Applied Ergonomics 44 (2013) 659–666

Contents lists available at SciVerse ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo

Predictive models of safety based on audit findings: Part 2: Measurement of model validity

Yu-Lin Hsiao [a,*], Colin Drury [b], Changxu Wu [b], Victor Paquet [b]

[a] Department of Industrial and Systems Engineering, Chung Yuan Christian University, Chung Li 32023, Taiwan
[b] Department of Industrial and Systems Engineering, State University of New York (SUNY) at Buffalo, Buffalo, NY 14260, USA

ARTICLE INFO

Article history:
Received 23 September 2011
Accepted 4 January 2013

Keywords:
Human error
HFACS-MA
Safety prediction
Neural network
Aviation maintenance

ABSTRACT

Part 1 of this study sequence developed a human factors/ergonomics (HF/E) based classification system (termed HFACS-MA) for safety audit findings and proved its measurement reliability. In Part 2, we used the human error categories of HFACS-MA as predictors of future safety performance. Audit records and monthly safety incident reports from two airlines submitted to their regulatory authority were available for analysis, covering over 6.5 years. Two participants derived consensus results of HF/E errors from the audit reports using HFACS-MA. We adopted Neural Network and Poisson regression methods to establish nonlinear and linear prediction models respectively. These models were tested for the validity of prediction of the safety data, and only Neural Network method resulted in substantially significant predictive ability for each airline. Alternative predictions from counting of audit findings and from time sequence of safety data produced some significant results, but of much smaller magnitude than HFACS-MA. The use of HF/E analysis of audit findings provided proactive predictors of future safety performance in the aviation maintenance field.

1. Introduction

In our lifetime, we have seen the change from basing safety on accident investigation to using more proactive safety measures such as recurrent audits. These proactive ways are carefully designed to focus attention on what their designers expect: to be the measurable precursors of accidents or incidents. It is assumed that the correct precursors are being measured, and the results are interpreted correctly as visible threats to system safety. However, in many ways, no quantifiable or data-driven efforts are undertaken to improve these audit findings before conditions are right for the same failures or critical events to occur. The phenomenon that threats often lie dormant is not because responsible managers didn't recognize the existence of these threats, but is for they have only limited resources to solve numerous problems or don't appreciate the necessity or urgency of correction. In other words, a more reasonable clue is required to convince managers to decide which threats are most eminent/serious.

To help improving this common occurrence among safety management in industry, we consider clarifying how these identifying failures in daily audits could possibly (probability) corrode the system safety in magnitude (severity) and result in economic loss (such as the cost of incidents). This idea is based on the risk management concept. We aim to investigate the topic from a more quantitative and objective standpoint. The first step is to quantitatively prove the causal relationship between audit findings as precursors and safety measures.

Do audit findings really predict the future safety performance of an organization? This question has not been answered for any system where audits are in use. What we could perceive is that audits are treated as indicators of systematic safety status in a qualitative way. This article, Part 2 of our study, shows how we tested the assumption about audits' predictive capability, using monthly data on safety performance regarding the maintenance activities of two different airlines.

In Part 1 (Hsiao et al., 2013), we developed a rationale for quantifying audit reports, using human factors/ergonomics (HF/E) analysis, rather than merely counting audit findings which had proven ineffective. Based on aviation safety research (Shappell and Wiegmann, 2003), on general system safety (Reason, 1990) and on management theories, e.g. Robbins and Coulter (2005), a classification system for human errors was developed and tested for reliability. We named our modified taxonomy the HFACS-Maintenance Audit (HFACS-MA, see Fig. 1) and found its reliability to be satisfactory. HFACS-MA used audit reports collected by the regulatory authority in the country where the airlines are

* Corresponding author.
E-mail address: yhsiao@cycu.edu.tw (Y.-L. Hsiao).

0003-6870/$ – see front matter © 2013 Elsevier Ltd and The Ergonomics Society. All rights reserved.
http://dx.doi.org/10.1016/j.apergo.2013.01.003

