PRIVACY ANALYTICS

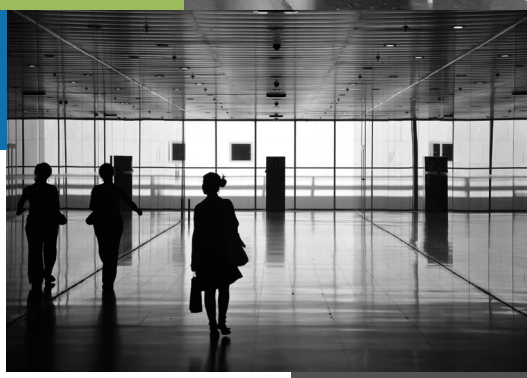
WHITE PAPER

CALCULATING THE ROI FROM THE DE-IDENTIFICATION OF HEALTH DATA

Beyond simply protecting the privacy of individuals, there is also a compelling business case for de-identification.

In this white paper, we present this case by performing a Return on Investment (ROI) analysis based on a series of typical scenarios.

This analysis illustrates that when considering the savings from avoiding a data breach, even modest investments in de-identification produce significant ROI.



PRIVACY ANALYTICS

a QuintilesIMS company

Calculating the ROI from the De-identification of Health Data

Breach Notification Costs and Likelihood

When a health dataset that has not been de-identified is involved in a breach, there is a need to notify the affected individuals, the appropriate Attorney General, regulators, and possibly the media, depending on the jurisdiction where the patients reside and the number of individuals affected by the breach. The total costs of a breach according to the Ponemon Institute were estimated to be approximately $200 per affected individual[1]. This cost covers investigation, direct notification costs, litigation, redress and compensation, penalties, loss of productivity to deal with the breach, and loss of business.

Using this figure, we can estimate the approximate cost of a breach if the dataset is not de-identified, as in the Ponemon Institute statistics on the cost of data breaches. The data comes from a series of Ponemon Institute reports:

- 2009 Annual Study – Cost of a Data Breach, Traverse City, MI
- 2010 Annual Study – U.S. Cost of a Data Breach, Traverse City, MI
- 2011 Cost of a Data Breach Study – United States, Traverse City, MI

Year	Average per Person	Organizational Average
2011	$194	$5,501,889
2010	$214	$7,241,899
2009	$204	$6,751,451
2008	$202	$6,655,758
2007	$197	$6,355,132
2006	$182	$4,789,637
2005	$138	$4,541,429

Table 1: Costs of a data breach. Values in USD

Year	Industry	Average per Person
2011	Healthcare	$240
	Services	$185
	Financial	$247
2010	Healthcare	$301
	Services	$301
	Financial	$353
2009	Healthcare	$294
	Services	$256
	Financial	$249
2008	Healthcare	$282
	Services	$283
	Financial	$240

Table 2: Costs by selected industries. Values in USD.

Year	Direct Costs	Indirect Costs
2011	$59	$135
2010	$73	$141
2009	$60	$144
2008	$50	$152

Table 3: Direct and Indirect Costs per data breach. Values in USD.

There is also evidence that suggests approximately 27% of organizations covered by the HIPAA Security Rule experience a reportable breach every year[2]. Given the methodology of the study that computed this number, it is arguably one of the more credible contemporary breach incidence estimates.

Breach Notification Costs and Likelihood

Although this number may seem high, it is actually lower than many other estimates produced by other organizations and that are also often quoted. Nevertheless, in our sensitivity analysis, we examine the impact of using an even smaller percentage of breaches per year.

But not every organization will experience a breach with a 27% probability. Therefore, to capture how the probability of a breach will vary across organizations, we can instead represent that probability as a distribution. For our purposes, we used what is called a Beta distribution, which represents an entire family of distributions. As parameters for the distribution are changed, so is the distribution's shape.

In this case, we used a mode of 0.27, with shape parameters α of 1.55 and β of 2.5, which result in a distribution skewed to the right. This particular distribution is shown in Figure 1. Using the mode implies that most of the time the probability of a breach will be 0.27, and by skewing the distribution to the right we assume that organizations are more likely to have a low probability of a breach.

