A MULTIVARIATE FREQUENCY-SEVERITY FRAMEWORK FOR HEALTHCARE DATA BREACHES

BY HONG SUN¹, MAOCHAO XU² AND PENG ZHAO³

¹School of Mathematics and Statistics, Lanzhou University, China sunh19@lzu.edu.cn
²Department of Mathematics, Illinois State University, USA mxu2@ilstu.edu
³School of Mathematics and Statistics, Jiangsu Normal University, China zhaop@jsnu.edu.cn

Data breaches in healthcare have become a substantial concern in recent years, and cause millions of dollars in financial losses each year. It is fundamental for government regulators, insurance companies, and stakeholders to understand the breach frequency and the number of affected individuals in each state, as these are directly related to the federal Health Insurance Portability and Accountability Act (HIPAA) and state data breach laws. However, an obstacle to studying data breaches in healthcare is the lack of suitable statistical approaches. We develop a novel multivariate frequency-severity framework to analyze breach frequency and the number of affected individuals at the state level. A mixed effects model is developed to model the square root transformed frequency, and the log-gamma distribution is proposed to capture the skewness and heavy tail exhibited by the distribution of numbers of affected individuals. We further discover a positive nonlinear dependence between the transformed frequency and the log-transformed numbers of affected individuals (i.e., severity). In particular, we propose to use a D-vine copula to capture the multivariate dependence among conditional severities given frequencies due to its inherent temporal structure and rich bivariate copula families. The rejection sampling technique is developed to simulate the predictive distributions. Both the in-sample and out-of-sample studies show that the proposed multivariate frequency-severity model that accommodates non-linear dependence has satisfactory fitting and prediction performances.

1. Introduction. Data breaches have become one of the most devastating cyber incidents and can lead to severe consequences, including the exposure of sensitive/confidential information, identity fraud, and financial losses. This situation is even worse in healthcare, as records including personal and medical information are highly valuable and can be easily used for identity fraud and ransomware. In the US, there were 3,054 healthcare data breach incidents involving more than 500 records between 2009 and 2019. These incidents resulted in the exposure of 230,954,151 healthcare records (HJ, 2020a). The year 2019 was a particularly bad one for the healthcare sector, as 41.2 million healthcare records were breached that, which is 195.61% more than 2018 (HJ, 2020b). According to the IBM 2020 Cost of a Data Breach Report (IBM, 2020), the average global cost of a data breach in 2019 was $3.86 million, while in the healthcare industry this cost was $7.13 million.

In healthcare, the federal HIPAA data breach notification rules require covered entities and business associates to ensure the security, confidentiality, and integrity of data. There are explicit rules for handling, storing, and transferring patient data. When patient data are breached, healthcare organizations have to adhere to a strict notification process and may even be liable for severe fines. The federal regulation is a key consideration in the data breach notification process. However, the data breach regulations at the state level cannot be overlooked, as each state has different regulations and laws. All the states in the US have

Keywords and phrases: Copula, Data breach, Heavy tail, Multivariate dependence, Score.
enacted legislation requiring private or governmental entities to notify individuals of security breaches involving personally identifiable information (ITG, 2020). Each state has its own data breach notification process, and when a data breach involving personal information occurs, the relevant businesses should inform the affected individuals. In healthcare, it often takes a long time to review the existing federal and state regulations for notification in the instance of a potential breach. Therefore, it is fundamental for government regulators, insurance companies, and stakeholders to understand the breach frequency and number of affected individuals at the state level, as these are directly related to the federal HIPAA and state data breach laws.

In the literature, there are several studies on the statistical modeling of data breaches facilitated by public data. For example, Maillart and Sornette (2010) studied the statistical properties of data breaches in US from 2000 to 2008 based on a dataset from the Open Security Foundation (OSF), and modeled the size of a data breach with a heavy-tailed power-law distribution. Using the breach dataset published by the Privacy Rights Clearinghouse (PRC), Edwards, Hofmeyr and Forrest (2016) showed that the daily breach size follows a log-normal distribution and the frequency follows a negative binomial distribution. They further showed that the frequency of large breaches (over 500,000 records breached) follows a different distribution, namely Poisson distribution, but the size of large breach still follows a log-normal distribution. By combining the datasets from the OSF and the PRC, Wheatley, Maillart and Sornette (2016) studied the tail properties of the distribution (more than 50,000 records breached) in which large breaches were modeled by a doubly truncated Pareto distribution and a linear regression model was used to study the frequency of data breaches. Based on Advisen\(^1\) data, Romanosky (2016) proposed a logistic regression model to study the relationship between the cost of data breach events and the number of exposed records. Eling and Jung (2018) studied the cross-sectional dependence among monthly losses based on breach events from 2005 to 2016 in the PRC dataset. The vine copula approach was proposed to model the dependence. Buckman et al. (2018) investigated the effect of data breach notification policy on data breaches by using the PRC data between 2005 and 2016. They developed a panel regression with a fixed effects model to test several hypotheses on the effect of policies. Xu et al. (2018) examined the hacking breach in terms of the incidents’ inter-arrival time and the breach size. They discovered that there is a nonlinear dependence between the inter-arrival time and the breach size, which can be captured well by the Gumbel copula. They further argued that both the inter-arrival time and breach size should be modeled with stochastic processes rather than distributions. Eling and Wirfs (2019) used an operational risk dataset provided by the SAS Institute to study cyber losses and applied the extreme value theory to analyze extreme cyber risks. They showed that the log-normal distribution may lead to a substantial overestimate of the actual cyber losses based on the analyzed data. More recently, Sun, Xu and Zhao (2020) developed a frequency-severity model by focusing on malicious hacking data breaches at the individual enterprise level, and further studied the application of the proposed model for ratemaking and pricing. They showed that breach frequency can be modeled by a hurdle Poisson model, and that breach severity which shows a heavy tail can be captured by a nonparametric-generalized Pareto distribution. Fang et al. (2021) studied the enterprise-level data breach risk based on the PRC dataset, in which a mixed D-vine dependence structure was employed to accommodate the complex dependence exhibited by the enterprise-level breach incident time series. They discovered that the enterprise-level breach sizes exhibit a negative temporal dependence, which can be well

\(^{1}\)Advisen is a US-based organization that collects and processes cyber loss and incident data reported by various sources including public government websites and private sources. Its cyber loss data has been widely used by insurance companies, brokers and cyber modeling firms.
modeled by the D-vine structure. One may also refer to Eling (2020) for reviews on the recent developments in cyber risk modelings.

However, there are few statistical models for healthcare breach data at the state level, although this perspective should never be overlooked. In this study, we aim to fill this gap by developing a novel multivariate frequency-severity framework to jointly model two important metrics, namely, the frequency and number of affected individuals each year at the state level. Those two fundamental metrics can help federal and state regulators, insurance companies, or other stakeholders to better understand the state-level data breach risk. On the one hand, those two metrics and their varieties (e.g., average severities) can help law-makers to enact better legislation at both state and federal levels (Bisogni, 2016). For example, the level and the limit of incident penalties based on HIPAA Omnibus Rule can be better adjusted and enforced for each state by using frequency and severity as individual state factors (Yaraghi and Gopal, 2018). On the other hand, our proposed model can be directly used by the U.S. Department of Health and Human Services Office for Civil Rights to efficiently budget and allocate resources to handle data breach incidents for each state (HHS, 2020).

The proposed multivariate frequency-severity framework has several intriguing properties, since it can model the arbitrary dependence between frequency and severity (i.e., log-transformed number of affected individuals), capture the temporal dependence among conditional severities given frequencies, and further accommodate the heterogeneity among frequencies. Specifically, we construct a parametric copula to model the joint dependence between frequency and severity variables, which is different from the standard independence assumption in the literature; see Frees, Lee and Yang (2016). We discover that there is a non-linear positive dependence between the transformed frequency and the severity, and that this dependence can be well characterized by Frank copulas. It should be mentioned that ignoring the due dependence between the frequency and the severity can lead to serious biases in the inference (Frees, Lee and Yang, 2016; Shi and Yang, 2018). In particular, the accuracy of the predictive distribution of severity can be significantly improved when the dependence is properly specified. Therefore, it is very valuable to model the dependence between the frequency and the severity. In addition, we introduce a linear mixed effects model to capture the heterogeneity among frequencies across states. We further employ a D-vine copula to capture the temporal dependence among severities given frequencies. The traditional approach to modeling multivariate dependence is to employ a multivariate distribution (e.g., Gaussian). However, there are several advantages to using a copula approach: i) It can accommodate any linear or nonlinear dependence between frequency and severity, and more specifically, it can capture any potential tail dependence (Joe, 2014). The independence or linear dependence models are included as special cases. ii) It separates marginals from the dependence structure based on Sklar’s theorem (Sklar, 1959). Therefore, the regression can be adapted into the marginal modeling to model complex scenarios such as heterogeneity and/or random effects. In particular, the D-vine approach can accommodate much more complex temporal dependence by using asymmetric bivariate copulas as building blocks (Czado, 2019). iii) It possesses computational efficiency due to the parametric nature of the copula approach, and makes the inference and prediction more straightforward. The Inference Function for Margins (IFM) computational approach associated with copula modeling can further facilitate the computational efficiency, which is extremely important in practice to handle complex data (Joe, 2014). For the severity, we develop a rejection sampling algorithm to simulate the predictive distribution. To the best of our knowledge, ours is the first study to use the D-vine copula to accommodate the temporal dependence among severities given frequencies while introducing a linear mixed effects model to accommodate the heterogeneity in the frequency-severity framework. The empirical analysis of state data breaches in the US and synthetic data study confirm the superior performance of the multivariate frequency-severity framework.
The rest of the paper is organized as follows. Section 2 presents some preliminaries on the linear mixed effects model and D-vine copula. In Section 3, we perform the exploratory data analysis on a real breach data in healthcare to motivate the model. Section 4 discusses the novel multivariate frequency-severity framework for modeling and predicting the breach risk. Section 5 presents an empirical study on the state level breach data in healthcare. In the last Section, we conclude our study and present some discussion.

2. Preliminaries.

2.1. Linear mixed-effects model. The linear mixed-effects model (LMM) is a very popular tool in data analysis, which extends linear regression models for data that are collected in groups. It is a linear regression model that typically includes random and/or fixed effects (West, Welch and Galecki, 2014). A natural application is to panel data where the random effects vary between subjects and induce within-subject dependence among repeated measurements after conditioning on observed covariates (Laird and Ware, 1982; Verbeke, 1997).

In statistical inference, the LMM is often represented as

\[ Y_{ij} = x_{ij}^\top \alpha + z_{ij}^\top b_i + \varepsilon_{ij}, \]

where \( Y_{ij} \) is a response for unit \( i = 1, 2, ..., M \) and a given subject \( j = 1, ..., n_i \), \( x_{ij} \) is a \( p \)-dimensional vector of covariates with fixed effects \( \alpha \), and \( z_{ij} \) is a \( q \)-dimensional vector of covariates with random effects \( b_i \), and \( \varepsilon_{ij} \) represents the residual error. Given the random effects and covariates, it follows that the conditional expectation of the response is

\[ \mu_{ij} = \mathbb{E}[Y_{ij}|b_i, x_{ij}, z_{ij}] = x_{ij}^\top \alpha + z_{ij}^\top b_i, \]

The random effects are assumed to have multivariate normal distributions \( b_i \sim N(0, G) \) where \( G \) is the variance matrix, and the residual error is also assumed to follow multivariate normal distributions \( \varepsilon_{ij} \sim N(0, \sigma^2_e) \). It is assumed that the responses are conditionally independent given the covariates and random effects, and have normal distributions

\[ f(y_{ij}|b_i, x_{ij}, z_{ij}) = \frac{1}{\sigma_e \sqrt{2\pi}} \exp \left[ -\frac{(y_{ij} - \mu_{ij})^2}{2\sigma^2_e} \right]. \]

For more details on LMM and its applications, please refer to Jiang (2007); Dobson and Barnett (2018).