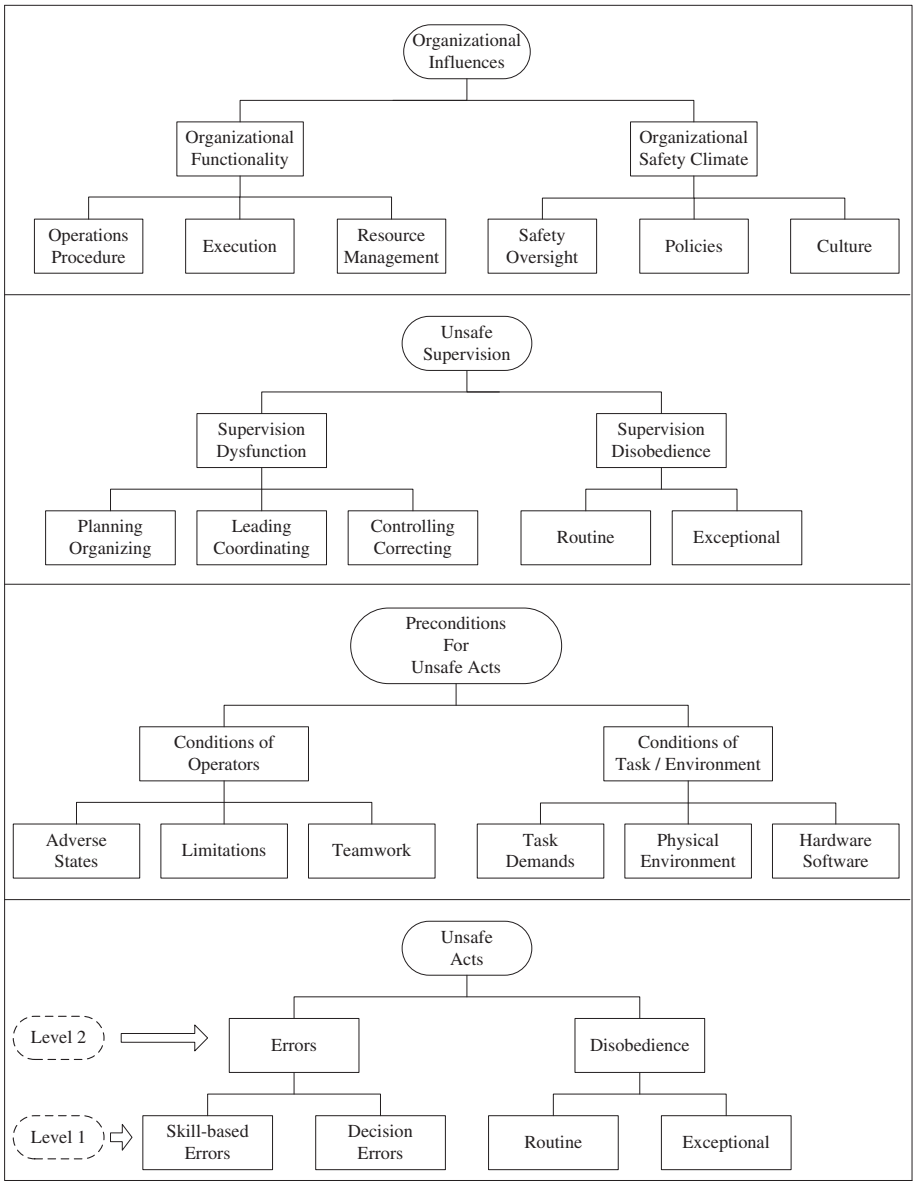

Fig. 1. The complete framework of HFACS-MA.

registered. In Part 2, we tested HFACS-MA against safety performance data collected by the same regulatory authority using a system independent of the audit data collection.

There have been attempts to validate human factors audit programs in the past but never against future safety performance. Drury (2006) noted that such validity can be assessed by content validity, concurrent validity, or construct validity. A later revision of the Drury's (2006) chapter (Drury and Dempsey, 2012) quotes a health and safety audit evaluated against safety standards (Robson et al., 2010) as an example of content validity, and a comparison between audit findings and expert findings in airline maintenance (Koli et al., 1995) as one of concurrent validity. Those reviews also noted that construct validity has not been prominent in the testing of audit systems. Overall, out of 17 audit systems or checklists for human factors issues, only 6 had been validated, and none of them had been validated against safety outcome data. Thus the results of the current study should be of interest beyond the aviation maintenance domain, although that is our primary concern.

Along with quantifying the threats of human errors in maintenance systems from audit reports, this research provides an opportunity to examine the relationship between human failures and overall safety performance at a more detailed level. Assuming that the safety performance of an organization reflects the external manifestation of their internal error environment, we postulate that a causal link should exist between overall safety performance and characteristics of the human failures. This answers the obvious question: "is safety performance predictable based on the analysis of human error from audit records?" and also potentially provides guidance on which HF/E failures to address.