Essentially, the Beta distribution gives us the probability that an organization will experience a breach, and therefore the probability that the organization will incur significant breach notification costs.

Return on Investment

In our analysis, we assume that an organization will invest between $100K to $500K to implement de-identification. We chose these numbers because they reflect the range we have observed in practice for medium-sized enterprises when they implement a corporate de-identification solution, including policies, training and tools.

In addition to meeting other compliance requirements, this de-identification can save the organization the costs of breach notification.

For example, if an organization experiences a breach of 5,000 records belonging to 5,000 individuals, and the dataset is not de-identified, then the total breach notification cost is estimated at $1M (5,000 records x $200 per record).

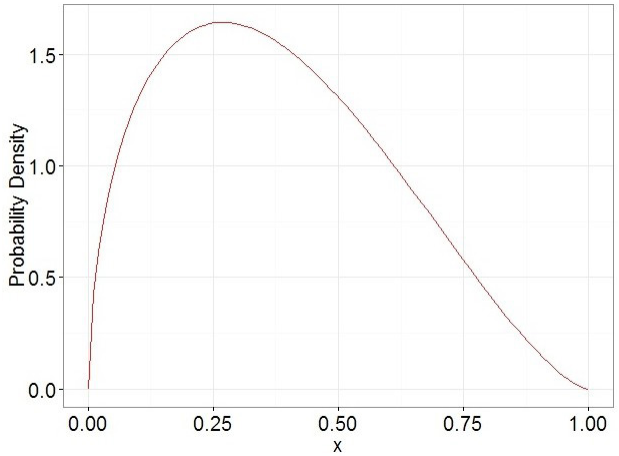


Figure 1: The probability density function for the Beta distribution. The x-axis is the probability of a health data custodian having a breach.

Return on Investment

However, if a breach occurs and the organization has been proactive in de-identifying its data, then the total cost to deal with the breach will be lower. The organization will still incur a cost of mobilizing an internal team, external counsel, and possible external security consultants to investigate the breach and to confirm that it is not a reportable breach. For the purposes of our model, we assumed that these expenses for a breach on previously de-identified data are a fixed cost of $20K. As you can see, these costs will be much lower than if the breach occurs on non-de-identified data ($1M versus $20K). Even if we change the $20K to a reasonably larger number, the main conclusions from our analysis are not changed. We used a standard ROI model as shown (in center)[3]:

$$\frac{\text{Costs Saved - Deid Costs}}{\text{Deid Costs}}$$

Using our example of a $1M breach notification cost, if that same organization de-identified its data beforehand then we can compute the savings. They would incur an estimated cost of $100K to de-identify the database and $20K to investigate the breach. Therefore, the costs saved would be $880K. If we plug those numbers in the equation above, we get 7.8 ROI (($880K – $100K)/$100K).Even in this simple example, the ROI numbers are significant because the costs of a breach are so high.

In our model, we assumed a one-year time horizon, and that the breach occurred at the end of that year. We used a discount rate of 5% since the value of money spent in the future is less today, which is when the investment in de-identification is made.

Simulation

To compute the overall ROI we performed a Monte Carlo simulation, which is a class of computer algorithms that use random sampling to get probabilities similar to what you would expect if you were collecting results from a casino game, hence the name. The simulation allowed us to model the probability distribution of a breach actually occurring. The simulation was run 1,000 times and the average ROI calculated across these runs (this is the expected ROI). In each simulation run, we draw a probability from the Beta distribution we described earlier (see Figure 1), and then draw from a binomial distribution with that probability. A binomial distribution allows us to simulate whether a breach does or does not occur. If a breach does occur, then we plug the numbers in the above equation to get the ROI.

Sensitivity Analysis

If no breach occurs, then the ROI is negative. The results of the simulation for two possible investments in de-identification are shown in Figure 2. The two investment values are $100K and $500K. The x-axis of the graph shows the database size (which affects the cost of the breach) going up to 1 million.

