The Statistical Invalidity SAFETY PER

By Matthew Hallowell, Mike Quashne, Rico Salas, Matt Jones, B

TOTAL RECORDABLE INCIDENT RATE (TRIR) has been used as the primary measure of safety performance for nearly 50 years. Simply, TRIR is the rate at which a company experiences an OSHA-recordable incident, scaled per 200,000 work hours. TRIR is based upon a standard definition of a recordable incident that was created and institutionalized in the recordkeeping requirements of the OSH Act of 1970 (BLS, 2019). According to the general criteria, an incident is recordable if it results in work-related injury or illness, requiring medical treatment beyond first aid, that results in loss of consciousness, days away from work, restricted work or transfer to another job (OSHA, 2010). Since organizations conform to this same definition, the TRIR metric has been used to compare industries, sectors, companies and even projects.

TRIR is used in many ways to measure safety performance, from the work site to the boardroom. For example, organizations use TRIR to report results, benchmark against peers, prequalify and select contractors, evaluate the performance of managers and track the impact of safety initiatives (Lofquist, 2010; Manuele, 2008; Wilbanks, 2019). TRIR is also a primary safety metric of concern among executives because it may impact workers' compensation insurance premiums, influence public image and be scrutinized by potential customers or investors (Lingard et al., 2017; Lofquist, 2010; Ritchie, 2013; Salas, 2020; Truitt, 2012). Although other lagging measures such as days away, restricted or transferred (DART) rates are also considered, no safety metric is as ubiquitous as TRIR.

Despite the pervasive use of TRIR, its limitations are being recognized. For example, some argue that TRIR is a poor reflection of safety performance because it does not account for the actual or potential severity of an incident (Toellner, 2001). For example, a four-stitch cut to the finger is counted in the same way as a fatality, and a near-hit with the potential to be fatal is not in the TRIR metric at all. Others point out that TRIR is reactive in nature as it only counts incidents and does not consider the underlying safety program (Lingard et al., 2017; Lofquist, 2010; Salas & Hallowell, 2016).

KEY TAKEAWAYS

- **This study challenges conventional wisdom of safety measurement with an empirical analysis of 3.26 trillion work hours of total recordable incident rate (TRIR) data and a statistical demonstration of proof.**
- **Parametric and nonparametric statistical analysis revealed that no discernable association exists between TRIR and fatalities.**
- **This analysis also found that the occurrence of recordable injuries is almost entirely random. Further, it found that TRIR is not precise and should not be communicated to multiple decimal points of precision.**
- **Finally, the analysis revealed that in nearly every practical circumstance, it is statistically invalid to use TRIR to compare companies, business units, projects or teams.**

More recently, some have begun to question the statistical validity of TRIR, suggesting that recordable injuries happen so infrequently that the metric is not stable or reliable. Since TRIR is typically reported over relatively short time frames (e.g., months, quarters, years), the number of recordable injuries in each period can be exceedingly small. Therefore, it is suspected that the confidence interval of typical reporting periods is so wide that it renders the metric useless. This potential limitation is implicitly recognized by those who criticize TRIR as unfairly biased against small companies.

To better understand the validity of TRIR as a performance metric, this study sought to answer the basic question: Given the way that it is used, to what extent is TRIR statistically valid? More specifically, the authors aimed to test whether TRIR is statistically stable, precise, predictive and indicative of high-severity events. The answer will help clarify whether TRIR should be used to make important business decisions such as comparing two contractors, evaluating the safety performance of managers or concluding that a new safety intervention is effective.

Background

A statistical analysis of any variable requires its decomposition into underlying components. Since TRIR is a rate, the metric is comprised of two variables: the number of events and time. Here, events are injuries and illnesses that conform to OSHA's definition of recordability. On the other hand, time is expressed as worker exposure hours. Because expressing TRIR as the number of incidents per work hour would yield an excessively small fraction, TRIR is scaled per 200,000 work hours. Since 200,000 work hours equates to approximately 100 employees working full time for 1 year, TRIR also reflects the percentage of workers who suffer a recordable incident in a year.

To compute a TRIR, an organization applies Equation 1 for a specific period. For example, a company that accounts for three recordable incidents over 350,000 work hours in a month would have a TRIR of 1.71 per 200,000 work hours for that month.

Equation 1:

$$TRIR = \frac{No. of\ recordable\ incidents \cdot 200{,}000}{No. of\ work\ hours}$$

TRIR is typically reported as a single number. Statistically, this means that the company takes the single TRIR value to represent safety performance over the period as if it was the only possible outcome. This number is also often reported to one or more decimal points of precision, where subtle differences in TRIR are assumed to be meaningful (e.g., the difference between 1.2 and 1.6 from one month to the next). Although

THINKHUBSTUDIO/ISTOCK/GETTY IMAGES PLUS

# of TRIR as a Measure of RFORMANCE

rad MacLean and Ellen Quinn

TRIR is used in this manner, the underlying assumptions have never been validated.