2. Data collection and quantification

In this study, there are two major data sources: historical audit records as predictors and incident reports as outcome measures. We transformed them into weighted human error rates and incident rates respectively to represent the status of human failures and the safety performance of maintenance systems. The weighted

human error rates were extracted from audit reports using the error categories of HFACS-MA, and the incident rates were directly counted from incident reports which were related to maintenance issues. The audit records and incident reports are independent data sources, and are obtained from different systems within the civil aviation authority which has requested confidentiality. There are total 1245 audit records covering a 6.5 years period from two airlines (called here airlines A and B). Airline A had 73 incidents, and airline B had 192 incidents during this specified time period.

The weighted human error rates and the incident rates were quantitatively measured within monthly periods, rather than weekly, quarterly or annually. This decision was made to balance the intrinsic characteristic (for weekly rate) and the statistical power (for quarterly and annual rate) of the raw data. Monthly period gives sufficient data points while avoids counting excessive numbers of days or weeks where no audits were taken. Although both the audit records and incident reports contained the accurate time at which the event happened, which made a transition to any time period possible, the number of flight departures were reported to the aviation authority at a monthly basis. Therefore, monthly rate was the natural choice.

2.1. Dependent variable – safety indicator

A quantitative safety indicator is needed to represent the future safety performance. As defined by the National Transportation Safety Board (NTSB) and Federal Aviation Administration (FAA) in the United States, an accident is "an occurrence associated with the operation of an aircraft which takes place between the time any person boards the aircraft with the intention of flight and all such persons have disembarked, and in which any person suffers death or serious injury, or in which the aircraft receives substantial damage". And incident is defined as "an occurrence, other than an accident, associated with the operation of an aircraft, which affects or could affect the safety of operations" (NTSB, 1988) (p. 1). Accident rate has been commonly accepted by the aviation industry as the universal safety performance indicator to depict a general safety picture of an airline, a country, or a geographic region such as Asia, Europe, or North America. However, because flight accidents rarely happen (usually less than one accident per airline per year), accident rate is not sensitive enough to be used as a safety indicator in our study: accident rates could remain constant or zero in whole year for an airline (Ding and Ru, 2008; Liou et al., 2008). Thus, incident rate is more typically used as it is based on far more data in specific time period, and incidents are seen as necessary precursors to accidents.

Briefly, the overall incident rate of the accessed time period (6.5 years) was 0.90 for airline A and 1.42 for airline B (per 1000 departures). And the accident rate of the two airlines over the same period was 0.012 and 0.015 respectively. It is noted that the accident/incident ratio was comparable for the two airlines at 0.013 for airline A, and 0.010 for airline B.

In this study, future incident rate served as the safety indicator (dependent variable), defined as:

$$\text{Incident Rate} = \frac{\sum \text{Incident}}{\text{Departure}} \times 1{,}000 \quad (1)$$

$\sum$ Incident: The number of incidents per month
Departure: The number of flight departure per month

In order to examine the predictive ability of the model over different time periods, i.e. to find out the best time range of forecasting, the safety indicator (incident rate) was tested for one-, two- and three-month following the time point of each analysis. For instance, the weighted human error rates in January rates were used to predict the future incident rate in either the next month (February), the next two months (February and March), or the next three months (February, March and April).

2.2. Independent variables

The independent variables of the prediction model comprise the weighted human error rates derived from audit findings via HFACS-MA, and both past and current incident rates. Incident rates were included to test whether future safety performance was simply predictable from past safety performance, i.e. whether or not the HF/E analysis using HFAC-MA add any value to the prediction. We recruited two raters to generate the weighted human error rates from qualitative text descriptions in the audit reports to quantitative data. The two raters read through each audit record to find existing failures, and diagnosed any flaws they found into specific human error categories in HFACS-MA. After their independent analysis, the raters recorded their results separately. If the results were different, the raters were to discuss the case to reach a consensus conclusion on each audit report.

Because the analysis process involved non-English documents and many human error concepts and definitions, two graduate students, who already possess both the native language and a human factors background, were recruited. HFACS-MA is the diagnostic framework of human error comprising four major human error tiers: Unsafe Act, Precondition of Unsafe Act, Unsafe Supervision, and Organizational Influence.

Since some specific errors in HFACS-MA were rare to find out in the historical audit reports, the appearance frequencies of these unusual failures were mostly zero while transformed to weighted human error rates in monthly period. This phenomenon would certainly increase the investigating difficulty of predictive validation in this study if the input values of these specific variables were mostly zero. Under this inherent constraint, we tried to alleviate it by integrating similar errors in the classification framework. In HFACS-MA, we set up "Class P" factors as the 'parent' category of its 'subordinate' errors, "Class S" (see Fig. 1 and Table 1 for details). These integrated "Class P" factors would accumulate the amount of its subsidiary errors in "Class S" to increase the quantities of non-zero monthly frequencies. In brief, there are eight "Class P" factors, e.g. Error, Condition of Operators, Supervision Dysfunction and

Table 1
Details of "Class P" and "Class S" factors of HFACS-MA.