2.2. Copula and D-vine. The copula has been an effective and popular tool for modeling high-dimensional dependence. It has been applied in many areas due to its flexibility and capability to model nonlinear dependence (Joe, 2014).

Let \( X_1, \ldots, X_d \) be continuous random variables with univariate marginal distributions \( F_1, \ldots, F_d \), respectively. Denote their joint cumulative distribution function (CDF) by

\[ F(x_1, \ldots, x_d) = P(X_1 \leq x_1, \ldots, X_d \leq x_d). \]

A \( d \)-dimensional copula, denoted by \( C \), is a CDF with uniform marginals in \([0, 1]\), namely the joint CDF of the random vector \((F_1(X_1), \ldots, F_d(X_d))\). By the well-known Sklar’s theorem (see Part A of Supplementary Material A), when the \( F_i \)’s are continuous, \( C \) is unique and satisfies

\[ F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)). \]

Let \( c(u_1, \ldots, u_n) \) be the \( d \)-dimensional copula density function and \( f_i \) be the marginal density function of \( X_i \) for \( i = 1, \ldots, d \). The joint density function of \((X_1, \ldots, X_d)\) is

\[ f(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \prod_{i=1}^{d} f_i(x_i). \]
The vine copula is a particular kind of copula that is very flexible and computationally tractable. The density of vine copula can be factored in terms of bivariate linking copulas and lower-dimensional margins; see Joe (2014); Czado (2019). A vine copula is described by a tree sequence on \( d \) elements, namely an ordered set of trees \( V = (\text{Tr}_1, \cdots, \text{Tr}_{d-1}) \), where \( \text{Tr}_i = (N_i, E_i) \) with node set \( N_i \) and edge set \( E_i \) for \( 1 \leq i \leq d-1 \), satisfying

a) \( \text{Tr}_1 \) is the first tree with node set \( N_1 = \{1, \cdots, d\} \) and edge set \( E_1 \).

b) For \( i = 2, \cdots, d-1 \), edge set \( E_{i-1} \) is the node set of tree \( \text{Tr}_i \).

c) For tree \( \text{Tr}_i \) (\( 2 \leq i \leq d-1 \)), if two nodes in \( E_{i-1} \) are connected by an edge in \( E_i \), then these two nodes share the same node in \( E_i \) as edges in \( \text{Tr}_{i-1} \).

In general, a \( d \)-dimensional vine copula is constructed by mixing \( d(d-1)/2 \) bivariate linking copulas on a tree. In this work, we use the D-vine copula, a special kind of vine copula in which the nodes on a tree only connect to their adjacent nodes (Brechmann and Schepsmeier, 2013). Figure 1 depicts the graphical specification of a 5-dimensional \((U_1, \cdots, U_5)\) D-vine in the form of a nested set of tree structures, where \( U_1, \ldots, U_5 \) are uniform random variables.

![Figure 1. Five-dimensional D-vine dependence structure.](image)

A D-vine with five variables has four trees \( \text{Tr}_j \), and tree \( \text{Tr}_j \) has \( 6-j \) nodes and \( 5-j \) edges, where \( 1 \leq j \leq 4 \). Each edge is associated with a pair-copula density used for modeling dependence between two variables, and the edge label represents the dependence parameter in the associated pair-copula density. In Tree 1, there are four pairs of variables, namely \((U_1, U_2), (U_2, U_3), (U_3, U_4)\) and \((U_4, U_5)\); the pair-wise dependencies are modeled by using four copulas \( c_{1,2}, c_{2,3}, c_{3,4}, c_{4,5} \), where \( c_{i,i+1} \) represents the copula density between \( U_i \) and \( U_{i+1} \) with \( 1 \leq i \leq 4 \). In Tree 2, three conditional dependencies are modeled: the one between \( U_1 \) and \( U_3 \) given \( U_2 \) using copula density \( c_{1,3|2} \); the one between \( U_2 \) and \( U_4 \) given \( U_3 \) using copula density \( c_{2,4|3} \); and the one between \( U_3 \) and \( U_5 \) given \( U_4 \) using copula density \( c_{3,5|4} \). In Tree 3, two conditional dependence are modeled: the one between \( U_1 \) and \( U_4 \) given \( U_2 \) and \( U_3 \) using copula density \( c_{1,4|2,3} \); and the one between \( U_2 \) and \( U_5 \) given \( U_3, U_4 \) using copula density \( c_{2,5|3,4} \). In Tree 4, only one conditional dependence is modeled, namely the one between \( U_1 \) and \( U_5 \) given \( U_2, U_3, U_4 \) using copula density \( c_{1,5|2,4} \). As a result, the joint density function of a 5-dimensional D-vine is given by

\[
 f_{1:5}(u_1, u_2, u_3, u_4, u_5) = \prod_{i=1}^{5} f_i(u_i) \prod_{j=2}^{5} \prod_{i=1}^{j-1} c_{i,j|(i+1):(j-1)}(u_i, u_j|u_{i+1}, \cdots, u_{j-1}),
\]

where \( f_i(u_i) \) is marginal density and \( c_{i,j|(i+1):(j-1)} \) is the bivariate copula density. It is worth mentioning that the D-vine structure has been popularly used in modeling the temporal dependence among data due to its special structure; see, for example, Smith (2015); Shi and Yang (2018); Fang et al. (2021). Please refer to Brechmann and Schepsmeier (2013); Czado (2019) for more details on D-vine copulas and their applications.
3. Data preprocessing and preliminary data analysis. The dataset we analyze in this research was obtained from the U.S. Department of Health and Human Services Office for Civil Rights (OCR) (HHS, 2020), which provides a publicly available breach dataset of unsecured protected health information. Although there are some other public sources (e.g., PRC) of healthcare data, the rationales for using the OCR data are as follows: i) The OCR data is the official data published by US government, which is directly related to the federal HIPAA regulations and breach notification rules. This reliable data fits the purpose of our study, which is to provide guidance for government regulators to make/adjust breach regulations for states. ii) The other data sources also collect the healthcare breach data from OCR. In particular, the OCR provides the number of affected individuals in a breach incident, while most of other data sources provide the number of breached records. Since many breaches reported within the last 24 months are currently under investigation by OCR, we consider the breach data from January 1, 2010 to December 31, 2018. After carefully removing redundant records, we have a total of 2,132 organizations reporting 2,473 breach records, with 194,485,233 individuals affected. To study the data breach risk at the state level, we aggregate breach frequency and affected individuals by state for each year. In the following, we perform a preliminary analysis on the processed breach data.

There are two common approaches to frequency modeling: i) The first one is to directly model the original frequency and develop a suitable model for the count data. For example, the classical statistical models for count data are Poisson, geometric, and negative binomial regression models. However, they are often of limited use when count datasets exhibit over-dispersion and/or an excess number of zeros. To model the excess zeros, zero-augmented models are developed by introducing a second model component to capture zero counts (Mullahy, 1986; Lambert, 1992). The most commonly used zero-augmented models are hurdle and zero-inflation type models (Zeileis, Kleiber and Jackman, 2008). Hurdle models combine a left-truncated count component with a right-censored hurdle component. Zero-inflation models take a somewhat different approach: they are mixture models that combine a count component and a point mass at zero (see Part B of Supplementary Material A). ii) The second approach is to model the transformed frequency (namely, square root transformation), and then develop a LMM model for the transformed data. Generally, if the frequency follows a Poisson-like distribution, square root transformation is often a good choice (Maindonald and Braun, 2006).

Table 1 shows the summary statistics of frequency and square root transformed frequency. The frequency has a minimum of 0 and a maximum of 41, and the standard deviation is 6.41. Those indicate that the frequency has relatively large variability and skewness. The transformed frequency has a maximum of 6.40 and standard deviation 1.29, and has a more close mean and median. Hence, the transformation reduces the skewness and the variability of frequency. In the rest of the analysis, the square root transformed frequency, referred to as s-frequency, is used for modeling purposes. The rationale for choosing the s-frequency is that it can result in better prediction results compared with those of using the frequency directly (see Section 5). In addition, this approach has a good interpretability that zeros can be explained as non-reported or non-detected events from the same data generation process (i.e., breach activity).

Figure 2 displays the boxplots of s-frequencies per year and per state, in which the means are indicated by the dots. Figure 2(a) shows the boxplots of s-frequencies across years. In terms of medians, the s-frequencies are relatively stable from 2010 to 2015, and are larger from 2016 to 2018. However, in terms of means, the s-frequency shows an overall increasing trend. This hints that the time effect should be incorporated into the modeling for frequency. Figure 2(b) shows the boxplots of s-frequencies across states where indices 1, · · · , 52 correspond to 50 states, the District of Columbia, and Puerto Rico in the US (see Part C of
DATA BREACH IN HEALTHCARE

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>$Q_5$</th>
<th>$Q_{25}$</th>
<th>Median</th>
<th>$Q_{75}$</th>
<th>$Q_{95}$</th>
<th>Max</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>17.65</td>
<td>41</td>
<td>5.29</td>
<td>6.41</td>
</tr>
<tr>
<td>S-frequency</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1.73</td>
<td>2.65</td>
<td>4.20</td>
<td>6.40</td>
<td>1.91</td>
<td>1.29</td>
</tr>
</tbody>
</table>

**Table 1**

Summary statistics of frequency and square root transformed frequency ("S-frequency"), where "SD" represents the standard deviation.

Supplementary Material A for the full names, abbreviations, and populations of these states, the District of Columbia, and Puerto Rico\(^2\). There are dramatic differences among states. For example, CA (index 5) has the largest mean, 5.52, and median, 5.83, while ID (index 14) has the smallest mean, 0.38, and median, 0, over 9 years. The largest s-frequency is 6.40 in CA in 2016, while the smallest frequency is 0. In fact, 15.17\% of the breach data is 0s. *This suggests that the breach s-frequency exhibits the heterogeneity across different states.* Table 2 shows the temporal dependence of s-frequencies across years. The upper triangular shows Spearman’s $\rho$s across years, and the lower triangle shows Kendall’s $\tau$s across years. All the values are positive and relatively large. This indicates that there is *positive dependence among s-frequencies across years*.

![Boxplots of s-frequencies per year and per state where the dots represent the means. The indices 1, \ldots, 52 correspond to 52 states in the US (see Part C of Supplementary Material A for the full names, abbreviations, and populations of these states).](image)

Table 3 shows the summary statistics of numbers of affected individuals (AIs). The number of AIs shows the extreme variability and skewness. Since the minimum positive number of AIs is 500\(^3\), the positive numbers of AIs are log-transformed to reduce the variability and skewness, which is referred to as *severity* in the following analysis. When the severity is 0, it indicates that there is no breach incident with 500 or more AIs\(^4\). It is seen that severity is much more symmetric, but the variability is still large, as indicated by the large sample

\(^2\)For the sake of brevity, 50 states, the District of Columbia, and Puerto Rico are collectively called ‘states’ in our discussion.

\(^3\)The minimum number of 500 is based on Section 13402(e)(4) of the Health Information Technology for Economic and Clinical Health Act: the Secretary must post a list of breaches of unsecured protected health information affecting 500 or more individuals.

\(^4\)Only incidents affecting 500 or more individuals are reported to the OCR.
An upper triangle table shows Spearman's $\rho$s and the lower triangle table shows Kendall's $\tau$s of s-frequencies across years. This suggests that a mixed distribution may be needed to model the severity.

Table 2 shows the temporal dependence of severities across years. It can be observed that all values are positive and relatively large. This hints that there is positive dependence among severities across years.

Figure 4(a) shows the boxplots of severities across years. It can be seen that the severities are more or less similar across years. Figure 4(b) displays the severities for states, which are quite different. Specifically, we observe that there are several states with smaller severities. These can be compared to their frequencies in Figure 2. When the frequency is small, the corresponding severity is also small. This suggests that the frequency and the severity are positively dependent.

Table 4 shows the temporal dependence of severities across years. The upper triangular shows Spearman’s $\rho$s while the lower triangle shows Kendall’s $\tau$s across years. It can be observed that all values are positive and relatively large. This hints that there is positive dependence among severities across years.