The ROI values are staggering, and are shown on the y-axis. The reason for such high ROI values is that the cost of a data breach can be so high. Even when we take into account that a breach occurring is only probabilistic — that is, it is not going to happen with every dataset or every institution — the expected ROI is still high and, as would be anticipated, it progressively gets higher as the databases get larger. For example, if the expected ROI is 20, then it means that the expected return, in terms of savings, is 20 times the investment in de-identification. As the graph shows, the expected ROI is much higher than 20 for large databases.

For a $100K investment in de-identification, the expected ROI is positive if the affected database has 1,300 or more individuals. For a $500K investment in de-identification, the expected ROI is positive if the affected database has 6,900 or more individuals. Therefore, the ROI becomes positive even for relatively small databases.

We can examine how these expected returns are affected if we vary some of the assumptions.

Sensitivity Analysis

The first assumption we examine is the cost of a data breach. While the cost of a data breach has been estimated to be approximately $200 per record at the low end, it may be argued that the total cost of a breach value does not keep growing linearly as the database size becomes very large — there should be a plateau at some point. However, justifying where such a plateau is reached is not easy since we were unable to find credible data.

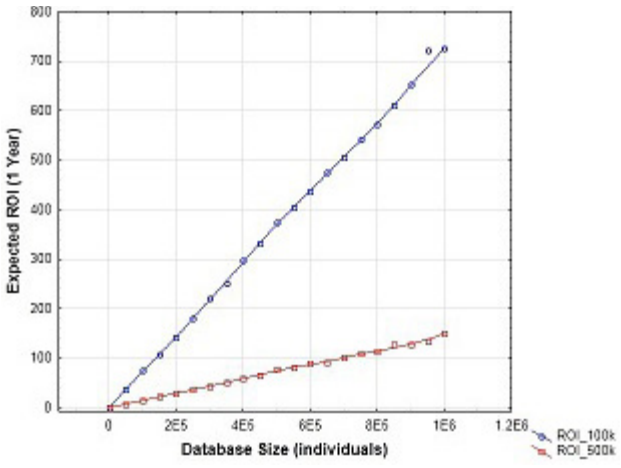


Figure 2: The expected ROI results for databases of various sizes affecting up to 1 million individuals.

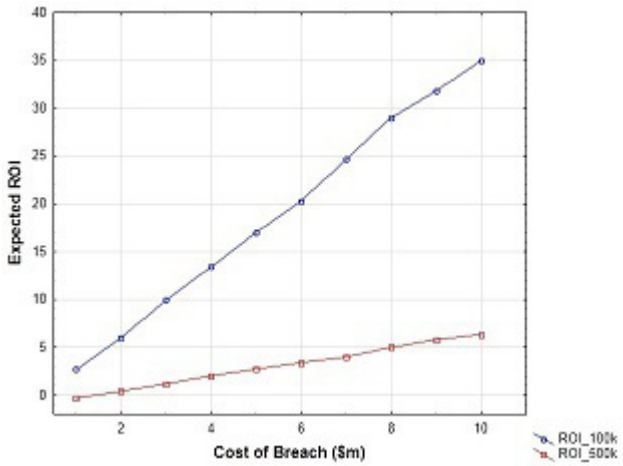


Figure 3: The expected ROI when the data breach costs are fixed values (the x-axis in $M).

Therefore, we assumed a fixed cost for a breach. In the graph below, we assumed a breach cost of $1M to $10M. This removes consideration of database size and simply assumes a lump sum within the range that has been reported in in Ponemon Institute Statistics on the Cost of Data Breaches. As can be seen in Figure 3, the expected ROI is still very high. For example, if a breach cost is $2M and the investment in de-identification is $100K, then the ROI is 6 times. For a de-identification investment of $500K, the expected ROI is approximately 0.4 times (an average of a 40% ROI).

We also modified the Beta distribution that we used to model the probability of a breach occurring. We set the beta shape parameters to of 1.1 and of 2.97, which means that the probability of a breach has an average at 0.27 and drops off quite rapidly beyond that (more skew to the right). In effect, this makes a breach less likely to occur than in the previous set of results. These new results are shown in Figure 4 as the database size increases.

As we can see, the expected ROI is reduced compared to the previous model, but it is still quite high.