In addition, TRIR is often used as a dependent variable by academic researchers. TRIR has been applied frequently as an objective, empirical metric of safety performance and is often conceptualized as an ideal response variable. In fact, the authors have used TRIR as a dependent variable when identifying which safety practices are more effective than others (Hallowell, 2010; Hallowell & Gambatese, 2009); validating safety leading indicators as predictive (Alruqi & Hallowell, 2019; Hallowell et al., 2013; Hinze et al., 2013; Lingard et al., 2017; Salas & Hallowell, 2016) and measuring intrinsic relationships between safety and other performance metrics such as productivity and quality (Wanberg et al., 2013).

Hypothetical Cases

Case examples were created to illustrate the potential limitations of using TRIR as a comparative metric:

- Company A has a recordable incident in the first 1,000 work hours that it is in business. At this point, the company's TRIR is 200 per 200,000 work hours.
- Company B has seven recordable incidents over 980,000 work hours in a given year. The company reports its yearly TRIR as 1.4 per 200,000 work hours.
- Company C has 24 recordable incidents over 6 million work hours in a given year. The company reports its yearly TRIR as 0.8 per 200,000 work hours.

The hypothetical cases were designed to vary greatly in the number of recordable injuries, the number of work hours accumulated and the resulting TRIR. Case A represents a new contractor that has a recordable incident early in its company history. In contrast, Case B is a medium-sized company with fewer than 500 employees that accumulates just less than 1 million work hours in a year. Finally, Company C is a large company that amasses millions of work hours in a year.

Company A is an extreme example that underscores the limitation of reporting TRIR over extremely short time frames or for small businesses. On the surface, the TRIR of 200 for Company A could be judged as more than 50 times worse than the average TRIR in the construction industry (OSHA, 2016). However, the short time frame makes it difficult to support this judgment. It also drives the general question: How many work hours of exposure are needed before TRIR becomes statistically meaningful?

Companies B and C offer interesting contrast as both accumulate relatively large numbers of work hours. Again, on the surface, it may appear that Company C is nearly twice as good as Company B. However, it is still unclear whether Company B is statistically different from Company C given that injuries do not appear to occur at a regular, predictable interval.

Later in this article, these three case companies are used to illustrate proper interpretation of TRIR and the implications of the results.

Analytical Approach

This study was performed via a collaboration among senior leaders of 10 construction companies and four academic researchers. The collaboration resulted in direct access to more than 3 trillion work hours of internally reported incident data, which were analyzed by the researchers using various diagnostic and predictive analytics.

Both parametric and nonparametric analyses were used to study TRIR. A parametric statistical analysis makes logical assumptions about the defining properties of the distributions (i.e., the metric follows a binomial distribution). Nonparametric statistical analyses make no assumptions about the underlying probability distributions, and instead estimate their distributions solely from the data. Both are important for statistical modeling because they help to answer different questions. For example, parametric analyses help to interpret the precision of TRIR by considering confidence intervals and the nonparametric analyses help to test whether past TRIR is predictive of future TRIR. Recognizing the significant implications of these results, both approaches were taken to gain a full understanding of when, if ever, TRIR can be used as a comparative or predictive metric.

Parametric Analysis

A parametric analysis starts by identifying the underlying distribution that governs the phenomenon to remove a layer of abstraction and create representative statistical equations. Most common metrics conform to well-known mathematical distributions such as normal, binomial or Poisson. When fitting a distribution, it is important to examine the assumptions about the metric to ensure that the chosen distribution is valid. Once a distribution is known, a mathematical function (equation) can be produced that allows us to interpret the precision of a given TRIR, or answer questions such as, how much exposure time is required to produce a precise measure of TRIR? For simplicity, we use the term "precision" to refer to the width of the confidence intervals, where a wide interval is considered to be less precise than a narrow interval.

The parametric analysis revealed important insight into how TRIR should be communicated. At present, most organizations report TRIR as a single number, often to multiple decimal points (e.g., 1.84). However, as will be shown, TRIR is subject to random variation and should not be communicated as a single number. For example, with the same underlying safety system, an organization would not expect the exact same number of recordable injuries every reporting period. Rather, it may be equally likely that long periods may exist between incidents or that incidents occur in clusters. Additionally, although TRIR is widely reported as a concrete measurement, it is actually a sample taken over a discrete period (e.g., 1 month or 1 year). To make statements about the possible outcomes of a safety management system, it is important to understand TRIR as a distribution of potential values rather than as a single point estimate.

TRIR as a Distribution

Two values are needed to compute TRIR: incidents and time. Incidents are discrete events; they happen or they do not, and there is no such thing as a fractional or negative incident. However, time is a continuous value and infinite possibilities exist. To work with time as a variable, it is often defined in ranges such as days, hours or minutes. If we look at each work hour (the base unit used in TRIR) as a discrete event and an incident as a binary possibility, this yields what is known as a Bernoulli trial.

A Bernoulli trial is built upon three assumptions (Wardrop, 2010):

•Assumption 1: Each trial results in one of two possible outcomes (occurrence or nonoccurrence). Incidents satisfy this assumption because they either occur or do not occur in one work hour. Although it is possible to have more than one recordable incident in a work hour, it is so rare as to be mathematically negligible.