To summarize, we observe that there are time effects, heterogeneity, and positive dependence among s-frequencies. In particular, the s-frequency has 15.17% 0s. The severity and s-frequency exhibit positive dependence, and the severities are positively correlated across years. All of those observed characteristics motivate us to develop the sequel multivariate frequency-severity model.
4. **Multivariate frequency-severity framework.** The multivariate frequency-severity model has been popularly studied in the domain of actuarial science and insurance. For example, Cummins and Wiltbank (1983) presented a model that utilizes multivariate frequency and severity distributions to analyze total annual losses when multiple loss processes are present. Unlike the traditional approach, the multivariate technique allows explicit recognition of dependence among the components of the overall process. Frees, Lee and Yang (2016) extended the literature on multivariate frequency-severity regression by using copula for modeling the dependence among number of claims and aggregate claim amount. In nonlife insurance, Lee and Shi (2019) proposed a dependent frequency-severity model to jointly examine frequency and severity in a longitudinal context. Their proposed model can accommodate temporal correlations within both frequency and severity. They also modeled the association between frequency and severity by using a copula regression. For more discussion of the multivariate frequency-severity models, one may refer to Shi, Feng and Ivantsova (2015); Frees, Lee and Yang (2016); Shi and Yang (2018); Sun, Xu and Zhao (2020); Shi et al. (2020). In the fol-
\begin{table}[h]
\centering
\begin{tabular}{c|c|cc|cc|cc|cc|cc|cc}
\hline
\hline
\multicolumn{1}{c|}{\multirow{2}{*}{$\tau$}} & \multicolumn{1}{c|}{$\rho$} & \multicolumn{5}{c}{Year 1} & \multicolumn{5}{c}{Year 2} & \multicolumn{5}{c}{Year 3} & \multicolumn{5}{c}{Year 4} & \multicolumn{5}{c}{Year 5} & \multicolumn{5}{c}{Year 6} & \multicolumn{5}{c}{Year 7} & \multicolumn{5}{c}{Year 8} & \multicolumn{5}{c}{Year 9} \\
\hline
Year 1 & \multicolumn{1}{c|}{-} & 0.42 & 0.60 & 0.57 & 0.57 & 0.54 & 0.44 & 0.51 & 0.60 \\
Year 2 & 0.31 & \multicolumn{1}{c|}{-} & 0.54 & 0.34 & 0.61 & 0.52 & 0.49 & 0.46 & 0.26 \\
Year 3 & 0.44 & 0.41 & \multicolumn{1}{c|}{-} & 0.58 & 0.65 & 0.57 & 0.58 & 0.44 & 0.54 \\
Year 4 & 0.41 & 0.24 & 0.42 & \multicolumn{1}{c|}{-} & 0.59 & 0.55 & 0.48 & 0.41 & 0.55 \\
Year 5 & 0.43 & 0.44 & 0.46 & 0.41 & \multicolumn{1}{c|}{-} & 0.58 & 0.51 & 0.43 & 0.50 \\
Year 6 & 0.38 & 0.39 & 0.42 & 0.38 & 0.44 & \multicolumn{1}{c|}{-} & 0.53 & 0.59 & 0.54 \\
Year 7 & 0.31 & 0.36 & 0.44 & 0.33 & 0.37 & 0.38 & \multicolumn{1}{c|}{-} & 0.59 & 0.60 \\
Year 8 & 0.35 & 0.32 & 0.32 & 0.29 & 0.30 & 0.41 & 0.42 & \multicolumn{1}{c|}{-} & 0.61 \\
Year 9 & 0.42 & 0.19 & 0.40 & 0.40 & 0.36 & 0.40 & 0.43 & 0.43 & \multicolumn{1}{c|}{-} \\
\hline
\hline
\end{tabular}
\caption{The upper triangle table shows Spearman’s $\rho$'s and the lower triangle table shows Kendall’s $\tau$'s of severities across years.}
\end{table}

Following, we develop a novel multivariate frequency-severity framework based on LMM and D-vine copula to accommodate the characteristics exhibited by the healthcare breach data.

For state $i$, let $N_{i,t}$ and $S_{i,t}$ denote the s-frequency and severity of reported breach incidents in $t$th year, respectively. Then, the healthcare breach data can be represented as $(N_{i,t}, S_{i,t})$, $i = 1, \ldots, M, t = 1, \ldots, T$.

Let $N = (N_1, N_2, \ldots, N_T)$ and $S = (S_1, S_2, \ldots, S_T)$ be the matrices of s-frequency and severity, where $N_t = (N_{1,t}, \ldots, N_{M,t})^\top$, and $S_t = (S_{1,t}, \ldots, S_{M,t})^\top$, $t = 1, \ldots, T$. Then, the joint density function of s-frequency and severity can be represented as

$$f_{N,S}(n, s) = f_N(n)f_{S|N}(s|n).$$

In the following, we discuss how to model the marginal distributions of s-frequency and severity, and accommodate the relevant dependence via copula.

4.1. Modeling marginal distributions. As seen from Section 3, the healthcare breach s-frequency, it not only includes 0s but also varies across years and states. Therefore, we propose to use the LMM for modeling the s-frequency. Specifically, it is assumed that

$$(N_{i,t}, S_{i,t}), i = 1, \ldots, M, t = 1, \ldots, T.$$

Let $N = (N_1, N_2, \ldots, N_T)$ and $S = (S_1, S_2, \ldots, S_T)$ be the matrices of s-frequency and severity, where $N_t = (N_{1,t}, \ldots, N_{M,t})^\top$, and $S_t = (S_{1,t}, \ldots, S_{M,t})^\top$, $t = 1, \ldots, T$. Then, the joint density function of s-frequency and severity can be represented as

$$f_{N,S}(n, s) = f_N(n)f_{S|N}(s|n).$$

In the following, we discuss how to model the marginal distributions of s-frequency and severity, and accommodate the relevant dependence via copula.

4.1. Modeling marginal distributions. As seen from Section 3, the healthcare breach s-frequency, it not only includes 0s but also varies across years and states. Therefore, we propose to use the LMM for modeling the s-frequency. Specifically, it is assumed that

$$N_{i,t} | b_i \sim N(\mu_{i,t}, \sigma^2_{\epsilon}),$$

$$\mu_{i,t} = x_i^\top \alpha + z_i^\top b_i,$$

where $\alpha$ and $b_i$ represent the fixed and random effects, respectively, $i = 1, \ldots, M, t = 1, \ldots, T$.

The preliminary analysis shows that zeros, skewness, and large values exist. Therefore, a natural idea is to consider a two-component mixture model to accommodate the mass probability at zero, the skewness, and the heavy tail of the distribution. Specifically, $S_{i,t}$ is assumed to be generated from a degenerate distribution at zero, with probability $p_{i,t}$ and a skewed and heavy-tailed distribution $G_{i,t}(\cdot)$ defined on $(0, \infty)$ with probability $1 - p_{i,t}$. Let $F_{S_{i,t}}(\cdot)$ and $f_{S_{i,t}}(\cdot)$ denote the distribution function and density function, respectively. Then, we have the following mixed distribution

$$F_{S_{i,t}}(s) = p_{i,t}1\{s = 0\} + (1 - p_{i,t})G_{i,t}(s),$$

$$f_{S_{i,t}}(s) = p_{i,t}\delta(0) + (1 - p_{i,t})g_{i,t}(s).$$
where $\mathbb{1}\{\cdot\}$ is the indicator function, $\delta(0)$ is the Dirac delta function, and $g_{i,t}(\cdot)$ is the density function associated with $G_{i,t}(\cdot)$. The mixture model in Eq. (4.3) has been extensively discussed in the literature; see, for example, Frees, Lee and Yang (2016); Eling and Jung (2018); Shi and Yang (2018); Lee and Shi (2019).

In the following, instead of modeling the severity by using the mixture model, we propose to use the following ‘pseudo’ mixture distribution. That is,

$$
F_{S_{i,t}}(s) = \begin{cases} 
H_{i,t}(s), & s < 1, \\
pi,t + (1 - pi,t)G_{i,t}^*(s), & s \geq 1,
\end{cases}
$$

(4.4)

where $H_{i,t}(s) = pi,tH(s)$ where $H(s)$ is a uniform distribution over $[0,1]$, and $G_{i,t}^*(s) = G_{i,t}(s|S_{i,t} \geq 1)$ is a left truncated distribution at $s = 1$. Then, the density function is

$$
f_{S_{i,t}}(s) = \begin{cases} 
pi,t, & s < 1, \\
(1 - pi,t)g_{i,t}^*(s), & s \geq 1,
\end{cases}
$$

(4.5)

where $g_{i,t}^*(s)$ is the density function of $G_{i,t}^*(s)$. That is, we replace zeros by randomly sampled values from $H(s)$ which does not change the essential structure of analyzed data. The main advantage of the ‘pseudo’ mixture approach is that the distribution is continuous which not only makes the relevant statistical inferences easier but also can take full advantage of multivariate dependence structure among severities.

4.2. Modeling multivariate dependence. Since we aim to model and predict the severity given the s-frequency, we discuss the conditional distribution and conditional density functions of severity in the following.

Assume that for state $i$, $i = 1, \ldots, M$, the severity in year $t$ only depends on the s-frequency in year $t$, $t = 1, \ldots, T$. Then, the conditional distribution of severity can be represented as

$$
F_{S_{i,t}|N_{i,t}}(s_i|n_i) = \mathbb{P}(S_i \leq s_i|N_i = n_i) \\
= \mathbb{P}(S_{i1} \leq s_{i1}, \ldots, S_{iT} \leq s_{iT}|N_1 = n_{i1}, \ldots, N_T = n_{iT}) \\
= C^S(F_{S_{i,t}|N_{i,t}}(s_{i1}|n_{i1}), \ldots, F_{S_{i,T}|N_{i,T}}(s_{iT}|n_{iT})),
$$

(4.6)

for $i = 1, \ldots, M$, where the copula $C^S$ represents the copula among the conditional severities.

Note that the joint distribution of $S_{i,t}$ and $N_{i,t}$ is

$$
F_{S_{i,t},N_{i,t}}(s_{i,t}, n_{i,t}) = \mathbb{P}(S_{i,t} \leq s_{i,t}, N_{i,t} \leq n_{i,t}) \\
= \begin{cases} 
H_{i,t}(s_{i,t}), & n_{i,t} = 0, \\
(1 - F_{N_{i,t}}(0))C^FS(F_{N_{i,t}}(n_{i,t}|N_{i,t} > 0), F_{S_{i,t}}(s_{i,t}|N_{i,t} > 0)), & n_{i,t} > 0,
\end{cases}
$$

(4.7)

where $C^FS$ is a bivariate copula to formulate the dependence between conditional s-frequency and severity.

Thus, the conditional distribution $S_{i,t}$ given $N_{i,t}$ can be represented as

$$
F_{S_{i,t}|N_{i,t}}(s_{i,t}|n_{i,t}) = \frac{\mathbb{P}(S_{i,t} \leq s_{i,t}|N_{i,t} = n_{i,t})}{f_{N_{i,t}}(n_{i,t})} \\
= \begin{cases} 
H(s_{i,t}), & n_{i,t} = 0, \\
(1 - F_{N_{i,t}}(0))C^FS(F_{N_{i,t}}(n_{i,t}|N_{i,t} > 0), F_{S_{i,t}}(s_{i,t}|N_{i,t} > 0)), & n_{i,t} > 0,
\end{cases}
$$

(4.8)

where $C^FS = \frac{\partial}{\partial u_2}C^FS(u_1, u_2)$ is the partial derivative of copula function $C^FS(u_1, u_2)$. The density function can be represented as

$$
f_{S_{i,t}|N_{i,t}}(s_{i,t}|n_{i,t}) = \begin{cases} 
1, & n_{i,t} = 0, \\
(1 - F_{N_{i,t}}(0))f_{S_{i,t}}(s_{i,t}|N_{i,t} > 0)c^FS(F_{N_{i,t}}(n_{i,t}|N_{i,t} > 0), F_{S_{i,t}}(s_{i,t}|N_{i,t} > 0)), & n_{i,t} > 0,
\end{cases}
$$

(4.9)
where \( c^{FS} = \frac{\partial^2}{\partial u_1 \partial u_2} C^{FS}(u_1, u_2) \) is the probability density function of the copula function \( C^{FS}(u_1, u_2) \).