	"Class P"	"Class S"
Unsafe act	Error	Skill-based error
		Decision error
	Noncompliance	Routine
		Exceptional
Precondition for unsafe act	Condition of operators	Adverse state
		Limitation
		Teamwork
	Condition of task/environment	Task demand
		Physical environment
		Hardware/software
Unsafe supervision	Supervision dysfunction	Planning/organizing
		Leading/coordinating
		Controlling/correcting
	Supervision noncompliance	Routine
		Exceptional
Organizational influence	Organizational functionality	Operation procedure
		Execution
		Resource management
	Safety climate	Safety oversight
		Safety policy
		Safety culture

Organizational Functionality. The measurement reliability of "Class P" factors was examined in Part 1 (Hsiao et al., 2013).

Take the quantitative transformation of airline A as an example: there were 78 unsatisfactory audit reports for airline A in 2002; they were analyzed via HFACS-MA and then computed to the 12 monthly rates of eight "Class P" factors per year. The human error rates were weighted using the severity criteria developed by the regulatory authority in Equation (2). Every audit result was categorized by its responsible inspector to address its severity. From negligible to serious problems, the degrees of severity were classified as 'Information', 'Recommendation', 'Concern', and 'Finding'. We adopted the corresponding weights of the above degrees into the calculation of human error rates to emphasize the importance of severity (see Table 2). The weights are quoted from the aviation authority; they are derived via the Analytic Hierarchy Process (AHP) method, which has been generally applied in a variety of decision situations (Forman and Gass, 2001; Saaty, 2001). Twelve inspectors from the authority performed a paired comparison task for each pair of degrees of severity, with respect to their impact on safety. These weights have already been adopted in the daily analysis system of the aviation authority.

In Equation (2), the numbers of each human error category were accumulated according to the inspection date for each month, and multiplied by these weights to calculate the weighted human error rates. These quantitative rates were then treated as the indicators of human failures of aviation maintenance systems, and served as independent variables to predict the subsequent safety performance. Note: for concision purpose, we abbreviated the "weighted human error rate" to "error rate" from here on.

$$\text{Weighted Human Error Rate} = \frac{\sum_i (w_i * n_i)}{W * n} \quad (2)$$

W: The highest weight of severity degree ($W = 11$, the designated weight of Finding)
w_i: The weight of the severity degree, $i = \{I, R, C, F\}$
n: The sum of the human failures with all severity degree per month
n_i: The sum of the human failures with specific severity degree, $i = \{I, R, C, F\}$

In brief, there are ten independent variables of the prediction model, i.e. eight from the error rates of the eight "Class P" factors and two from the past and current incident rates. The next step is to establish a safety prediction model using both Neural Network and Poisson Regression methods and measure their predictive capability.

3. Method

3.1. Neural network

An Artificial Neural Network (NN) involves nonlinear functions to handle complex problems and is superior to other nonlinear techniques in its generality and practical ease of implementation (Azoff, 1994). Because the NN method is capable of coping with the challenges of non-normality, nonlinearity, and collinearity in forecasting area (Kappert and Omta, 1997; Jensen et al., 1999; Ranaweera et al., 2002; Wang et al., 2010), we utilized it in this study to avoid any non-normality problems.

In this study, we adopted the Back-Propagation Network (BPN) which is popular for performing difficult prediction problems (Werbos, 1994). Our model comprises one input layer, one hidden layer, and one output layer (see Fig. 2). Although more than one hidden layer might be beneficial for some applications, literature reviews showed that a single hidden layer is sufficient for a back-propagation network to solve most function approximation or input–output mapping problem for an arbitrary accuracy (Azoff, 1994; Fausett, 1994; Shyur et al., 1996; Ricotti and Zio, 1999; Nefti and Oussalah, 2004; Ung et al., 2006).

The input layer includes ten input neurons (the independent variables: eight error rates of the "Class P" factors, one past and one current incident rate), and the output layer possesses only one output neuron (the dependent variable: future incident rate) (Fig. 2). Information processing occurs at each neuron and is passed between neurons over connection links. Each link has an associated weight, which, in a typical NN, multiplies the transmitted signal values. A transfer function (or called activation function) would then be applied to the hidden neurons' net input (sum of weighted input values) to determine their output values.