•Assumption 2: The probability (p) of an injury remains constant from one work hour to the next. It is assumed each work hour has approximately the same (exceedingly low) probability of injury. Although certain situations have a higher probability of injury than others, over enough time exposure each work hour can be represented as approximately the same. This is not to say that all work types pose equal risk, only that the risk is evenly distributed over long periods.

•Assumption 3: The trials (work hour) are independent. Independence in a data set means that there is not a well-established connection in the outcomes among trials (work hours). For TRIR, there are no known patterns in occurrence, especially when the data are considered in an extremely high number of trials (e.g., hundreds of thousands of work hours). Although an argument can be made that an incident in one work hour makes an incident in the next more or less likely, independence remains a reasonable assumption.

Now that TRIR is understood to be logically represented as a series of Bernoulli trials, the distribution that best represents the context must be identified. Because recordable incidents can only occur or not occur, the distribution of TRIR must be discrete. In other words, the number of incidents observed over any period must be a whole number. Although the binomial distribution represents the distribution of potential outcomes from a sample of Bernoulli trials, the Poisson distribution was used instead of a binomial distribution because it is an even more accurate representation of TRIR.

A Poisson distribution can be thought of as a unique case of the binomial distribution where the probability of occurrence is extremely small. The Poisson distribution is reasonable to use when there are at least 20 trials and the probability of occurrence is less than or equal to 5% (Prins, 2012). In the case of TRIR at a basic unit level (incidents and work hours), both are true because incidents are rare and TRIR is typically measured over at least thousands of work hours. Another key benefit of the Poisson distribution is that it expresses the probability of a given number of events occurring in a fixed interval of time. This means that the Poisson distribution can be scaled to virtually any time range.

Computing Confidence Intervals

When a value is observed from a distribution of possible observations, the result should be communicated as a confidence interval. A confidence interval is the range of values that is likely to contain the true value with some degree of confidence. The level of confidence is expressed as a probability that the true value lies between an upper and lower interval. For example, TRIR could be expressed as a value that is contained in a range between values X and Y with 95% confidence. This can also be practically interpreted as a 5% probability that a long-term TRIR would be outside this range. Confidence intervals are computed using equations that best represent the underlying distribution. For a Poisson distribution, the Wilson confidence interval is the most appropriate approximation (Wallis, 2013). The upper and lower bounds of the Wilson confidence interval are represented by Equation 2 (Wallis, 2013).

Equation 2:

$$\frac{\hat{p} + \frac{z^2}{2n}}{1 + \frac{z^2}{n}} \pm \frac{z}{1 + \frac{z^2}{n}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}$$

where $\hat{p}$ is the number of actual events (incidents) divided by the number of trials (work hours) and z is the critical value of a standard normal distribution corresponding to the target confidence interval (e.g., $\alpha = 0.05$ for a 95% confidence interval). In this analysis a significance level (α) of 0.05 was always selected so the corresponding z is 1.96.

Although there are other approximations of Poisson confidence intervals that are more accurate (e.g., Garwood, 1936; Ulm, 1990), the Wilson confidence interval is simple to compute with a basic calculator and provides an approximation within 1% of the more computationally demanding methods. For example, the "exact" method from Ulm (1990) requires a statistical package to analyze and produces results that are practically indistinguishable from the Wilson interval shown in Equation 2.

The Wilson confidence interval allows one to judge the precision of an observed TRIR using only two pieces of information: the number of incidents and the number of work hours in the sample. Once this information is known, the confidence interval can be approximated using Equation 2. Multiplying the endpoints by 200,000 shows the range in terms of TRIR as it is normally reported.

An example provides some clarity for the layperson. If a theoretical company had one recordable injury in 200,000 work hours (n = 200,000, $\hat{p}$ = 1/200,000), we can calculate the 95% confidence interval using Equation 2 as 0.18 to 5.66 injuries per 200,000 work hours. Theoretically, this corresponds to the range of results that the company's safety system is designed to produce that month. A TRIR below 0.18 would be interpreted as unusually low and a TRIR above 5.66 as unusually high. Therefore, reporting a TRIR of 1.00 per 200,000 is not appropriate or meaningful. Rather, the TRIR should be reported as an interval of 0.18 to 5.66 with the most likely true value of 1.00. We can also ask the complementary question: Assuming the company's true injury rate is one per 200,000 work hours, how many injuries are likely to occur over future periods of 200,000 work hours if the safety system remains the same? Figure 1 shows the range of potential results and their probabilities. Figure 1 was created using Equation 3, which is the probability mass function of the Poisson distribution. The equation determines the probability of exactly N injuries based on an observed rate of λ injuries within a given period (e.g., 200,000 work hours).

FIGURE 1

PROBABILITY OF EXPERIENCING *N* INCIDENTS IN 200,000 WORK HOURS

Probability of experiencing *N* incidents in the next 200,000 work hours assuming a true TRIR of one recordable injury in 200,000 work hours.