Then the joint probability density function of \( S_i \) given \( N_i \) can be rewritten as

\[
f_{S_i|N_i}(s_i|n_i) = \prod_{t=1}^{T} f_{S_i|N_i}(s_{i,t}|n_{i,t}) c_{i,T}^{S}(F_{S_i|N_i}(s_{i,1}|n_{i,1}), \ldots, F_{S_i|N_i}(s_{i,T}|n_{i,T}))
\]

for \( i = 1, \ldots, m \), where \( c_{i,T}^{S} \) is defined as

\[
c_{i,T}^{S}(u_1, \ldots, u_T) = \frac{\partial^T}{\partial u_1, \ldots, \partial u_T} C^S(u_1, \ldots, u_T).
\]

Since the vine copula is very flexible, we propose to model \( C^S \) by using a D-vine structure. For the notation simplicity, denote \( Z_i = (Z_{i,1}, \ldots, Z_{i,T})^T \), where \( Z_{i,t} = S_{i,t}|N_{i,t}, i = 1, \ldots, M, t = 1, \ldots, T \). Then, we have

\[
f_{Z_i}(z_i) = f_i(z_{i,T}|z_{i,T-1}, \ldots, z_{i,1}) \cdots f_i(z_{i,2}|z_{i,1}) \cdot f_i(z_{i,1})
\]

\[
= \prod_{t=1}^{T} f_{i}(z_{i,t}) \prod_{t=1}^{T} \prod_{s=1}^{t-1} f_{i,s,t|s,t}(z_{i,s}, z_{i,t}|z_{i,s,t};(t-1)),
\]

where \( z_{i,t} = s_{i,t}|n_{i,t} \) represents the realization of \( Z_{i,t} \) (Joe, 2014; Brechmann and Scheipers, 2013; Shi and Yang, 2018), \( f_i(\cdot) \) is defined in Eq. (4.9), and

\[
\tilde{f}_{i,s,t|s,t}(z_{i,s}, z_{i,t}|z_{i,s,t};(t-1)) = c_{s,t}(s+1):(t-1) \left( F_{i,s}(s+1):(t-1), F_{i,s}(s+1):(t-1) \right),
\]

and \( c_{s,t}(s+1):(t-1)(u_1, u_2) \) is the bivariate copula density function associated with conditional distributions \( F_{i,s}(s+1):(t-1)(z_{i,s}|z_{i,s+1};(t-1)) \) and \( F_{i,s}(s+1):(t-1)(z_{i,t}|z_{i,s+1};(t-1)) \) given the realizations of \( (z_{i,s+1}, \ldots, z_{i,t-1}) \). It should be noted that when \( s = t-1 \), it represents the unconditional density, i.e.,

\[
\tilde{f}_{i,t-1,t|t-1}(z_{i,t-1}, z_{i,t}|z_{i,t};(t-1)) \equiv \tilde{f}_{i,s,t}(z_{i,t-1}, z_{i,t}).
\]

Further,

\[
F_{i,s}(s+1):(t-1)(z_{i,s}|z_{i,s+1};(t-1)) = c_{2,s,t-1}(s+1):(t-2) \left( F_{i,s}(s+1):(t-2), F_{i,s}(s+1):(t-2) \right),
\]

and

\[
F_{i,t|s}(s+1):(t-1)(z_{i,t}|z_{i,s+1}, \ldots, z_{i,t-1}) = c_{1,s+1,t}(s+2):(t-1) \left( F_{i,s+1}(s+2):(t-1), F_{i,t|s}(s+2):(t-1) \right),
\]

where

\[
c_{k,s,t}(s+1):(t-1)(u_1, u_2) = \frac{\partial}{\partial u_k} C_{s,t}(s+1):(t-1)(u_1, u_2), \quad k = 1, 2.
\]
4.3. Model estimation and prediction. Due to the parametric nature of the proposed model, we employ the likelihood-based method for estimation. The total log-likelihood function for the conditional severity is

\[
ll(\theta, \eta) = \sum_{i=1}^{M} \sum_{t=1}^{T} \log f_{Z_{i,t}}(z_{i,t}) + \sum_{i=1}^{M} \sum_{t=2}^{T-1} \sum_{s=1}^{T} \hat{f}_{s,t}(s_{i+1};(t-1)) (z_{i,s}, z_{i,t} | \tilde{z}_{i,(s+1):(t-1)}) ,
\]

where \( f_{Z_{i,t}}(z_{i,t}) \equiv f_{S_{i,t}|N_{i,t}}(s_{i,t}|n_{i,t}) \) is specified by Eq. (4.9), \( \hat{f}_{s,t}(s_{i+1};(t-1)) (\cdot | \cdot) \) is specified by Eq. (4.13), and \( \theta \) and \( \eta \) summarize the parameters in marginals and the D-vine, respectively. In the literature, two commonly used methods to estimate parameters are the joint maximum likelihood estimation (MLE) and the IFM approaches (Joe, 2014). The joint MLE is a full likelihood approach and estimates all model parameters simultaneously. The IFM approach has two steps: (i) estimation of the parameters of the marginal models (\( \theta \)) by assuming independence, and (ii) estimation of the parameters of the D-vine copula (\( \eta \)) by fixing the parameters obtained in step (i). Note that although the IFM may lose some statistical efficiency, it is much more computationally efficient than the joint MLE. In general, the IFM approach has good efficiency except for possible extreme dependence near the Frechet bounds (Joe, 2005). Since the IFM approach is especially convenient for the copula selection when marginals are fixed and much easier when the total number of parameters is large, it is favored in practice. In our study, the dependence is not near the Frechet bounds, and hence, statistical efficiency is not a concern. This is also confirmed in Part D of Supplementary Material A, where we conduct the comparison of the joint MLE and the IFM approaches for our data. Therefore, we use the IFM approach for the parameter estimations.

To predict the severity, we use the proposed D-vine model which has a natural temporal structure. Denote \( \pi_{i,t} = F_{S_{i,t}|N_{i,t}}(s_{i,t}|n_{i,t}) \), \( t = 1, \ldots, T \), \( S_{i,1:t} = (S_{i,1}, \ldots, S_{i,t}) \top \), and \( N_{i,1:t} = (N_{i,1}, \ldots, N_{i,t}) \top \), \( i = 1, \ldots, M \). Then, the conditional distribution of \( S_{i,t+1} \) given \( (S_{i,1:t}, N_{i,1:(t+1)}) \) can be represented as

\[
F_{S_{i,t+1}|S_{i,1:t},N_{i,1:(t+1)}}(s|s_{i,1:t}, n_{i,1:(t+1)}) = \frac{\mathbb{P}(S_{i,t+1} = s_{i,1:t} | N_{i,1:(t+1)} = n_{i,1:(t+1)})}{\mathbb{P}(S_{i,1:t} = s_{i,1:t} | N_{i,1:(t+1)} = n_{i,1:(t+1)})}
\]

\[
= \frac{\mathbb{P}(S_{i,t+1} \leq s, S_{i,1:t} = s_{i,1:t} | N_{i,1:(t+1)} = n_{i,1:(t+1)})}{\mathbb{P}(S_{i,1:t} = s_{i,1:t} | N_{i,1:t} = n_{i,1:t})}
\]

\[
= \frac{c_{\pi_{i,t}}(\pi_{i,t}, \ldots, \pi_{i,t}, F_{S_{i,t+1}|N_{i,1:(t+1)}}(s|n_{i,t+1}))}{c^S(\pi_{i,1}, \ldots, \pi_{i,t})},
\]

where \( c_{\pi_{i,t}}(u_{1}, \ldots, u_{t}, u_{t+1}) = \frac{\partial^t}{\partial u_1 \cdots \partial u_t} C^S(u_{1}, \ldots, u_{t+1}) \), and \( c^S(u_{1}, \ldots, u_{t}) \) is the copula density corresponding to \( t \) dimensional copula \( C^S(u_{1}, \ldots, u_{t}) \). The associated conditional density function is

\[
f_{S_{i,t+1}|S_{i,1:t},N_{i,1:(t+1)}}(s|s_{i,1:t}, n_{i,1:t+1}) = \frac{\partial}{\partial s} F_{S_{i,t+1}|S_{i,1:t},N_{i,1:(t+1)}}(s|s_{i,1:t}, n_{i,1:t+1})
\]

\[
= f_{S_{i,t+1}|N_{i,1:(t+1)}}(s|n_{i,t+1}) \frac{c^S(\pi_{i,1}, \ldots, \pi_{i,T}, F_{S_{i,T+1}|N_{i,T+1}}(s|n_{i,T+1}))}{c^S(\pi_{i,1}, \ldots, \pi_{i,T})}.
\]

Since it is infeasible to derive the analytic results for the prediction, we propose to use the rejection sampling technique to simulate the predictive distribution (Robert and Casella,
The rejection sampling method generates sampling values from a target distribution $X$ with an arbitrary probability density function $f(x)$ by using an alternative distribution $Y$ with probability density $g(x)$. The basic idea is that one can generate a sample value from $X$ by sampling from $Y$ and accepting the sample from $Y$ with probability $f(x)/(cg(x))$, where $c$ is a scalar such that $f(x) \leq cg(x)$ for all values of $x$. Algorithm 1 presents the details of simulating the predictive distributions of breach s-frequency and severity. In our study, the targeted density is Eq. (4.20), and the average acceptance rate is 33.15%.

**Algorithm 1 Simulating predictive distributions of breach s-frequency and severity**

**INPUT:** In-sample breach data $(n_{i,t}, s_{i,t})$, $i = 1, \ldots, M$, $t = 1, \ldots, T_1$; Out-of-sample breach data $(n_{i,t}, s_{i,t})$, $i = 1, \ldots, M$, $t = T_1 + 1, \ldots, T_2$; Number of simulations $r = 5,000$.

1. for $t$ in $T_1 : (T_2 - 1)$ do
2. for $i$ in $1 : M$ do
3. Predict the s-frequency mean $\mu_{i,t+1}$ based on LMM in Section 4.1;
4. Simulate s-frequencies $\{n_{i,t+1,h} | h = 1, \ldots, r \}$ based $\mu_{i,t+1}$ via Eq. (4.2);
5. for $h$ in $1 : r$ do
6. if $n_{i,t+1,h} \leq 0$ then
7. Set s-frequency $n_{i,t+1,h} = 0$ and severity $s_{i,t+1,h} = 0$;
8. else
9. Randomly sample 1,000 values $\{s_{w,i,j} | w = 1, \ldots, 1,000 \}$ from $[a, b]$ where $a_i = \min\{s_{i,t} | s_{i,t} > 0, t = 1, \ldots, T_1 \}$ and $b_i = \max\{s_{i,t} | s_{i,t} > 0, t = 1, \ldots, T_1 \}$;
10. Compute $f_{i,j}^w = \max\{f(s_{w,i,j}) | w = 1, \ldots, 1,000 \}$, where $f(\cdot)$ is the conditional density in Eq. (4.20);
11. Randomly sample a value $s$ from $[a_i, b_i]$;
12. Randomly generate an observation $u$ from $U(0,1)$;
13. if $u < f(s)/f_{i,j}^w$ where $f(\cdot)$ is the conditional density in Eq. (4.20); then
14. Accept this sample and set $s_{i,t+1,h} = s$;
15. else
16. Reject this sample and repeat the sampling procedure from line 11 to line 17;
17. end if
18. end if
19. end for
20. return $\{(n_{i,t+1,h}, s_{i,t+1,h})\}_i=1,\ldots,M_2; t=T_1+1,\ldots,T_2; h=1,\ldots,r$.