In Figure 5, we show the results for the case where the cost of the breach is fixed. We can see that for a $500K investment in de-identification, the ROI is negative unless the cost of the breach approaches $2M (which would be equivalent to 10,000 individuals in the breached database, using the $200 per individual assumption).

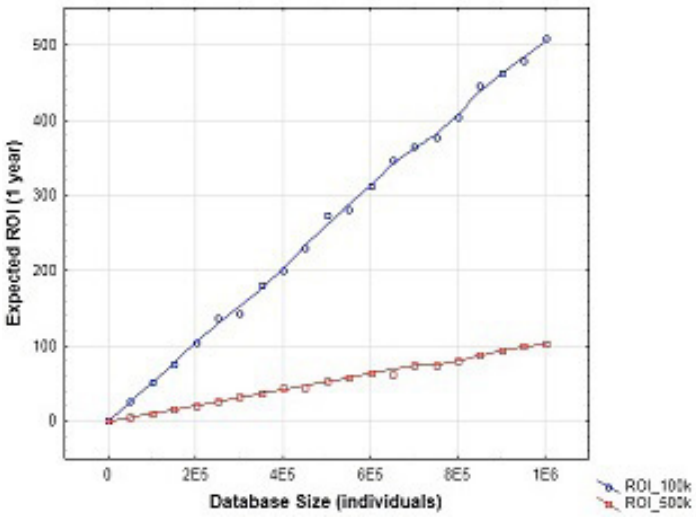


Figure 4: The expected ROI when the likelihood of a breach occurring is further skewed to the right for different database sizes.

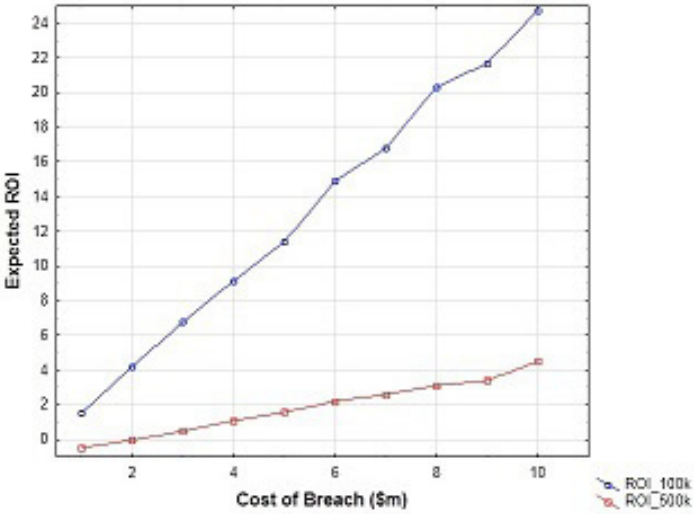


Figure 5: The expected ROI when the data breach costs are fixed values (the x-axis in$M) when the likelihood of a breach occurring is further skewed to the right.

Summary

CONTACT US

251 Laurier Ave W
Suite 200
Ottawa, Ontario, Canada
K1P 5J6

Phone: 613.369.4313

www.privacy-analytics.com

sales@privacy-analytics.com

When considered as an investment, de-identification produces significant expected returns, and this conclusion is quite robust across variations in ROI model assumptions. For the two investment sizes we considered:

- $100K investment in de-identification – The returns are positive for databases around 2,000 or more individuals.
- $500K investment in de-identification – The returns are positive for databases of around 10,000 or more individuals under the more conservative assumptions that are still consistent with current evidence on breach costs.

De-identification is a good investment because data breaches are becoming increasingly likely and the costs of notification when a data breach occurs are staggering. Any modest investment that would reduce these costs would pay for themselves relatively quickly.
We did not show the results when we model ROI over multiple years because the expected ROI numbers just increase further - the one - year numbers already make a strong case. If we included other savings from de-identification such as monetary benefits from being able to disclose data, then the ROI value can only go up.

References

1. El Emam, K., Guide to the De-identification of Personal Health Information. CRC Press (Auerbach), 2013.

2. HIMSS Analytics, 2012 HIMSS Analytics Report: Security of Patient Data, 2012.

3. El Emam, K., The ROI from Software Quality. CRC Press (Auerbach), 2005.