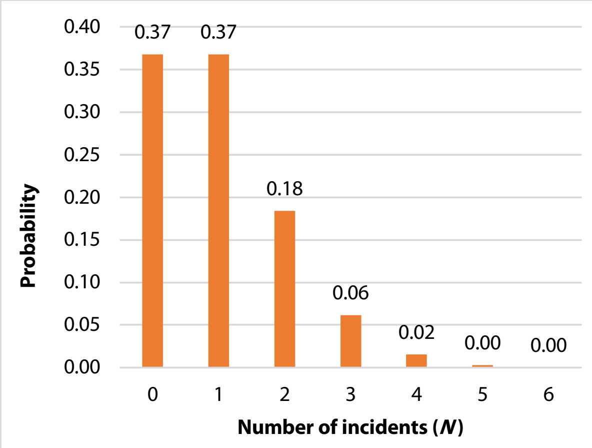
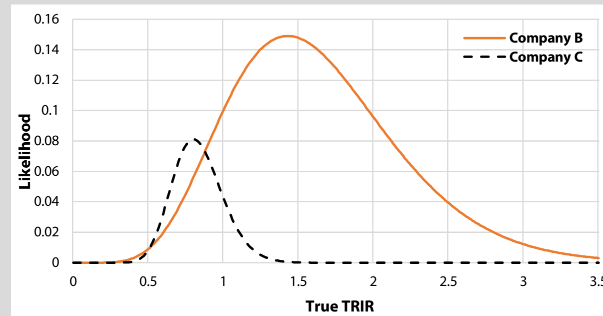


FIGURE 2

TRIR DISTRIBUTIONS FOR COMPANIES B & C

Likelihood of the true injury rates for Companies B and C (per 200,000 hours), based on their observed histories.



Equation 3:

$$\frac{\lambda^N e^{-\lambda}}{N!}$$

where N equals the number of injuries on the horizontal axis of Figure 1 and λ equals the observed number of injuries

As an extension of the preceding example, if the same company experienced two injuries in the next 200,000 work hours, the 95% interval would then be 0.55 to 7.29 per 200,000 work hours. The company might be concerned that it doubled its injuries from the previous interval from 1.00 to 2.00 per 200,000 work hours. However, a test for significance shows that there is no statistical difference between the months even though the number of injuries doubled. This result means that the difference in the count of injuries alone does not reveal anything significant about the difference in the safety system between the two periods.

Analysis of Case Examples

Using the three case examples from the background (Companies A, B and C), we can explore how reporting TRIR as a confidence interval vastly changes its interpretation and meaning.

- Company A has a recordable injury in the first 1,000 work hours that the company is in business. TRIR range: 35.31 to 1,128.51 per 200,000 work hours.
- Company B has seven recordable injuries over 980,000 work hours in a given year. TRIR range: 0.69 to 2.95 per 200,000 work hours.
- Company C has 24 recordable injuries over 6 million work hours in a given year. TRIR range: 0.54 to 1.19 per 200,000 work hours.

For example, the TRIR of Company A, which had a TRIR of 200 resulting from one recordable injury in 1,000 work hours, would be correctly reported as 35 to 1,128 with 95% confidence. This statistically confirms the logical interpretation that this new small company does not have enough exposure time to return a meaningful TRIR. The comparison of Companies B and C is also interesting. Since both companies reported their TRIR over many work hours, it may seem appropriate to recognize their difference in TRIR as meaningful. However, as shown in Figure 2, the distribution of Company C fits entirely within the distribution of Company B indicating that the two injury records are statistically indistinguishable. That said, another important conclusion can be made: the TRIR for Company C can be stated with a much higher precision (smaller interval) than Company B. Therefore, there is a distinct advantage to having a greater number of work hours accumulated. However, it still remains unclear whether the historical TRIR has any bearing on future TRIR, which was the primary subject of the nonparametric empirical analysis.

Tables 1 and 2 (p. 32) are designed to assist practitioners with the interpretation of TRIR as a range. Table 1 shows the relationship between precision (width of the confidence interval), TRIR and exposure time (work hours), computed using Equation 2. Simply, it shows how many work hours are needed to achieve a given precision level for various TRIR scenarios. The number of hours needed for a given precision can be approximated by Equation 4, which is based upon the Wilson binomial formula (Krishnamoorthy & Peng, 2007).

Equation 4:

$$n \cong \frac{z_\alpha^2\left[(pq - 2d^2) + \sqrt{(pq - 2d^2)^2 - d^2(4d^2 - 1)}\right]}{2d^2}$$

where:

$$d = \frac{desired\ TRIR\ precision}{200{,}000 \cdot 2}$$

and n is the required number of work hours to achieve precision d given a probability of injury in each hour p, and $q = 1 - p$.

Table 1 (p. 32) reveals some potentially surprising results. For example, if an organization wishes to report a TRIR of 1.00 with a precision of 0.1 (e.g., 0.95 to 1.05 per 200,000 work hours), about 300 million work hours of exposure time is required. In other words, unless the TRIR of 1.0 is derived from 300 million work hours of exposure time, it should not be reported to even one decimal point. Furthermore, reporting two meaningful decimal places for TRIR requires approximately 30 billion work hours of data. The required number of work hours is so high because recordable injuries occur, on average, so infrequently that they do not produce statistical stability. This raises questions about reporting TRIR to two decimal points and making important business decisions with this level of granularity.