4.4. **Prediction assessment.** For the prediction evaluation, the common metrics such as mean square error only provide limited information on the performance of predictive distribution. To evaluate the predictive distribution, which is more informative, we propose the following two approaches:

- **Predictive likelihood.** The predictive likelihood function assesses the likelihood of observed values based on the predictive density. It has been widely used in the literature for the evaluation of density forecasting (Hinkley et al., 1979; Tay and Wallis, 2000; González-Rivera, Lee and Mishra, 2004). We compute the following predictive log-likelihood (PLL) as a metric to assess the predictive ability,

\[
(4.21) \quad PLL = \sum_{i=1}^{M} \log f_{s_{i,t+1}|S_{i,1:t},N_{i,1:t+1}}(s_{i,t+1}|s_{i,t},n_{i,1:t+1}),
\]

where $s_{i,t+1}$ is the observed severity for state $i$ at year $t+1$. 

Scoring rule. The second metric to evaluate the probability forecasting is the scoring rule, which assigns a numerical score based on the predicted distribution and observations. The scoring rule evaluates the accuracy of a forecast distribution given the observations. Conceptually, the scoring rule can be considered an error measures for distribution functions. The continuous ranked probability score (CRPS) (Matheson and Winkler, 1976) is defined in terms of the predictive CDF \( F \) and is given by

\[
\text{CRPS}(F, x) = \int_{\mathbb{R}} (F(z) - \mathbb{1}\{x \leq z\})^2 \, dz,
\]

where \( \mathbb{1}\{x \leq z\} \) denotes the indicator function. In our study, the predictive distribution of interest \( F \) is not available in an analytic form, but only through a simulated sample \( X_1, \ldots, X_m \sim F \). Therefore, we use the empirical distribution

\[
\hat{F}_m(z) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{X_j \leq z\},
\]

which is a natural approximation of the predictive CDF. In this case, the CRPS reduces to

\[
\text{CRPS}(\hat{F}_m(z), y) = \frac{1}{m} \sum_{j=1}^{m} |X_j - y| - \frac{1}{2m^2} \sum_{j=1}^{m} \sum_{k=1}^{m} |X_j - X_k|,
\]

which allows us to compute the CRPS directly from the simulated sample; see Grimit et al. (2006).

5. Empirical study. In this section, we analyze the healthcare breach data in Section 3 by following the framework developed in the previous section. We use the first 5 years as the in-sample data, and the remaining 4 years as the out-of-sample data for assessing the prediction performance. The rolling approach is used for assessing the fitting and prediction performances for each year (i.e., \( T = 5, \ldots, 8 \)). There are two reasons to use the rolling evaluation: i) Since the empirical data is at the state level, there are only 52 observations for each year. The rolling approach can provide more data for prediction assessment. ii) The rolling approach can clearly show the robustness of proposed model without limiting it to a specific year. That is, the robustness is reflected in the consistent satisfactory fitting and prediction performance in all years, as shown below. This further confirms the superiority of developed model.

5.1. Modeling the frequency. For the s-frequency modeling, we consider the following mixed effects for the states and years, i.e.,

\[
N_{i,t} | M_i \sim N(\mu_{i,t}, \sigma^2_e) \\
\mu_{i,t} = \alpha_0 + \alpha_1 t + \alpha_2 c_i + M_i,
\]

where \( c_i \) is the covariate of the state population (unit: 100,000) based on 2010 census data, with \( i = 1, \ldots, 52, t = 1, \ldots, T, T = 5, \ldots, 8 \). \( \alpha_0, \alpha_1, \) and \( \alpha_2 \) are the fixed effects, and \( M_i \sim N(0, \sigma^2_M) \) represents the random effect for the \( i \)th state.

Table 5 shows the estimated parameters and standard errors for the fixed effects and random effects for each time \( T, T = 5, \ldots, 8 \). It can be seen that the fixed effects are all very significant, and nonzero values of variances for the random effects warrant incorporating random effects.

To assess the goodness of fit, we adopt the simulation-based approach used in Hartig (2020) to create the scaled residuals and perform the statistical tests. The basic procedure
<table>
<thead>
<tr>
<th>$T$</th>
<th>$\alpha_0$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\sigma_M$</th>
<th>$\sigma_e$</th>
<th>KS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.55*</td>
<td>0.11*</td>
<td>0.01*</td>
<td>0.42</td>
<td>0.63</td>
<td>0.44</td>
</tr>
<tr>
<td>6</td>
<td>0.60*</td>
<td>0.08*</td>
<td>0.01*</td>
<td>0.41</td>
<td>0.62</td>
<td>0.38</td>
</tr>
<tr>
<td>7</td>
<td>0.56*</td>
<td>0.09*</td>
<td>0.02*</td>
<td>0.42</td>
<td>0.62</td>
<td>0.14</td>
</tr>
<tr>
<td>8</td>
<td>0.54*</td>
<td>0.09*</td>
<td>0.02*</td>
<td>0.41</td>
<td>0.62</td>
<td>0.11</td>
</tr>
</tbody>
</table>

**Table 5**

Estimated parameters for $T = 5, \ldots, 8$. ‘KS’ represents the $p$-value of Kolmogorov-Smirnov test. * represents the significance at 0.01.

is 1) simulating data from the fitted model for each observation; 2) computing the cumulative empirical density function for the simulated observations; and 3) calculating the values of cumulative empirical density function for the observed data, which are named the scaled residuals. In a correctly specified model, it is expected that the scaled residuals will follow a uniform distribution, which is tested by the Kolmogorov-Smirnov (KS) test. Table 5 shows the $p$-values of the KS test, and it can be seen that the $p$-values are all no less than 0.11. This indicates that the fitting performances are satisfactory for all $T$’s. The difference between the expected and observed residuals of LMM for each $T$ is displayed in Figure 5. We observe that the differences are all very small, which further confirms the satisfactory fitting performances of the proposed model.

![Figure 5](image)

*Fig 5. Difference between expected and observed residuals of LMM fitting of healthcare breach data for $T = 5, \ldots, 8$.*

For the illustration, Figure 6(a) shows the fitting s-frequencies versus the observed s-frequencies for $T = 8$. It is seen that the fitting can capture the observed pattern well. Figure 6(b) shows the fitting performances of some representative states across different years. It is seen that the fitted lines generally follow the observed trends very well.
5.1.1. Model comparison. To further compare the fitting performance of proposed model to those of other models, we consider various mixed effects models in Table 6, where M4 is the proposed model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Formula</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>$\alpha_0 + M_i$</td>
<td>Random effect among states with fixed mean effect</td>
</tr>
<tr>
<td>M2</td>
<td>$\alpha_0 + \alpha_1 c_i + M_i$</td>
<td>Random effect among states with fixed population and mean effects</td>
</tr>
<tr>
<td>M3</td>
<td>$\alpha_0 + \alpha_2 t + M_i$</td>
<td>Random effect among states with fixed year and mean effects</td>
</tr>
<tr>
<td>M4</td>
<td>$\alpha_0 + \alpha_1 t + \alpha_2 c_i + M_i$</td>
<td>Random effect among states with fixed year, population, and mean effects</td>
</tr>
<tr>
<td>M5</td>
<td>$\alpha_0 + M_i + T_t$</td>
<td>Random effects among states and years with fixed mean effect</td>
</tr>
<tr>
<td>M6</td>
<td>$\alpha_0 + \alpha_1 c_i + M_i + T_t$</td>
<td>Random effects among states and years with fixed population and mean effects</td>
</tr>
</tbody>
</table>

Table 6
Various models with different fixed and random effects, where $T_t$ represents the random effect among years, and $c_i$ represents the covariate of state population (unit: 100,000).

Table 7 shows Bayesian Information Criterions (BICs) of those models for each $T$. It is seen that M4 has the smallest BIC for each $T$. Therefore, the proposed mixed model is superior to the others.

<table>
<thead>
<tr>
<th>BIC</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M5</th>
<th>M6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 5$</td>
<td>669.33</td>
<td>605.64</td>
<td>665.88</td>
<td><strong>602.19</strong></td>
<td>668.72</td>
<td>605.03</td>
</tr>
<tr>
<td>$T = 6$</td>
<td>772.80</td>
<td>703.85</td>
<td>770.96</td>
<td><strong>702.01</strong></td>
<td>772.26</td>
<td>703.31</td>
</tr>
<tr>
<td>$T = 7$</td>
<td>888.87</td>
<td>817.55</td>
<td>874.17</td>
<td><strong>802.85</strong></td>
<td>879.20</td>
<td>807.88</td>
</tr>
<tr>
<td>$T = 8$</td>
<td>1014.05</td>
<td>939.95</td>
<td>984.50</td>
<td><strong>910.40</strong></td>
<td>993.59</td>
<td>919.49</td>
</tr>
</tbody>
</table>

Table 7
BICs of various mixed models in Table 6 for $T = 5, \ldots, 8$.

For comparison purposes, we consider the commonly used zero-inflated negative binomial (ZINB), hurdle negative binomial (HNB), zero-inflated Poisson (ZIP), and hurdle Poisson
(HP) regression models with the covariate of state population (i.e., $c_i$) and the year (i.e., $t_i$). We use three popular metrics, MSE (mean square error), SMAPE (symmetric mean absolute percentage error), and PMAD (percent mean absolute deviation) to compare the performance:

$$\text{MSE}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

$$\text{SMAPE}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|},$$

$$\text{PMAD}(y, \hat{y}) = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i|},$$

where $y_i$ is the real value, and $\hat{y}_i$ is the predicted value, $i = 1, \ldots, n$. The fitting performances of various models are presented in Table 8. It can be seen that the proposed model M4 has the smallest MSEs, SMAPEs, and PMADs for all the years. This confirms that the proposed model has a better fitting performance. In Figure 7, the QQ-plots of those traditional models for $T = 8$ are displayed (The QQ-plots for other years are similar. See Part E of Supplementary Material A for those plots). It is seen that the traditional modeling approaches cannot provide satisfactory fittings, particularly in the tails.

<table>
<thead>
<tr>
<th></th>
<th>M4</th>
<th>ZINB</th>
<th>HNB</th>
<th>ZIP</th>
<th>HP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MSE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$T = 5$</td>
<td>5.39</td>
<td>25.11</td>
<td>28.54</td>
<td>9.67</td>
<td>9.69</td>
</tr>
<tr>
<td>$T = 6$</td>
<td>5.44</td>
<td>29.02</td>
<td>32.77</td>
<td>10.31</td>
<td>10.33</td>
</tr>
<tr>
<td>$T = 7$</td>
<td>5.42</td>
<td>29.27</td>
<td>32.33</td>
<td>10.42</td>
<td>10.44</td>
</tr>
<tr>
<td>$T = 8$</td>
<td>5.52</td>
<td>31.92</td>
<td>34.60</td>
<td>11.21</td>
<td>11.23</td>
</tr>
<tr>
<td><strong>SMAPE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$T = 5$</td>
<td>0.67</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>$T = 6$</td>
<td>0.66</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>$T = 7$</td>
<td>0.65</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>$T = 8$</td>
<td>0.63</td>
<td>0.69</td>
<td>0.68</td>
<td>0.69</td>
<td>0.69</td>
</tr>
<tr>
<td><strong>PMAD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$T = 5$</td>
<td>0.34</td>
<td>0.53</td>
<td>0.54</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>$T = 6$</td>
<td>0.34</td>
<td>0.54</td>
<td>0.55</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>$T = 7$</td>
<td>0.32</td>
<td>0.52</td>
<td>0.53</td>
<td>0.45</td>
<td>0.45</td>
</tr>
<tr>
<td>$T = 8$</td>
<td>0.32</td>
<td>0.51</td>
<td>0.52</td>
<td>0.44</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Table 8

*Fitted MSEs, SMAPEs, PMADs of proposed LMM (M4), zero-inflated negative binomial (ZINB), hurdle negative binomial (HNB), zero-inflated Poisson (ZIP), and hurdle Poisson (HP) regression models, for $T = 5, \ldots, 8$.*

