(6):625–65.

Mack, T. 1993. Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin* 23:213–25. doi: 10.2143/AST.23.2.2005092

Mack, T. 1994. Which stochastic model is underlying the chain ladder method? *Insurance: Mathematics and Economics* 15:133–38. doi: 10.1016/0167-6687(94)90789-7

Maillart, T., and D. Sornette. 2010. Heavy-tailed distribution of cyber-risks. *The European Physical Journal B* 75:357–64. doi: 10.1140/epjb/e2010-00120-8

Maine Attorney General. 2024. Privacy, identity theft and data security breaches. https://www.maine.gov/ag/consumer/identity_theft/index.shtml.

Maine Legislature. 2019. Maine legislature, s.p. 209, 129th legislature, 1st regular session. https://legislature.maine.gov/legis/bills/getPDF.asp?paper=SP0209&item=5&snum=129&PID=.

Maine Legislature. 2024. The notice of risk to personal data act. https://legislature.maine.gov/statutes/10/title10sec1346.html.

Malavasi, M., G. W. Peters, P. V. Shevchenko, S. Trück, J. Jang, and G. Sofronov. 2022. Cyber risk frequency, severity and insurance viability. *Insurance: Mathematics and Economics*.

Marsh. 2022. The state of cyber resilience report. https://www.marsh.com/us/services/cyber-risk/insights/the-state-of-cyber-resilience.html#sizetracker.

Maryland Attorney General. 2023. Maryland information security breach notices. https://www.marylandattorneygeneral.gov/Pages/IdentityTheft/breachnotices.aspx.

Maryland Legislature. 2024. Maryland code, commercial law, section 14-3504. https://mgaleg.maryland.gov/mgawebsite/laws/StatuteText?article=gcl&section=14-3504&enactments=False&archived=False.