To further illustrate how TRIR should be interpreted, Table 2 (p. 32) shows the 95% confidence interval for a series of scenarios with varying TRIR and number of work hours. For example, Table 2 shows that if an organization measured its TRIR to be 1.0 over a period of 1 million work hours (i.e., five recordable injuries

TABLE 1

RELATIONSHIPS AMONG PRECISION (WIDTH OF CONFIDENCE INTERVAL), TRIR & EXPOSURE TIME (WORK HOURS)

Precision	TRIR	Work hours	Precision	TRIR	Work hours	Precision	TRIR	Work hours
0.1	0.20	62,409,083	0.25	0.20	10,715,491	0.5	0.20	3,197,046
0.1	0.40	123,404,751	0.25	0.40	20,137,222	0.5	0.40	5,357,737
0.1	0.60	184,709,031	0.25	0.60	29,819,030	0.5	0.60	7,682,889
0.1	0.80	246,092,232	0.25	0.80	39,575,018	0.5	0.80	10,068,588
0.1	1.00	307,507,116	0.25	1.00	49,361,749	0.5	1.00	12,481,762
0.1	1.25	384,297,072	0.25	1.25	61,616,226	0.5	1.25	15,517,884
0.1	1.50	461,099,608	0.25	1.50	73,883,277	0.5	1.50	18,566,029
0.1	1.75	537,909,256	0.25	1.75	86,157,534	0.5	1.75	21,621,176
0.1	2.00	614,723,280	0.25	2.00	98,436,299	0.5	2.00	24,680,748
0.1	3.00	922,000,301	0.25	3.00	147,573,796	0.5	3.00	36,941,358

TABLE 2

95% CONFIDENCE INTERVALS FOR A SERIES OF TRIR SCENARIOS

	TRIR 95% confidence interval								
Work hours	**0.1**	**0.2**	**0.5**	**1.0**	**1.5**	**2.0**	**3.0**	**4.0**	**5.0**
100,000	0.00 to 7.68	0.00 to 7.68	0.00 to 7.68	0.00 to 7.68	0.00 to 7.68	0.35 to 11.33	0.35 to 11.33	1.10 to 14.59	1.10 to 14.59
250,000	0.00 to 3.07	0.00 to 3.07	0.00 to 3.07	0.14 to 4.53	0.14 to 4.53	0.44 to 5.83	0.82 to 7.06	1.71 to 9.36	2.20 to 10.47
500,000	0.00 to 1.54	0.00 to 1.54	0.07 to 2.27	0.22 to 2.92	0.41 to 3.53	0.85 to 4.68	1.36 to 5.78	2.17 to 7.36	2.75 to 8.39
1 million	0.00 to 0.77	0.04 to 1.13	0.11 to 1.46	0.43 to 2.34	0.68 to 2.89	1.09 to 3.68	1.82 to 4.95	2.59 to 6.18	3.39 to 7.38
2.5 million	0.01 to 0.45	0.04 to 0.58	0.22 to 1.05	0.55 to 1.68	0.91 to 2.28	1.35 to 2.95	2.15 to 4.08	3.03 to 5.27	3.87 to 6.36
5 million	0.02 to 0.29	0.09 to 0.47	0.27 to 0.84	0.68 to 1.48	1.07 to 2.04	1.52 to 2.64	2.39 to 3.76	3.29 to 4.86	4.20 to 5.96
10 million	0.04 to 0.23	0.11 to 0.37	0.34 to 0.74	0.76 to 1.32	1.20 to 1.88	1.64 to 2.43	2.56 to 3.52	3.48 to 4.59	4.42 to 5.66
20 million	0.05 to 0.18	0.13 to 0.31	0.38 to 0.66	0.82 to 1.22	1.28 to 1.76	1.74 to 2.30	2.68 to 3.36	3.63 to 4.41	4.58 to 5.46
50 million	0.07 to 0.15	0.15 to 0.26	0.42 to 0.60	0.88 to 1.13	1.36 to 1.66	1.83 to 2.18	2.79 to 3.22	3.76 to 4.26	4.73 to 5.28

occurred over the span of 1 million work hours), the correct interpretation is that the safety system is designed to produce a TRIR between 0.43 and 2.34. Note that in Table 2 some scenarios are not possible because they would not correspond to a whole number of incidents; however, they are included to maintain continuity.

Nonparametric Analysis

To complement the parametric analysis and to add a degree of validation, an empirical analysis was performed with injury data provided by 10 construction organizations that are members of the Construction Safety Research Alliance. The partner organizations represent infrastructure, power generation and delivery, and commercial building sectors. The partners provided monthly counts of recordable injuries, fatalities and work hours for a 15-year period. In total, the data set included 3.26 trillion work hours of data. The empirical analysis of these data is summarized here, and the analytical approach is explained in fine detail in Salas (2020).