It should be remarked that the other possible approach to modeling the frequency is via generalized zero-inflated/hurdle type regression models with random effects (Brooks et al., 2017). In Part F of Supplementary Material A, we have also conducted the comparison of the proposed model with generalized zero-inflated/hurdle type regression models with random effects. The results show that the zero-inflated/hurdle type with random effect models have slightly better fitting performance than that of M4. However, M4 outperforms those models in terms of prediction in most of years. Since accurate prediction is more important in our study, M4 is recommended for modeling the breach frequency. The synthetic data study in Section H of Supplementary Material A also confirms the superior performance of M4. Please see Part F and Part H of Supplementary Material A for detailed comparison and discussion.
5.2. Modeling the severity. From the preliminary analysis, it is seen that for the number of affected individuals, a skewed and heavy tail distribution is desired. Therefore, for the nonzero severity (i.e., log-transformed number of affected individuals), we compare the fitting performances of various distributions. Those distributions include gamma, normal, Weibull, and mixed distributions. For the mixed distributions, we consider fitting the upper tail by using the generalized Pareto distribution (GPD) outlined in the extreme value theory (De Haan and Ferreira, 2007), and the left portion by using other distributions (Xu et al., 2018; Sun, Xu and Zhao, 2020), including gamma, normal, and Weibull. Please see Part G of Supplementary Material A for more details on the mixed distributions. The Bayesian information criterion (BIC) is used for model selection purposes (Schwarz, 1978). Table 9 displays all BICs of a variety of models, and gamma shows the smallest BIC among these distributions. It is seen that the gamma distribution is preferred for all $T$’s. Therefore, we recommend the truncated gamma distribution as the marginal distribution in Eq. (4.4), and its density function in (4.5) can be represented as

$$g^*_i,t(s; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} s^{\alpha-1} e^{-\beta s} \left/ \left(1 - \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta}\right)\right., \ s > 1,$$

where $\alpha$ and $1/\beta$ are shape and scale parameters for $T = t$, respectively, and $\Gamma(\cdot)$ is the gamma function. It is seen that all the estimates in Table 10 are very significant.

<table>
<thead>
<tr>
<th>BIC</th>
<th>Gamma</th>
<th>Normal</th>
<th>Weibull</th>
<th>Gamma-GPD</th>
<th>Normal-GPD</th>
<th>Weibull-GPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 5$</td>
<td>901.26</td>
<td>910.02</td>
<td>926.41</td>
<td>910.10</td>
<td>912.82</td>
<td>918.79</td>
</tr>
<tr>
<td>$T = 6$</td>
<td>1113.72</td>
<td>1130.39</td>
<td>1154.89</td>
<td>1125.06</td>
<td>1127.89</td>
<td>1134.00</td>
</tr>
<tr>
<td>$T = 7$</td>
<td>1301.14</td>
<td>1319.03</td>
<td>1348.06</td>
<td>1313.14</td>
<td>1311.34</td>
<td>1318.57</td>
</tr>
<tr>
<td>$T = 8$</td>
<td>1492.93</td>
<td>1508.73</td>
<td>1540.64</td>
<td>1509.00</td>
<td>1506.29</td>
<td>1510.84</td>
</tr>
</tbody>
</table>

**Table 9**

BICs of various distributions for fitting the nonzero severity.

Difference between expected and observed positive severities are shown in Figure 8. It can be observed that the overall fitting performances are satisfactory as all the points are around 0.

5.3. Modeling multivariate dependence. For the dependence modeling, recall that there are two types of dependence: (i) the dependence between the nonzero s-frequency and severity (i.e., $C^{FS}$), and (ii) the dependence among the conditional severities (i.e., $C^S$). In the following, we discuss the modeling procedures in detail.
<table>
<thead>
<tr>
<th></th>
<th>α</th>
<th></th>
<th>β</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 5$</td>
<td>25.75†</td>
<td>2.48</td>
<td>2.56†</td>
<td>0.25</td>
</tr>
<tr>
<td>$T = 6$</td>
<td>23.29†</td>
<td>2.04</td>
<td>2.30†</td>
<td>0.20</td>
</tr>
<tr>
<td>$T = 7$</td>
<td>24.32†</td>
<td>1.97</td>
<td>2.38†</td>
<td>0.19</td>
</tr>
<tr>
<td>$T = 8$</td>
<td>25.08†</td>
<td>1.89</td>
<td>2.45†</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Table 10  
Parameter estimates ("Est.") and standard errors ("SE") of truncated gamma marginal modeling of non-zero severities. † represents the significance at 0.001.

5.3.1. Dependence among frequency and severity. To examine the dependence between the nonzero s-frequency and the severity, the scatter plots are shown in Figure 9. It can be observed that there is dependence between them for each $T$.

To formally model the dependence between the nonzero s-frequency and severity, both marginals are transformed to uniform marginals by using the aforementioned s-frequency and truncated gamma distributions. Various commonly used bivariate copulas from

\[ \Omega = \{ \text{Gaussian, Student t, Clayton, Joe, Gumbel, and Frank} \} \]

are used for fitting the dependence structure (Joe, 2014). The BIC is used to select the optimal bivariate copula structure for each $T$.

Table 11 shows the fitting results. It can be seen that the Frank copula is selected for $T = 5, 7, 8$ and Gaussian copula is selected for $T = 6$. The estimated parameters are reported in Table 12. All the estimates are very significant, and the positive values of Kendall’s $\tau$ indicate the positive dependence between the nonzero s-frequency and the severity. Those selected dependence structures are used for modeling the dependence between the nonzero s-frequency and the severity.
5.3.2. Dependence among conditional severities. Now given the conditional distributions

\( (F_{S_1|N_1}(s_1|n_1), \ldots, F_{S_T|N_T}(s_T|n_T)) \),

we fit the dependence among the conditional severities via the D-vine copula in Section 4.2.
<table>
<thead>
<tr>
<th>BiCop</th>
<th>Est.</th>
<th>SE</th>
<th>Kendall’s τ</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 5$: Frank</td>
<td>1.14Δ</td>
<td>0.46</td>
<td>0.12</td>
</tr>
<tr>
<td>$T = 6$: Gaussian</td>
<td>0.20*</td>
<td>0.07</td>
<td>0.13</td>
</tr>
<tr>
<td>$T = 7$: Frank</td>
<td>1.28*</td>
<td>0.41</td>
<td>0.14</td>
</tr>
<tr>
<td>$T = 8$: Frank</td>
<td>1.41†</td>
<td>0.37</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Table 12
Parameter estimates, standard errors, and Kendall’s $\tau$s of bivariate copulas between nonzero s-frequency and severity. $\dagger$ represents the significance at 0.001; * represents the significance at 0.01; $\Delta$ represents the significance at 0.05.

To accommodate the temporal dependence, the dependence structure for each tree is assumed to be the same. Specifically, the copulas for tree $T_1$ are all the same as $C_{1,2}$. That is, $C_{1,2} = C_{2,3} = \cdots = C_{T-1,T}$. Similarly, the copulas for tree $T_2$ are all the same as $C_{1,3;2} = \cdots = C_{T-2,T;T-1}$. On the same pattern continues up to tree $T_{T-1}$, with copula $C_{1,T;2:T-1}$. We use the sequel maximum likelihood estimate to select the best copula structure for each tree. Specifically, for the first tree, we use the maximum likelihood estimate to select the best copula from $\Omega$ in Eq. (5.3). Then, we fix the selected copula for the first tree, and perform the same copula selection procedure for the second tree. This sequel selection approach is continued until the last tree. After we select the copulas for all trees, we jointly estimate the corresponding parameters.

It is interesting to find that the same dependence structure is preferred across different trees, although they are allowed to be different. Specifically, the selected copula families are Frank for $T = 5, \ldots, 8$ as shown in Table 13. The Frank copula is perhaps selected for two reasons: (i) Frank copula is very flexible and can accommodate both positive dependence and negative dependence. Note that the positive dependence is exhibited by the data analyzed. (ii) The Frank copula has a nebulous but uniform cloud along the full correlation path. This makes it suitable for fitting the data that exhibits uniform dependence. The dependence in the data we analyze does not concentrate at any part of the distribution; rather, it exhibits uniformity to some extent. This makes the Frank copula able to capture the dependence well. For comparison purposes, we also use the different copula structures for the D-vine fitting. BIC values are reported in Table 13, where it can be seen that the selected copula families have the smallest BICs.

<table>
<thead>
<tr>
<th>BIC</th>
<th>Gaussian</th>
<th>Student t</th>
<th>Clayton</th>
<th>Gumbel</th>
<th>Frank</th>
<th>Joe</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 5$</td>
<td>1082.33</td>
<td>1260.38</td>
<td>884.07</td>
<td>887.41</td>
<td>879.45</td>
<td>895.62</td>
</tr>
<tr>
<td>$T = 6$</td>
<td>1114.51</td>
<td>1292.57</td>
<td>1079.54</td>
<td>1098.17</td>
<td>1078.79</td>
<td>1111.30</td>
</tr>
<tr>
<td>$T = 7$</td>
<td>1295.93</td>
<td>1472.53</td>
<td>1280.64</td>
<td>1258.67</td>
<td>1258.67</td>
<td>1295.93</td>
</tr>
<tr>
<td>$T = 8$</td>
<td>1483.69</td>
<td>1472.53</td>
<td>1461.54</td>
<td>1433.26</td>
<td>1433.26</td>
<td>1484.29</td>
</tr>
</tbody>
</table>

Table 13
BICs of D-vine copula with different pair-copula families.

Table 14 reports the estimated parameters and their significances. Most of estimated parameters are significant, and the estimated parameters for Trees 1 and 2 are very significant, which indicates that there exist strong dependence across the years. The corresponding Kendall’s $\tau$s in Table 14 show that there is positive dependence within each tree. This hints that the severities are positively correlated across the years, which coincides with the conclusion from Table 4.
5.4. Prediction. The prediction is performed based on Algorithm 1. The predictive distribution of severity for each state is derived based on 5,000 simulations. The rolling forecasting is performed for the out-of-sample data by setting $T_1 = 5$ and $T_2 = 9$ in Algorithm 1. That is, we predict the severities for each state from 2015 to 2018, and examine the predictive ability of the proposed model for each year.

For the frequency prediction assessment, we report the predicted MSEs, SMAPEs, and PMADs based on the proposed M4 in Table 15. For comparison purposes, we also report the predicted values for ZINB, HNB, ZIP, and HP, respectively. It can be seen that the proposed M4 has a robust predictive performance. That is, M4 has the smallest predictive MSEs, SMAPEs, and PMADs across all the evaluation years. In Part F of Supplementary Material A, we also present a comparison of M4 with zero-inflated/hurdle type regression models with/without random effects. Again, we observe that M4 has a better prediction performance overall. Therefore, M4 is recommended for predicting the frequency. The synthetic data study in Section H of Supplementary Material A also confirms the superior performance of M4.