With too few or too many hidden neurons, the network might either fail to model the underlying function or result in over-fitting (Kim and Park, 2001). Currently there is no specific rule prescribing how to decide the optimum number of hidden neurons. According to the experience gained by NN practitioners, it is suggested that the number of hidden neurons should vary as part of the investigation. Therefore, in this study, we used a trial and error process to determine the most appropriate number of hidden neurons based on balancing issues of simplicity and accuracy. On the other hand, the choice of appropriate transfer function is dependent on the numerical characteristics and ranges of the inputand output. Since all variables of this study have been standardized, a tangent sigmoid or log sigmoid function was appropriate considering the constraints.

We divided the variables (ten inputs and one output) into two groups: one training set and one testing set. Such a split-sample technique has been successfully adopted in safety related

Table 2
Weights of the severity degree of audit reports.

Severity	Information (I)	Recommendation (R)	Concern (C)	Finding (F)
Weight	1	5	8	11

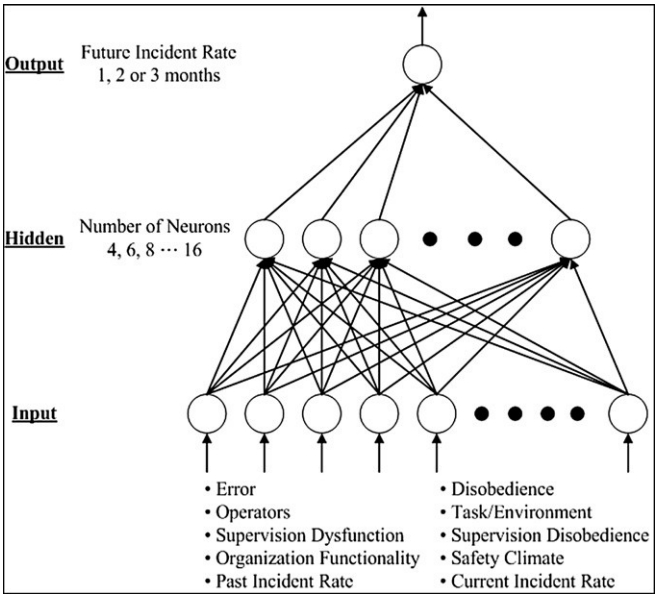

Fig. 2. The architecture of the neural network model.

research (Hashemi et al., 1995; Ung et al., 2006). In NN prediction studies, researchers would consider use more data set for training to assure its training performance and leave reasonable samples for test or validation (Zhang et al., 1999; Khosravi et al., 2011). Because the shortage of training data could seriously affect the learning performance of NN model, we decided to use reasonable sample size (i.e. two years, $n = 24$) for testing, and leave all other samples ($n = 53$) for training. The training set was presented in a random order to provide a stochastic weight update scheme.

3.2. Poisson regression

Because the response variable of our study, future incident rate, is discrete random ratio data and fits with Poisson distribution rather normality, we have chosen Poisson regression method as the generalized linear model (GLM) for modeling safety performance using the set of error rates. The same ten independent variables and one dependent variable as used NN model were added to the GLM. We aimed to explore the influential effects of the human error categories on the future incident rate and to compare the predictive capability of GLM with the performance of nonlinear NN method. The sample size of Poisson regression model has the same arrangement as the previous NN model: 53 to establishing the GLM model and 24 for predictive validation.

4. Analysis

4.1. Neural network results

The NN toolbox of Matlab® was used to establish the safety prediction model. After utilizing trial and error method to explore the best time period to obtain the highest performance of prediction, we found that the future incident rate with a one-month period had the best training and predictive performance compared to the two- and three-month ranges. The resultant correlation coefficients of both two- and three-month's incident rates were around 0.4, which are hardly sufficient to effectively support the prediction of the safety indicator.

Both airlines attained a strong level of overall correlation of the training sets ($r > 0.6$). The testing set was then used to predict the corresponding future incident rate. The predictive results of NN model were compared with the actual incident rate (target or desired output) to examine the predictive performance of the net. For the testing sets, the correlation coefficients both reached a significant level ($r_A = 0.61$, $r_B = 0.57$), which were also similar to the overall results of training set ($r_A = 0.61$, $r_B = 0.60$), giving R^2 values 0.38 for airline A, and 0.32 for airline B. The coefficient of determination, R^2, is usually used as a measure of the proportion of variability in a data set that is accounted for in prediction models. Therefore, R^2 value of 0.38 or 0.33 implies that about 30% of the variance is accounted for.