Random & Unpredictable Nature of TRIR

To assess the significance of TRIR, the data for each organization were fit to a generalized linear model using different distribution functions to determine the best model (i.e., the one with the lowest Akaike information criterion). The best fit model was then subjected to 100 repeated sampling and tenfold cross-validation to measure the strength, stability and response of the model to the observed data. Then, a Monte Carlo simulation was used to evaluate how much of the final result is due to random variation. Holding the data generating process constant, the Monte Carlo simulation tested how different estimators fare in trying to uncover underlying parameters. Put simply, this method involved using the historical TRIR data for each organization to create a predictive equation, then testing how well that equation correctly estimated future TRIR.

The results indicated that TRIR is 96% to 98% random. This means that the best model was only able to predict the observed TRIR in 2% to 4% of the trials. Since safety is a chaotic system due to the interfaces between people, culture, policies, regulations, equipment and other external factors (e.g., economy, weather, natural events), this is not a surprising finding. However, it does have important implications. First, it provides empirical evidence to support the logical assumptions made in the parametric analysis. Second, the fact that TRIR is random in nature provides further evidence that it must be reported as a range (i.e., confidence interval) and absolutely should not be expressed as a single number.

In addition to measuring the significance, the models were also tested to assess the extent to which historical TRIR predicts future TRIR. The results showed that at least 100 months of data were required to achieve reasonable predictive power because of the high degree of random variation. Since TRIR is generally used to make comparisons or decisions on the order of months or years, this finding indicates that for all practical purposes, TRIR is not predictive. For example, a client that hires a contractor with a TRIR of 0.75 cannot reasonably expect that the contractor will achieve that same performance on its upcoming project.

Lack of Relationship Between TRIR & Fatalities

Effect measurements revealed that variation in TRIR has no association with fatalities. That is, trends in TRIR do not associate statistically with fatality occurrence. Instead, fatalities appear to follow different patterns, suggesting that they occur for different reasons. This finding challenges the long-standing assumption, derived from the Heinrich (1959) safety pyramid, that injuries of different severity levels exist at fixed ratios and have the same underlying causes. It also debunks the notion that reducing TRIR is a surrogate for mitigating the risk of high-impact events.

Key Findings

This study challenges conventional wisdom of safety measurement with an empirical analysis of 3.26 trillion work hours of TRIR data and a statistical demonstration of proof. The results reveal strong evidence that TRIR is almost entirely random and is not indicative of future performance unless millions of work hours are amassed. The specific conclusions that follow logically and empirically are as follows:

1. TRIR is not associated with fatalities. The effect measurements revealed that there is no discernable association between fatalities and TRIR. Recordable injuries and fatalities follow different patterns and occur for different reasons. Thus, TRIR trends are not a proxy for high-impact incidents. Hence, it can be inferred that safety interventions, including policies, regulations and management systems, associated with the improvement in TRIR performance may not necessarily prevent fatalities.

2. TRIR is almost entirely random. Empirical analysis revealed that changes in TRIR are due to 96% to 98% random variation. This is logically confirmed by the fact that recordable injuries do not occur in predictable patterns or regular intervals. This is likely because safety is a complex phenomenon that is impacted by many factors.

3. TRIR cannot be represented as a single point estimate. Since TRIR is almost entirely random, a single number does not represent the true state of safety performance. Instead, TRIR is best expressed as a confidence interval and studied over extended periods. For example, a yearly TRIR value of 1.29 is statically meaningless for almost every organization. This finding was initially explored in the parametric analysis and was empirically validated in the nonparametric analysis.

4. TRIR is not precise and should not be communicated to multiple decimal points. Unless hundreds of millions of work hours are amassed, the confidence bands are so wide that TRIR cannot be accurately reported to even one decimal point. The implication is that the TRIR for almost all companies is virtually meaningless because they do not accumulate enough work hours.

5. If an organization uses TRIR for performance evaluations, it is likely rewarding nothing more than random variation. Because of the random nature of TRIR, it is unclear whether a change in performance (positive or negative) is due to an underlying change in the safety system or if the organization is simply observing random variation.

6. TRIR is predictive only over extremely long periods. Other researchers have postulated that TRIR can be used as a predictor of performance when taken over extremely long periods (Alruqi & Hallowell 2019; Lingard et al., 2017; Salas & Hallowell, 2016; Wilbanks 2019). The empirical results of this study confirmed that TRIR is only predictive when more than 100 months of TRIR data are accumulated.

The results of this study may not be surprising to some professionals who have made these postulations for years. However, this is the first scientific evidence that explains why TRIR is not a valid comparative measure of safety performance.

Conclusions & Recommendations

The conclusions of this research may disrupt how the profession approaches safety measurement and reporting. Therefore, along with the senior executives from our industry partners, the authors offer the following practical interpretation of the results.

1. TRIR should not be used as a proxy for serious injuries and fatalities. For years, many practitioners have used declining trends in TRIR to indicate the mitigation of fatality risk. However, the lack of statistical association between TRIR and fatalities suggests that the assumption holds no scientific merit. Practitioners should consider creating and testing targeted measurement, learning and prevention efforts for serious injuries and fatalities.