Massachusetts Legislature. 2024. Massachusetts general laws, part i, title xv, chapter 93h, section 1. https://malegislature.gov/Laws/GeneralLaws/PartI/TitleXV/Chapter93H/Section1.
Microsoft. 2020. Microsoft report shows increasing sophistication of cyber threats. https://blogs.microsoft.com/on-the-issues/2020/09/29/microsoft-digital-defense-report-cyber-threats/.
Montana Department of Justice. 2023. Security breaches. https://dojmt.gov/consumer/databreach/.
Montana Legislature. 2024. Montana code annotated, title 30, chapter 14, part 17. https://leg.mt.gov/bills/mca/title_0300/chapter_0140/part_0170/sections_index.html.
National Academy of Social Insurance. 2022. Workers' compensation: Benefits, costs, and coverage. https://www.nasi.org/wp-content/uploads/2022/11/2022-Workers-Compensation-Report-2020-Data.pdf.
National Association of Attorneys General. 2024. Data breaches. https://www.naag.org/issues/consumer-protection/consumer-protection-101/privacy/data-breaches/.
National Association of Insurance Commissioners. 2022. Report on the cyber insurance market. https://content.naic.org/sites/default/files/cmte-c-cyber-supplement-report-2022-for-data-year-2021.pdf.
National Association of Insurance Commissioners. 2024a. Financial statement filing & step through guide. https://content.naic.org/industry_financial_filing.htm.
National Association of Insurance Commissioners. 2024b. State filing instructions and checklists. https://content.naic.org/industry_filing_state_instructions.htm.
The National Conference of State Legislatures. 2024. Security breach notifications laws. https://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx.
New Hampshire Department of Justice. 2024. Security breach notifications. https://www.doj.nh.gov/bureaus/consumer-protection-antitrust-bureau/security-breach-notifications-0.
New Jersey Cybersecurity & Communications Intergration Cell. 2024. Public data breaches. https://www.cyber.nj.gov/threat-center/public-data-breaches/.
North Dakota Attorney General. 2024. Data breach notices. https://attorneygeneral.nd.gov/consumer-resources/data-breach-notices.
North Dakota Legislature. 2024. North dakota century code, title 51, chapter 30. https://www.ndlegis.gov/cencode/t51c30.html.
Office of the Australian Information Commissioner. 2019. Notifiable data breaches scheme 12-month insights report. https://www.oaic.gov.au/privacy/notifiable-data-breaches/notifiable-data-breaches-statistics/notifiable-data-breaches-scheme-12month-insights-report.
Oklahoma Office of Management & Enterprise Services. 2023. Cybersecurity breaches. https://oklahoma.gov/omes/services/information-services/is/cyber-command/cybersecurity-breaches.html.
Oladimeji, S., and S. M. Kerner. 2023. Solarwinds hack explained: Everything you need to know. https://www.techtarget.com/whatis/feature/SolarWinds-hack-explained-Everything-you-need-to-know.
Oregon Department of Justice. 2024. Search data breaches. https://justice.oregon.gov/consumer/DataBreach/.
Oregon Legislature. 2024. Oregon revised statutes, section 646a.600. https://oregon.public.law/statutes/ors_646a.600.
Organization for Economic Cooperation and Development. 2020. Sustainable cyber insurance markets. https://web-archive.oecd.org/2020-08-18/546659-building-a-sustainable-cyber-insurance-market.htm.
Palsson, K., S. Gudmundsson, and S. Shetty. 2020. Analysis of the impact of cyber events for cyber insurance. *The Geneva Papers on Risk and Insurance-Issues and Practice* 45:564–79. doi: 10.1057/s41288-020-00171-w
K. Peremans, P. Segaert, S. van Aelst, and T. Verdonck. 2017. Robust bootstrap procedures for the chain-ladder method. *Scandinavian Actuarial Journal* 2017:870–97. doi: 10.1080/03461238.2016.1263236
Pinheiro, P. J. R., J. M. Andrade e Silva, and M. de Lourdes Centeno. 2003. Bootstrap methodology in claim reserving. *Journal of Risk and Insurance* 70: 701–14. doi: 10.1046/j.0022-4367.2003.00071.x
Ponemon Institute. 2022. Ponemon 2022 study: Data risk in the third-party ecosystem. https://www.riskrecon.com/ponemon-report-data-risk-in-the-third-party-ecosystem-study.
Poyraz, O. I., M. Canan, M. McShane, C. A. Pinto, and T. S. Cotter. 2020. Cyber assets at risk: Monetary impact of U.S. personally identifiable information mega data breaches. *The Geneva Papers on Risk and Insurance-Issues and Practice* 45:616–38. doi: 10.1057/s41288-020-00185-4
Prevalent. 2024. Third-party data breaches: What you need to know. https://www.prevalent.net/blog/third-party-data-breaches/.
Privacy Rights Clearinghouse. 2023. Data breach notification in the United States 2022 report. https://privacyrights.org/resources/data-breach-notification-united-states-and-territories.
Privacy Rights Clearinghouse. 2024. Data breach chronology database. https://privacyrights.org/data-breaches.
Renshaw, A. E., and R. J. Verrall. 1998. A stochastic model underlying the chain-ladder technique. *British Actuarial Journal* 4:903–23. doi: 10.1017/S1357321700000222
Romanosky, S. 2016. Examining the costs and causes of cyber incidents. *Journal of Cybersecurity* 2:121–35. doi: 10.1093/cybsec/tyw001
Sangari, S., E. Dallal, and M. Whitman. 2022. Modeling reporting delays in cyber incidents: An industry-level comparison. *International Journal of Information Security* 22(1):63–76. doi: 10.1007/s10207-022-00623-5
SAS. 2024. SAS oprisk var. https://support.sas.com/en/software/oprisk-var-support.html#documentation.
Security Boulevard. 2022. Dwell time can impact the outcome of a cyberattack. https://securityboulevard.com/2022/05/dwell-time-can-impact-the-outcome-of-a-cyberattack/.
SecurityBrief Australia. 2021. Why the biggest cyber-attacks go undetected. https://securitybrief.com.au/story/why-the-biggest-cyber-attacks-go-undetected.
Shevchenko, P. V., J. Jang, M. Malavasi, G. W. Peters, G. Sofronov, and S. Trück. 2023. The nature of losses from cyber-related events: Risk categories and business sectors. *Journal of Cybersecurity* 9:tyac016. doi: 10.1093/cybsec/tyac016
Shi, P. 2017. A multivariate analysis of intercompany loss triangles. *Journal of Risk and Insurance* 84:717–37. doi: 10.1111/jori.12102
Sriram, K., and P. Shi. 2021. Stochastic loss reserving: A new perspective from a Dirichlet model. *Journal of Risk and Insurance* 88:195–230. doi: 10.1111/jori.12311
Strupczewski, G. 2019. What is the worst scenario? Modeling extreme cyber losses. In *Multiple perspectives in risk and risk management*, ed. P. Linsley, P. Shrives, and M. Wieczorek-Kosmala, 211–30. Cham, Switzerland: Springer.