The tangent sigmoid transfer function used in the model had better and more significant predictive performance than the log sigmoid. However, it produced an unexpected artifact: negative incident rates, which were illogical. Therefore, we exponentiated the original prediction result transforming the values from negative to positive. Even though exponentiation would raise the concern of mathematical manipulation, we have presented the transformed results with the original ones as a comparison. The fitted line plots of the original and exponential prediction results are represented with 95% prediction interval for both airlines in Fig. 3a and b. Most of the prediction results were fitted within the prediction interval. We used three measures: correlation, standard error and R^2 to evaluate the predictive performance of the NN model and summarize in Table 3.

It should be noted that exponentiation transformation does not improve the predictive performance in general. Although the exponentiated result of airline A had better performance than the original prediction ($r_A = 0.77$, standard error = 1.09, $R^2 = 0.59$), the exponentiated result for airline B showed a decreased performance instead ($r_B = 0.50$, standard error = 1.05, $R^2 = 0.25$).

Any NN model requires the choice of several parameters, such as the transfer function, in what is essentially an empirical procedure with guidelines. Although there are several manipulations between the raw audit reports and the final prediction model, their numbers are not excessive as some values of the parameters have to be chosen. What we have done is to explore these parameters around the typical values, rather than in a random manner. We selected the parameters with the strongest predictive capability, but some others came close in their predictions, as shown in Table 3 for example.

4.2. Poisson regression results

For airline A, the R^2 of Poisson regression was 0.12 with $F(1, 23) = 3.03$, $p = 0.096$. Only the standardized coefficient, beta, of Error, Noncompliance and Condition of Operators were significant ($p < 0.05$) as 10.76, 16.08 and −13.40 respectively. The R^2 of airline B was negligible as 0.04 with $F(1, 23) = 0.97$, $p = 0.33 > 0.05$. No beta value was statistically significant in the regression analysis for airline B. Subsequently, the results of Poisson regression method were inadequate to establish a safety performance prediction model for our study. Although the intent of using Poisson regression method to predict future incident rate was unsuccessful, it provided a standpoint demonstrating the difficulty of using linear model to predict safety performance while comparing with the nonlinear NN model.

In conclusion, based on the comparison of the predictive performance of the NN and Poisson regression models, we not only confirmed that nonlinear methods such as NN could possess better predictive capability than generalized linear model, but also provided solid evidence of the causality between human error and safety performance from a social science and quantitative perspective. The prediction validation of the safety forecasting model founded on a HF/E classification framework (i.e. HFACS-MA) is confirmed in this study with acceptable performance criteria.

5. Discussion

The major contribution of this study is to develop a prediction model of future safety performance from past audit data from a HF/E standpoint. Generalization and validity of a safety prediction model is rare to find similar literature in field study (Raouf and Dhillon, 1993; Miyagi, 2005; Shyur, 2008). Current studies are mostly based on analysis or evaluation of the individual attributes and systematic conditions, such as policy, climate, attitude, and training factors (Ding and Ru, 2008), or procedures, culture and management maturity (OGP, 2009), or age and experience (McFadden, 2003). No real-time prediction validity has been quantitatively confirmed thorough utilization of human factors concepts and safety audit data. Therefore this study is a timely validation of proactive collection of data that purports a priori to be safety-related.

The results obtained here are encouraging as a first development and test of HFACS-MA model for deriving valid predictors from on-going audits. Following Cohen (1988) where the value of correlation coefficient between 0 and 0.09 as "none", 0.1–0.3 as "small", 0.3–0.5 as "moderate", and 0.5–1.0 as "large" correlation, for the two airlines, their safety performance were largely correlated with human failures ($r = 0.62$ and 0.57). This is not changed