2. TRIR should not be used to track internal performance or compare companies, business units, projects or teams. Since the average company requires tens of millions of work hours to return a confidence interval with one decimal point of precision, organizations should be especially careful making any comparisons using TRIR. For example, most companies do not even have enough work hours to detect statistically significant changes in TRIR from one year to the next. Therefore, the authors challenge practices where TRIR is used to compare companies, projects or teams. At best, TRIR is only useful for comparing industries or sectors of the U.S. economy over long periods. For the same reason, TRIR should not be used as the primary safety metric when incentivizing organizational performance or comparing or prequalifying contractors.

3. The safety profession must change how it communicates TRIR. TRIR is almost always communicated as a precise number as if it was the only possible outcome (Stricoff, 2000). Since recordable injuries are so infrequent and are a product of so much random variation, a single precise number is meaningless. Instead, TRIR should be accompanied by the range of potential outcomes that the safety system could have reasonably produced. The implication is that small companies would report large confidence intervals (high uncertainty) and large companies would report smaller confidence intervals (lower uncertainty). However, almost no company would be able to appropriately report TRIR to the level of precision most commonly used today.

4. Companies should not place much emphasis on short-term changes in TRIR. It is tempting to track TRIR over time to identify when performance is improving or degrading. However, as observed in the empirical analysis, changes that occur from month to month are mostly random and do not necessarily reflect any actual change in the safety system.

5. TRIR should not be used to measure the impact of safety interventions. Managers are often compelled to show reductions in TRIR resulting from safety initiatives and investments. However, controlled experiments and longitudinal data are needed to establish causal inference between safety interventions and TRIR trends. This renders TRIR entirely inadequate for attributing change because most companies drastically shift their management approaches, safety programs and even their business models over the long time frames needed to produce a precise and stable TRIR.

6. New approaches to safety measurement are needed. In addition to statistical invalidity, the use of TRIR also does not describe why the performance, good or bad, was achieved and what can be done to improve. This leaves organizations wondering, "Are we truly good or simply lucky?" or worse, "Are we truly bad, or do we simply need to log more work hours?" The academic and professional community should consider alternative measures of safety performance that assess the actual safety system at high frequency. Increasing the number of reliable measurements could drastically improve the stability, precision and predictive nature of safety metrics. To be comparative, however, these metrics must be standardized and consistently reported.

Since TRIR has remained the most pervasive measure of safety for nearly 50 years, this study underscores the need to scientifically test even the most basic assumptions of the safety profession. In the spirit of scientific inquiry, the authors recommend that other researchers propose alternative hypotheses about TRIR, conduct independent tests and challenge the assumptions made in this article. Although the authors stand by our conclusions, they recognize that other perspectives may generate different models and results. PSJ

References

Alruqi, W.M. & Hallowell, M.R. (2019). Critical success factors for construction safety: Review and meta-analysis of safety leading indicators. *Journal of Construction Engineering and Management, 145*(3). https://doi.org/10.1061/(ASCE)CO.1943-7862.0001626

Garwood, F. (1936). Fiducial limits for the Poisson distribution. *Biometrika, 28*(3-4), 437-442. https://doi.org/10.1093/biomet/28.3-4.437

Hallowell, M. (2010). Cost effectiveness of construction safety program elements. *Construction Management and Economics, 28*(1), 25-34. https://doi.org/10.1080/01446190903460706

Hallowell, M.R. & Gambatese, J.A. (2009). Construction safety risk mitigation. *Journal of Construction Engineering and Management, 135*(12), 1316-1323. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000107

Hallowell, M.R., Hinze, J.W., Baud, K.C. & Wehle, A. (2013). Proactive construction safety control: Measuring, monitoring, and responding to safety leading indicators. *Journal of Construction Engineering and Management, 139*(10). https://doi.org/10.1061/(ASCE)CO.1943-7862.0000730

Heinrich, H.W. (1959). *Industrial accident prevention: A scientific approach* (4th ed.). McGraw-Hill.

Hinze, J., Thurman, S. & Wehle, A. (2013). Leading indicators of construction safety performance. *Safety Science, 51*(1), 23-28. https://doi.org/10.1016/j.ssci.2012.05.016

Krishnamoorthy, K. & Peng, J. (2007). Some properties of the exact and score methods for binomial proportion and sample size calculation. *Communications in Statistics—Simulation and Computation, 36*(6), 1171-1186. https://doi.org/10.1080/03610910701569218

Lingard, H., Hallowell, M., Salas, R. & Pirzadeh, P. (2017). Leading or lagging? Temporal analysis of safety indicators on a large infrastructure construction project. *Safety Science, 91*, 206-220. https://doi.org/10.1016/j.ssci.2016.08.020

Lofquist, E.A. (2010). The art of measuring nothing: The Paradox of measuring safety in a changing civil aviation industry using traditional safety metrics. *Safety Science, 48*(10), 1520-1529. https://doi.org/10.1016/j.ssci.2010.05.006

Manuele, F.A. (2008, Dec.). Serious injuries and fatalities A call for a new focus on their prevention. *Professional Safety, 53*(12), 32-39.