### Table 14

<table>
<thead>
<tr>
<th>D-vine</th>
<th>Par.</th>
<th>Tree 1</th>
<th>Tree 2</th>
<th>Tree 3</th>
<th>Tree 4</th>
<th>Tree 5</th>
<th>Tree 6</th>
<th>Tree 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 5$: Frank</td>
<td>Est.</td>
<td>1.56</td>
<td>1.88</td>
<td>0.51</td>
<td>1.66</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\tau$</td>
<td>0.17</td>
<td>0.20</td>
<td>0.06</td>
<td>0.18</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>$T = 6$: Frank</td>
<td>Est.</td>
<td>1.60</td>
<td>2.07</td>
<td>0.68</td>
<td>1.55</td>
<td>0.53</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\tau$</td>
<td>0.17</td>
<td>0.22</td>
<td>0.08</td>
<td>0.17</td>
<td>0.06</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>$T = 7$: Frank</td>
<td>Est.</td>
<td>1.33</td>
<td>2.11</td>
<td>0.45</td>
<td>1.48</td>
<td>0.68</td>
<td>0.68</td>
<td>-</td>
</tr>
<tr>
<td>$\tau$</td>
<td>0.14</td>
<td>0.23</td>
<td>0.05</td>
<td>0.16</td>
<td>0.03</td>
<td>0.09</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>$T = 8$: Frank</td>
<td>Est.</td>
<td>1.37</td>
<td>2.23</td>
<td>0.46</td>
<td>1.49</td>
<td>0.51</td>
<td>0.05</td>
<td>2.75</td>
</tr>
<tr>
<td>$\tau$</td>
<td>0.15</td>
<td>0.24</td>
<td>0.05</td>
<td>0.16</td>
<td>0.06</td>
<td>0.01</td>
<td>0.29</td>
<td></td>
</tr>
</tbody>
</table>

*Estimated parameters (Est.) and Kendall’s $\tau$s for each tree. $^*$ represents the significance at 0.01; $\triangle$ represents the significance at 0.05; $\triangledown$ represents the significance at 0.1.*

### Table 15

<table>
<thead>
<tr>
<th>Year</th>
<th>MSE</th>
<th>M4</th>
<th>ZINB</th>
<th>HNB</th>
<th>ZIP</th>
<th>HP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>4.49</td>
<td>45.67</td>
<td>39.06</td>
<td>12.71</td>
<td>12.78</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>6.59</td>
<td>39.99</td>
<td>34.39</td>
<td>10.70</td>
<td>10.61</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>7.17</td>
<td>64.12</td>
<td>56.87</td>
<td>18.31</td>
<td>18.27</td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td>8.06</td>
<td>49.54</td>
<td>54.58</td>
<td>19.82</td>
<td>16.12</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>6.58</td>
<td>47.86</td>
<td>54.53</td>
<td>15.00</td>
<td>15.03</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>SMAPE</th>
<th>M4</th>
<th>ZINB</th>
<th>HNB</th>
<th>ZIP</th>
<th>HP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>0.62</td>
<td>0.68</td>
<td>0.67</td>
<td>0.71</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>0.53</td>
<td>0.59</td>
<td>0.60</td>
<td>0.58</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>0.55</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td>0.48</td>
<td>0.54</td>
<td>0.54</td>
<td>0.54</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>0.55</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>PMAD</th>
<th>M4</th>
<th>ZINB</th>
<th>HNB</th>
<th>ZIP</th>
<th>HP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>0.33</td>
<td>0.54</td>
<td>0.56</td>
<td>0.50</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>0.27</td>
<td>0.44</td>
<td>0.46</td>
<td>0.36</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>0.32</td>
<td>0.47</td>
<td>0.48</td>
<td>0.40</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td>0.30</td>
<td>0.46</td>
<td>0.47</td>
<td>0.39</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>0.31</td>
<td>0.48</td>
<td>0.49</td>
<td>0.41</td>
<td>0.41</td>
<td></td>
</tr>
</tbody>
</table>

*Predicted MSEs, SMAPEs, and PMADs of the proposed LMM (M4), zero-inflated negative binomial (ZINB), hurdle negative binomial (HNB), zero-inflated Poisson (ZIP), and hurdle Poisson (HP) regression models for different years.*
For the severity prediction assessment, we use the two metrics introduced in Section 4.4, as they can provide more predictive information.

- PLL. For comparison purposes, in addition to the independence (i.e., benchmark) model, we also study the performance of the following panel data model (Baltagi, 2008),

\[ S_{i,t} = \beta_0 + \beta_1 t + \beta_2 c_i + \beta_3 \text{State}_i + \beta_4 t\text{c}_i + \beta_5 t\text{State}_i + \beta_6 c_i\text{State}_i + \beta_7 t\text{c}_i\text{State}_i, \]

with exchangeable and AR(1) correlation structures. That is,

\[
(S_1, \ldots, S_t) \sim \text{MVN}(\mu, V),
\]

where $\mu$ is the mean vector, and $V$ is the variance-covariance matrix with the block diagonal form,

\[
V = \begin{pmatrix}
V_1 & 0 & \cdots & 0 \\
0 & V_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & V_M
\end{pmatrix}.
\]

For the commonly used correlation structures, we have

- exchangeable structure,

\[
V_i = \sigma^2 \begin{pmatrix}
1 & \rho & \cdots & \rho \\
\rho & 1 & \cdots & \rho \\
\vdots & \vdots & \ddots & \vdots \\
\rho & \rho & \cdots & 1
\end{pmatrix},
\]

where $\sigma^2$ is the variance, and $\rho$ is the correlation.

- AR(1) structure,

\[
V_i = \sigma^2 \begin{pmatrix}
1 & \rho & \cdots & \rho^{T-1} \\
\rho & 1 & \cdots & \rho^{T-2} \\
\vdots & \vdots & \ddots & \vdots \\
\rho^{T-1} & \cdots & \cdots & 1
\end{pmatrix},
\]

where $\sigma^2$ is the variance, $\rho$ is the correlation, and $T$ is the number of years in the fitting model.

<table>
<thead>
<tr>
<th>Year</th>
<th>PLL</th>
<th>D-vine</th>
<th>Benchmark</th>
<th>Exchangeable</th>
<th>AR(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>-65.06</td>
<td>-103.05</td>
<td>-141.05</td>
<td>-140.48</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>-58.58</td>
<td>-90.11</td>
<td>-138.61</td>
<td>-138.33</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>-48.31</td>
<td>-92.09</td>
<td>-134.44</td>
<td>-134.10</td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td>-47.29</td>
<td>-100.61</td>
<td>-133.25</td>
<td>-133.29</td>
<td></td>
</tr>
</tbody>
</table>

| TABLE 16 |
| PLL values of severities based on the proposed D-vine, benchmark, exchangeable, and AR(1) models for different years. |

Table 16 shows the PLL values of severities for the proposed D-vine, benchmark, exchangeable, and AR(1) models. It is seen that for each year, the proposed D-vine model has the largest PLL, which suggests that it has the best predictive ability.
• **Scoring rule.** For comparison purposes, we compute the CRPS for each model over different years, and compare the CRPS of proposed D-vine model to those of other models. Table 17 shows the percentages of the CRPS of proposed models that are smaller than the CRPS of the benchmark, exchangeable, and AR(1) models, respectively. That is, 

\[
\text{Percentage} = \frac{1}{52} \sum_{i=1}^{52} \mathbb{1}(\text{CRPS}_{i}^{\text{D-vine}} < \text{CRPS}_{i}^{\text{Other}}),
\]

where \( \text{Other} \in \{\text{benchmark, exchangeable, AR(1)}\} \). It is seen that for all the evaluation years, the proposed model is superior to the other models. In terms of the overall prediction performance, the proposed model can improve 12.50%, 25.96%, and 25.96% compared to the benchmark, exchangeable, and AR(1) models, respectively.

<table>
<thead>
<tr>
<th>Year</th>
<th>CRPS</th>
<th>Benchmark</th>
<th>Exchangeable</th>
<th>AR(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>76.92%</td>
<td>71.15%</td>
<td>71.15%</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>59.62%</td>
<td>86.54%</td>
<td>84.62%</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>63.46%</td>
<td>76.92%</td>
<td>78.85%</td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td>50.00%</td>
<td>69.23%</td>
<td>69.23%</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>62.50%</td>
<td>75.96%</td>
<td>75.96%</td>
<td></td>
</tr>
<tr>
<td>Improvement</td>
<td>12.50%</td>
<td>25.96%</td>
<td>25.96%</td>
<td></td>
</tr>
</tbody>
</table>

**Table 17**

Percentages of CRPS of the proposed model smaller than the CRPS of the benchmark, exchangeable, and AR(1) models, respectively.

In Table 18, the CRPS of proposed D-vine model and benchmark are listed for each state over evaluation years. We further highlight the worst top 10 states in terms of the severity for each year. It is seen that in 2015, 70% of the top 10 states has the smaller CRPS for the proposed model. This percentage increases to 80% in 2016. In 2017, 70% of the top 10 states has the smaller CRPS for the proposed model, and the percentage increases to 80% in 2018. This further shows that the proposed model can perform much better for the worst top 10 states. In Part H of Supplementary Material A, we also evaluate the predictive performance of proposed model on two synthetic data. The additional study confirms the superior prediction performance of developed model as well.