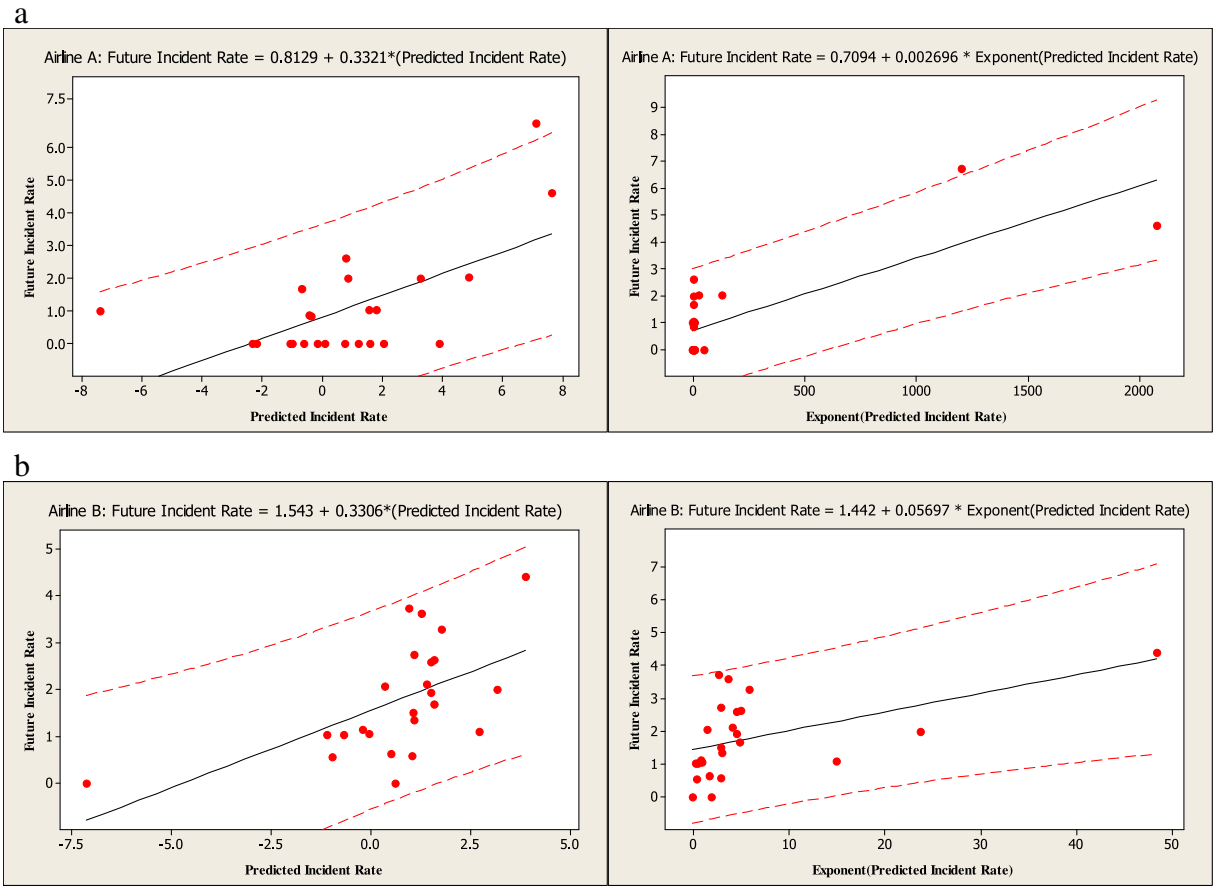

Fig. 3. (a). Fitted regression line with 95% prediction interval for airline A. The left figure illustrates the original predictive results, and the right figure shows the exponential ones. (b). Fitted regression line with 95% prediction interval for airline B. The left figure illustrates the original predictive results, and the right figure shows the exponential ones.

robustly after exponentiating the original output to avoid negative prediction artifacts ($r = 0.77$ and 0.50 for two airlines respectively). While any study would prefer to find very large correlations, the ones we did find ($\sim$0.6) are encouraging for an initial field study. We used existing independent data sets (audit and incident reports) rather than being able to develop our own data collection. Proactive data collection, for example collecting data based directly on HFACS or HFACS-MA rather than pre-existing audit checklist or questions, would be expected to help any predictions. However, it would be difficult to justify such a new study without the positive results of this study.

Before continuing, it is important to test alternative explanations for these significant results. In Part 1, we showed that merely counting audit findings did not predict future safety performance, leading to our development of a model based on HF/E concepts (HFACS-MA). We repeated the same analysis on our current data using both the unweighted and weighted audit records counts. The best results were both for the weighted data: for airline A the results were statistical significant ($p < 0.01$) though small with $R^2 = 0.172$. For airline B, both analysis was not significant ($p > 0.2$) and showed negligible prediction performance (best $R^2 = 0.06$). Compared with the data analyzed by HFACS-MA and NN model (airline A: $R^2 = 0.38$ and airline B: $R^2 = 0.32$) in this study, the improved prediction performance of our study is evident.

Table 3
Prediction performance of the neural network model.

		Correlation	Standard error	R^2	F (1, 23)	p value
Airline A	Original	0.62	1.33	0.38	13.69	0.001
	Exponential	0.77	1.09	0.59	32.20	0.000
Airline B	Original	0.57	0.99	0.32	10.61	0.004
	Exponential	0.50	1.05	0.25	7.34	0.013