Sun, H., M. Xu, and P. Zhao. 2021. Modeling malicious hacking data breach risks. *North American Actuarial Journal* 25:484–502. doi: 10.1080/10920277.2020.1752255

Taylor, G. 2012. *Loss reserving: An actuarial perspective*. Vol. 21. Dordrecht, Netherlands: Springer Science & Business Media.

Taylor, G., and G. McGuire. 2016. Stochastic loss reserving using generalized linear models. *CAS Monograph* 3.

Texas Attorney General. 2024. Data security breach reports. https://oag.my.site.com/datasecuritybreachreport/apex/DataSecurityReportsPage.

Texas Legislature. 2024. Texas business and commerce code, title 11, chapter 521. https://statutes.capitol.texas.gov/Docs/BC/htm/BC.521.htm.

U.S. Department of Health and Human Services. 2024. Breach portal: Notice to the secretary of hhs breach of unsecured protected health information. https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf.

U.S. General Services Administration. 2018. The value of federal government data. https://digital.gov/2018/03/14/data-briefing-value-federal-government-data/.

U.S. Government Accountability Office. 2021. Cyber insurance: Insurers and policyholders face challenges in an evolving market. https://www.gao.gov/products/gao-21-477.

USAFacts. 2023. Why should you trust U.S. government data? https://usafacts.org/articles/why-should-you-trust-us-government-data/.

Verizon. 2024. 2024 Data breach investigations report. https://www.verizon.com/business/resources/reports/dbir/.

Vermont Attorney General. 2024. Security breach notices. https://ago.vermont.gov/categories/security-breach-notices.

Verrall, R. J. 1994. Statistical methods for the chain-ladder technique. In *Casualty Actuarial Society forum*, 393–446. Dordrecht, Netherlands.

Verrall, R. J. 2000. An investigation into stochastic claims reserving models and the chain-ladder technique. *Insurance: Mathematics and Economics* 26:91–99. doi: 10.1016/S0167-6687(99)00038-4

Washington, A. L. 2014. Government information policy in the era of big data. *Review of Policy Research* 31:319–25. doi: 10.1111/ropr.12081

Washington Legislature. 2024. Revised code of Washington, chapter 19.255.010. https://app.leg.wa.gov/RCW/default.aspx?cite=19.255.010.

Washington State Office of the Attorney General. 2024. Data breach notifications directory. https://www.atg.wa.gov/data-breach-notifications.

Wei, L., J. Li, and X. Zhu. 2018. Operational loss data collection: A literature review. *Annals of Data Science* 5:313–37. doi: 10.1007/s40745-018-0139-2

Wheatley, S., A. Hofmann, and D. Sornette. 2021. Addressing insurance of data breach cyber risks in the catastrophe framework. *The Geneva Papers on Risk and Insurance-Issues and Practice* 46:53–78. doi: 10.1057/s41288-020-00163-w

Wheatley, S., T. Maillart, and D. Sornette. 2016. The extreme risk of personal data breaches and the erosion of privacy. *The European Physical Journal B* 89:1–12. doi: 10.1140/epjb/e2015-60754-4

Wisconsin Department of Agriculture, Trade and Consumer Protection. 2024. Data breaches. https://datcp.wi.gov/Pages/Programs_Services/DataBreaches.aspx.

Woodruff Sawyer. 2020. You can outsource a service, but not cyber risk. https://woodruffsawyer.com/cyber-liability/you-can-outsource-service-not-cyber-risk/.

Wüthrich, M. V., and M. Merz. 2008. *Stochastic claims reserving methods in insurance*. Vol. 435. Chichester, West Sussex: John Wiley & Sons.

Xu, M., K. M. Schweitzer, R. M. Bateman, and S. Xu. 2018. Modeling and predicting cyber hacking breaches. *IEEE Transactions on Information Forensics and Security* 13:2856–71. doi: 10.1109/TIFS.2018.2834227

Yan, A., and N. Weber. 2018. Mining open government data used in scientific research. In *Transforming digital worlds: 13th International conference, iConference 2018*, *Sheffield, UK*, March 25–28, 2018, *Proceedings 13*, ed. G. Chowdhury, J. McLeod, V. Gillet, and P. Willett, 303–13. Cham, Switzerland: Springer.