OSHA. (2010, Feb. 16). Standard interpretation: Clarification of assigned working hours when recording work-related injuries/illnesses. www.osha.gov/laws-regs/standardinterpretations/2010-02-16

OSHA. (2016, Aug. 23). Standard interpretation: Clarification on how the formula is used by OSHA to calculate incident rates. www.osha.gov/laws-regs/standardinterpretations/2016-08-23

Prins, J. (2012). Counts control charts. In C. Croarkin & P. Tobias (Eds.), *NIST/SEMATECH e-Handbook of Statistical Methods* (Chapter 6.3.3.1.). National Institute of Standards and Technology. https://doi.org/10.18434/M32189

Ritchie, N.W. (2013). Journey to zero: Aspiration versus reality. Paper presented at the European HSE Conference and Exhibition, London, U.K. https://doi.org/10.2118/164963-MS

Salas, R. & Hallowell, M. (2016). Predictive validity of safety leading indicators: Empirical assessment in the oil and gas sector. *Journal of Construction Engineering and Management, 142*(10). https://doi.org/10.1061/(ASCE)CO.1943-7862.0001167

Salas, R.I. (2020). *Empirical evaluation of construction safety: Insights from economic trends, national cultures and workplace injury rates* (Publication No. 27835552) [Doctoral dissertation, University of Colorado at Boulder]. ProQuest Dissertations and Theses Global.

Stricoff, R.S. (2000, Jan.). Safety performance measurement: Identifying prospective indicators with high validity. *Professional Safety, 45*(1), 36-39.

Toellner, J. (2001, Sept.). Improving safety and health performance: Identifying and measuring leading indicators. *Professional Safety, 46*(9), 42-47.

Truitt, D. (2012, March). Contractor prequalification. *Professional Safety, 57*(3), 34-35.

U.S. Bureau of Labor Statistics (BLS). (2019). Injuries, illnesses and fatalities. www.bls.gov/iif

Ulm, K. (1990). Simple method to calculate the confidence interval of a standardized mortality ratio (SMR). *American Journal of Epidemiology, 131*(2), 373-375. https://doi.org/10.1093/oxfordjournals.aje.a115507

Wallis, S. (2013). Binomial confidence intervals and contingency tests: Mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics, 20*(3), 178-208. https://doi.org/10.1080/09296174.2013.799918

Wanberg, J., Harper, C., Hallowell, M.R. & Rajendran, S. (2013). Relationship between construction safety and quality performance. *Journal of Construction Engineering and Management, 139*(10). https://doi.org/10.1061/(ASCE)CO.1943-7862.0000732

Wardrop, R. (2010). [Course notes on Bernoulli trials]. Department of Statistics, University of Wisconsin-Madison. http://pages.stat.wisc.edu/~wardrop/courses/371achapter2a.pdf

Wilbanks, D.W. (2019, July). Contractor safety prequalification: The significant limitations of loss rates. *Professional Safety, 64*(7), 22-26.

Matthew Hallowell, Ph.D., is executive director of the Construction Safety Research Alliance (CSRA). He is also a president's teaching scholar and endowed professor of construction engineering at the University of Colorado. Outside academia, he is the is the founder and executive director of Safety Function LLC and serves as a technical advisor to the Edison Electric Institute and Interstate Natural Gas Association of America (INGAA). Hallowell is a professional member of ASSP's Rocky Mountain Chapter.

Mike Quashne, CUSP, is a member of the CSRA Board of Advisors and a manager of business transformation and performance assessment for Baltimore Gas and Electric. He holds a B.S. in Aeronautical Engineering from the U.S. Air Force Academy and an M.S. in Operations Research from the Air Force Institute of Technology. Quashne is a member of ASSP's Chesapeake Chapter and Utilities Practice Specialty.

Rico Salas, Ph.D., is a research associate with CSRA. He earned an M.S. from the University of Cambridge and a Ph.D. from the University of Colorado at Boulder. He has an extensive technical and work experience in design, construction and process safety in the oil and gas industry. He is the asset integrity engineering manager for a leading U.S. oil and gas company.

Matt Jones, Ph.D., is associate professor of psychology and neuroscience and director of the Center for Research on Training at the University of Colorado at Boulder. His research develops mathematical and statistical methods for understanding human behavior.

Brad MacLean is chair of CSRA and senior vice president of safety for the Wolfcreek Group. He has been a health, safety and environment practitioner in the energy sector since 1990, working in Canada and the U.S. MacLean is a founding member and past chair of the INGAA Foundation Safety Committee. MacLean is a member of ASSP's Gulf Coast Chapter.

Ellen Quinn has more than 20 years of executive experience serving as the lead for EHS in global commercial operations and utilities including Carrier and The Otis Elevator Co. Her specialties include mitigation of high-risk activities to eliminate fatalities and serious injuries. She served as the vice chair of CSRA and was a founding member of the National Safety Council's Campbell Institute.

Acknowledgments

This research was made possible by the generous funding provided by the members of the Construction Safety Research Alliance (CSRA), a nonprofit organization that supports collaboration between industry and academia on safety research and development. The authors thank the CSRA industry partners for providing data, advice and inspiration for this work.