6. **Conclusion and discussion.** In this study, we propose a novel multivariate frequency-severity framework to jointly model the frequency and the number of affected individuals for the healthcare breach data at the state level. It is found that the square root transformed frequency can be well modeled by a linear mixed-effects model. The log-gamma distribution is found to be suitable for modeling the heavy tail phenomenon exhibited by the number of affected individuals. The dependence among s-frequency and severity can be modeled with the Frank copula. The multivariate dependence among conditional severities can be well captured by the D-vine copula. By accommodating the nonlinear positive dependence, we evaluate the proposed frequency-severity model with both in-sample and out-of-sample data, which show that the proposed approach is very promising. The proposed multivariate frequency-severity model has several implications. First, the breach frequency varies significantly across states. Therefore, the breach notification law should be carefully designed at the state level, although the federal HIPAA law is already in place. Second, the severity has a heavy tail and shows strong dependence across years. Therefore, the government regulators should take it into account for when drafting legislation. The proposed model can provide useful guidance for insurance companies for modeling and pricing the breach risks in healthcare. Third, some
<table>
<thead>
<tr>
<th>State</th>
<th>D-vine</th>
<th>Benchmark</th>
<th>State</th>
<th>D-vine</th>
<th>Benchmark</th>
<th>State</th>
<th>D-vine</th>
<th>Benchmark</th>
<th>State</th>
<th>D-vine</th>
<th>Benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>AK</td>
<td>1.90</td>
<td>2.51</td>
<td>AL</td>
<td>1.43</td>
<td>1.32</td>
<td>AR</td>
<td>0.39</td>
<td>0.61</td>
<td>AZ</td>
<td>0.70</td>
<td>0.71</td>
</tr>
<tr>
<td>CA</td>
<td>2.11</td>
<td>4.20</td>
<td>CO</td>
<td>0.86</td>
<td>1.20</td>
<td>CT</td>
<td>0.54</td>
<td>0.84</td>
<td>DC</td>
<td>0.43</td>
<td>0.51</td>
</tr>
<tr>
<td>DE</td>
<td>6.64</td>
<td>7.18</td>
<td>FL</td>
<td>0.51</td>
<td>0.69</td>
<td>GA</td>
<td>1.88</td>
<td>2.64</td>
<td>HI</td>
<td>0.46</td>
<td>1.09</td>
</tr>
<tr>
<td>IA</td>
<td>7.85</td>
<td>8.74</td>
<td>ID</td>
<td>8.87</td>
<td>7.93</td>
<td>IL</td>
<td>0.44</td>
<td>0.62</td>
<td>IN</td>
<td>6.46</td>
<td>7.08</td>
</tr>
<tr>
<td>KS</td>
<td>0.74</td>
<td>0.80</td>
<td>KY</td>
<td>1.37</td>
<td>0.90</td>
<td>LA</td>
<td>0.61</td>
<td>0.47</td>
<td>MA</td>
<td>0.48</td>
<td>0.53</td>
</tr>
<tr>
<td>MD</td>
<td>3.08</td>
<td>2.80</td>
<td>ME</td>
<td>1.44</td>
<td>1.49</td>
<td>MI</td>
<td>0.55</td>
<td>0.51</td>
<td>MN</td>
<td>1.53</td>
<td>1.35</td>
</tr>
<tr>
<td>MO</td>
<td>0.46</td>
<td>0.49</td>
<td>MS</td>
<td>0.95</td>
<td>0.48</td>
<td>MT</td>
<td>0.48</td>
<td>0.53</td>
<td>NC</td>
<td>0.41</td>
<td>0.47</td>
</tr>
<tr>
<td>ND</td>
<td>7.14</td>
<td>7.74</td>
<td>NE</td>
<td>0.71</td>
<td>1.52</td>
<td>NH</td>
<td>8.15</td>
<td>3.99</td>
<td>NJ</td>
<td>2.24</td>
<td>1.26</td>
</tr>
<tr>
<td>NM</td>
<td>1.87</td>
<td>2.58</td>
<td>NV</td>
<td>0.45</td>
<td>0.48</td>
<td>NY</td>
<td>3.28</td>
<td>3.98</td>
<td>OH</td>
<td>0.41</td>
<td>0.50</td>
</tr>
<tr>
<td>OK</td>
<td>0.42</td>
<td>0.50</td>
<td>OR</td>
<td>1.41</td>
<td>1.23</td>
<td>PA</td>
<td>0.52</td>
<td>0.88</td>
<td>PR</td>
<td>2.33</td>
<td>1.74</td>
</tr>
<tr>
<td>RI</td>
<td>0.44</td>
<td>0.52</td>
<td>SC</td>
<td>0.42</td>
<td>0.67</td>
<td>SD</td>
<td>0.65</td>
<td>0.56</td>
<td>TN</td>
<td>1.12</td>
<td>1.19</td>
</tr>
<tr>
<td>TX</td>
<td>0.50</td>
<td>1.31</td>
<td>UT</td>
<td>8.66</td>
<td>8.83</td>
<td>VA</td>
<td>1.65</td>
<td>2.41</td>
<td>VT</td>
<td>0.82</td>
<td>1.29</td>
</tr>
<tr>
<td>WA</td>
<td>4.98</td>
<td>5.00</td>
<td>WI</td>
<td>1.53</td>
<td>1.75</td>
<td>WV</td>
<td>7.94</td>
<td>8.46</td>
<td>WY</td>
<td>7.95</td>
<td>8.30</td>
</tr>
<tr>
<td>AK</td>
<td>0.46</td>
<td>0.55</td>
<td>AL</td>
<td>0.51</td>
<td>0.49</td>
<td>AR</td>
<td>1.60</td>
<td>0.89</td>
<td>AZ</td>
<td>3.56</td>
<td>4.01</td>
</tr>
<tr>
<td>CA</td>
<td>1.08</td>
<td>3.03</td>
<td>CO</td>
<td>0.41</td>
<td>0.49</td>
<td>CT</td>
<td>1.17</td>
<td>0.71</td>
<td>DC</td>
<td>0.40</td>
<td>0.84</td>
</tr>
<tr>
<td>DE</td>
<td>1.02</td>
<td>0.80</td>
<td>FL</td>
<td>2.17</td>
<td>3.54</td>
<td>GA</td>
<td>1.48</td>
<td>2.43</td>
<td>HI</td>
<td>0.79</td>
<td>0.66</td>
</tr>
<tr>
<td>IA</td>
<td>0.87</td>
<td>0.50</td>
<td>ID</td>
<td>7.24</td>
<td>7.48</td>
<td>IL</td>
<td>0.44</td>
<td>0.67</td>
<td>IN</td>
<td>0.90</td>
<td>1.50</td>
</tr>
<tr>
<td>KS</td>
<td>0.93</td>
<td>0.65</td>
<td>KY</td>
<td>0.49</td>
<td>0.65</td>
<td>LA</td>
<td>0.47</td>
<td>0.50</td>
<td>MA</td>
<td>0.42</td>
<td>0.50</td>
</tr>
<tr>
<td>MD</td>
<td>1.91</td>
<td>2.26</td>
<td>ME</td>
<td>7.56</td>
<td>7.50</td>
<td>MI</td>
<td>0.72</td>
<td>0.63</td>
<td>MN</td>
<td>0.52</td>
<td>0.51</td>
</tr>
<tr>
<td>MO</td>
<td>1.40</td>
<td>0.93</td>
<td>MS</td>
<td>1.05</td>
<td>0.79</td>
<td>MT</td>
<td>0.47</td>
<td>0.55</td>
<td>NC</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>ND</td>
<td>7.24</td>
<td>7.28</td>
<td>NE</td>
<td>0.91</td>
<td>0.57</td>
<td>NH</td>
<td>0.86</td>
<td>0.55</td>
<td>NJ</td>
<td>0.79</td>
<td>1.12</td>
</tr>
<tr>
<td>NM</td>
<td>0.72</td>
<td>0.50</td>
<td>NV</td>
<td>0.95</td>
<td>1.22</td>
<td>NY</td>
<td>2.38</td>
<td>3.81</td>
<td>OH</td>
<td>2.10</td>
<td>2.06</td>
</tr>
<tr>
<td>OK</td>
<td>0.41</td>
<td>0.71</td>
<td>OR</td>
<td>0.54</td>
<td>0.56</td>
<td>PA</td>
<td>0.45</td>
<td>0.71</td>
<td>PR</td>
<td>9.50</td>
<td>8.96</td>
</tr>
<tr>
<td>RI</td>
<td>0.51</td>
<td>0.53</td>
<td>SC</td>
<td>1.86</td>
<td>0.93</td>
<td>SD</td>
<td>7.52</td>
<td>8.03</td>
<td>TN</td>
<td>0.53</td>
<td>0.55</td>
</tr>
<tr>
<td>TX</td>
<td>0.59</td>
<td>1.53</td>
<td>UT</td>
<td>0.47</td>
<td>0.60</td>
<td>VA</td>
<td>0.43</td>
<td>0.56</td>
<td>VT</td>
<td>7.45</td>
<td>7.32</td>
</tr>
<tr>
<td>WA</td>
<td>2.27</td>
<td>2.10</td>
<td>WI</td>
<td>0.42</td>
<td>0.52</td>
<td>WV</td>
<td>7.66</td>
<td>8.24</td>
<td>WY</td>
<td>0.55</td>
<td>1.09</td>
</tr>
<tr>
<td>AK</td>
<td>1.40</td>
<td>2.67</td>
<td>AL</td>
<td>0.87</td>
<td>0.49</td>
<td>AR</td>
<td>1.98</td>
<td>1.10</td>
<td>AZ</td>
<td>0.49</td>
<td>0.53</td>
</tr>
<tr>
<td>CA</td>
<td>0.45</td>
<td>1.86</td>
<td>CO</td>
<td>1.23</td>
<td>0.65</td>
<td>CT</td>
<td>0.67</td>
<td>0.49</td>
<td>DC</td>
<td>1.62</td>
<td>1.99</td>
</tr>
<tr>
<td>DE</td>
<td>0.75</td>
<td>0.60</td>
<td>FL</td>
<td>0.60</td>
<td>1.03</td>
<td>GA</td>
<td>0.68</td>
<td>1.48</td>
<td>HI</td>
<td>7.12</td>
<td>8.04</td>
</tr>
<tr>
<td>IA</td>
<td>0.38</td>
<td>0.68</td>
<td>ID</td>
<td>8.30</td>
<td>7.69</td>
<td>IL</td>
<td>0.45</td>
<td>0.62</td>
<td>IN</td>
<td>0.47</td>
<td>1.11</td>
</tr>
<tr>
<td>KS</td>
<td>0.91</td>
<td>0.81</td>
<td>KY</td>
<td>2.77</td>
<td>2.27</td>
<td>LA</td>
<td>0.41</td>
<td>0.53</td>
<td>MA</td>
<td>0.42</td>
<td>0.51</td>
</tr>
<tr>
<td>MO</td>
<td>0.45</td>
<td>0.63</td>
<td>ME</td>
<td>0.53</td>
<td>0.82</td>
<td>MI</td>
<td>2.52</td>
<td>2.11</td>
<td>MN</td>
<td>0.44</td>
<td>0.50</td>
</tr>
<tr>
<td>ND</td>
<td>1.18</td>
<td>1.01</td>
<td>MS</td>
<td>0.40</td>
<td>0.58</td>
<td>MT</td>
<td>0.48</td>
<td>0.53</td>
<td>NC</td>
<td>0.94</td>
<td>0.99</td>
</tr>
<tr>
<td>NM</td>
<td>0.68</td>
<td>1.12</td>
<td>NE</td>
<td>1.05</td>
<td>0.51</td>
<td>NH</td>
<td>0.66</td>
<td>1.47</td>
<td>NJ</td>
<td>0.61</td>
<td>0.54</td>
</tr>
<tr>
<td>OK</td>
<td>0.84</td>
<td>0.67</td>
<td>OR</td>
<td>0.58</td>
<td>0.66</td>
<td>PA</td>
<td>1.12</td>
<td>1.77</td>
<td>PR</td>
<td>8.75</td>
<td>9.04</td>
</tr>
<tr>
<td>RI</td>
<td>0.84</td>
<td>0.52</td>
<td>SC</td>
<td>3.11</td>
<td>2.71</td>
<td>SD</td>
<td>0.43</td>
<td>0.63</td>
<td>TN</td>
<td>0.58</td>
<td>0.94</td>
</tr>
<tr>
<td>TX</td>
<td>0.83</td>
<td>2.03</td>
<td>UT</td>
<td>1.84</td>
<td>2.09</td>
<td>VA</td>
<td>0.57</td>
<td>0.50</td>
<td>VT</td>
<td>0.65</td>
<td>0.95</td>
</tr>
<tr>
<td>WA</td>
<td>0.55</td>
<td>0.53</td>
<td>WI</td>
<td>1.39</td>
<td>0.75</td>
<td>WV</td>
<td>0.76</td>
<td>0.73</td>
<td>WY</td>
<td>8.23</td>
<td>8.39</td>
</tr>
</tbody>
</table>

**Table 18**

CRPS of severity for each state for the proposed D-vine and benchmark models. The top 10 states in terms of the severity for each year are bolded.

application-oriented metrics (e.g., a linear/nonlinear function of frequency and severity) can be created for the practical use based on the predictive distributions. For example, the predictive distribution of mean severity can be directly derived from our prediction results, which can be used for pricing purposes.

Although the developed frequency-severity framework is motivated by the state-level healthcare breach data, it applies to more than this kind of data. For example, in insurance, the frequency-severity model has been widely adopted in ratemaking and underwriting for
the premium determination (Kaas et al., 2008). It is very important in insurance applications due to several requirements, such as terms of contracts, insurers’ databases, and regulatory requirements; see Frees, Derrig and Meyers (2014). The proposed framework can be equally applied to the SIC (Standard Industrial Classification) or NAICS (North American Industry Classification System) code-based breach data for group risk rating and premium writing, which is badly needed in the current cyber insurance market. For instance, Advisen provides cyber risk data with NAICS codes. The developed model can be used for grouping cyber risks based on the code. The proposed model is also applicable in business decision-making. For example, there are four basic categories of business risk: i) low frequency, low severity; ii) high frequency, low severity; iii) low frequency, high severity; and iv) high frequency, high severity. Based on the frequency-severity model, the business decision-maker can take the corresponding strategy. For instance, if the business risk is predicted to be category i), then the decision-maker can decide to accept the risk, and retain and budget for it. However, if the business risk is predicted to be category iii), the decision-maker should transfer the risk to a third party (e.g., an insurance company). We have employed the proposed framework to study two synthetic data in Part H of Supplementary Material A. The satisfactory fitting and prediction performances of proposed model further indicate that the framework is broadly applicable.

The current study has some of the same limitations as other studies. First, the model is geared towards predicting the distributions of frequencies and severities one year ahead of time. It does not predict the timing of occurrences of breaches, which may be of interest to some insurance companies. Second, the dataset for the case study may not be complete because some breach incidents may not be reported or even detected. Nevertheless, it is, to the best of our knowledge, the most comprehensive, publicly available, official healthcare dataset that is directly related to HIPAA and state breach regulations.

Acknowledgments. The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that significantly improved the quality of this paper. The authors particularly thank the Editor-In-Chief for suggesting the approach for modeling the frequency.

SUPPLEMENTARY MATERIAL

Supplementary Material A: Supporting Material
It provides supporting material consisting of Sklar’s theorem (Part A), details on zero-inflated/hurdle type count models (Part B), index, full Name, abbreviation, and population of US States (Part C), performance comparison by using IFM and full MLE (Part D), QQ-plots of fitted zero-inflated and hurdle regression models for $T = 5, 6, 7$ (Part E), fitting and predictive performances of zero-inflated and hurdle regression models with random effects (Part F), mixed distribution with GPD tail (Part G), and additional studies on two synthetic data (Part H).

Supplementary Material B: Healthcare Breach Dataset and R code
It contains the state level healthcare breach dataset and R code.

REFERENCES


DATA BREACH IN HEALTHCARE