However, future safety performance could also be a mere reflection of past performance. The prior and current monthly safety performance levels were included in our regression modeling, but did not result in significant contributions to prediction. In an attempt to determine whether there was any predictive validity in past safety performance, we also examined the whole data sequence for each airline. A standard method of time series analysis used autocorrelation to determine whether the nth order autoregressive models apply. While first-order autoregressive models have occasionally been used in HF/E (Drury and Corlett, 1975), a more advanced technique known as Autoregressive Integrated Moving Average (ARIMA) can be applied. This method uses moving averages of the data sequence rather than raw data after differencing to remove non-stationary in the data. Applying ARIMA to each airline's monthly safety performance data showed a first-order autoregressive model ($p = 0.01$ for Airline A, $p = 0.02$ for Airline B). The autoregressive equations were:

$$X_{i+1} = 0.29\,Xi + e$$
$$X_{i+1} = 0.27\,Xi + e$$

for airline A and B respectively, where X_i is the current data and ε is an error term. Clearly, with advanced techniques, it is possible to

extract some predictive validity from the time sequence of safety performance data, but the first-order autoregression coefficients are not particularly large (0.29, 0.27) and the predictive power was not large compared to HFACS-MA models. The conclusion from all of these alternative analyses is that HFACS-MA model provided the highest predictive validity. Merely using audit findings counts (either unweighted or weighted), or time sequence properties of the monthly safety performance data gives much less predictive validity.

This study demonstrated that the incident rate recorded each month could be predicted at a statistically significant, but not close to perfect level from the independently-collected audit results. Furthermore, that may not be the whole picture of airline safety. Current researchers (Janic, 2000; Ding and Ru, 2008; Rose, 2008) consider incident rate to represent only a partial view of overall safety, one that should be integrated with other concerns. Accident rate is another possible safety performance measure, although here the data were (thankfully) too few to use this measure in our analysis. Other enterprises may still have high accident rates and be able to use accident rate. Note that here the accident/incident ratios were similar between the two airlines despite the paucity of accident data, providing as least passing support to the traditional idea of the "accident pyramid" concept (Heinrich, 1950; Bird, 1975).

Similarly, the audit system used in this research is only one of many, even in aviation maintenance. Audit reports are already available in aviation maintenance, and in other fields (Drury, 2006). They are (or should be) based on sound error principles to find error precursors. Although some audit systems have been developed specifically for HF/E functions (Drury, 1990, 1999; Latorella and Prabhu, 2000) or organizational issues, e.g. the Air Transportation Oversight System (ATOS) of the FAA, most contemporary audit systems still lack suitable human error taxonomies to further investigate or analyze human failures. For aviation regulatory authorities, the contribution of this study is to eliminate the ambiguity between human factors and audit systems and emphasize the need for human error detection and diagnosis capability of audit systems. Our methodology can be used to test the effectiveness of current audit system, and improve their effectiveness. Inspectors could potentially focus their attention and resources on precursors of more serious problems and make appropriate alterations to their audit schedule. As a result, it can not only assist the assignment of limited inspection resources, but more importantly it can potentially contribute to the prevention of those human errors that are the leading factors in the lack of aviation safety.

In this study, because most of the independent variables did not fit the assumption of the normal distribution, we used Poisson regression as a comparison basis for the NN method. Generally, Poisson regression modeling proved insufficient capability to explain the causality between the error rates and the future incident rate. The NN model was expected to be better as it overcome some challenges existed in linear methods. Here it did indeed provide superior prediction performance to Poisson regression, similar to other results in the aviation field (Shyur et al., 1996; Miyagi, 2005; Shyur, 2006). We have carefully chosen the neural network parameters based on past practice and initial testing. The structure of our NN model is conventional (1 hidden layer etc.) and is consistent with other studies solving the function approximation problem (Azoff, 1994; Fausett, 1994; Shyur et al., 1996; Ricotti and Zio, 1999; Nefti and Oussalah, 2004; Ung et al., 2006).

This study used existing audit data as a field study in a manner more consistent with HF principles. It is not a laboratory or simulation research that we could try to control variables to focus on the only interested topic, e.g. Fitts' Law, and expect to represent high percentage of total variance. From a standpoint of field study, the performance of the NN method reached an acceptable level ($r = 0.62$ and 0.57), and the values of the performance criterion R^2 were in a reasonable range ($R^2 = 0.38$ and 0.32). Though this is not a great achievement yet, in social science, we do consider it as a fair start point of this innovative subject. We understand that the contradictory between negative incident rates and exponential transformation is a dilemma of this study and will be looking for ways to solve it in our future studies. To improve the prediction capability, ways such as using different NN model, having more safety data, e.g. near miss reports, in the analysis pool, or adopting other multivariate analysis techniques to adjust the amounts and values of the predictors will be taken into account.

On the other hand, although the result showed that future incident rate with a one-month period had the best training and predictive performance than the ranges of two- and three-month, the reasons behind this phenomenon are still unknown. Because of the novelty of this subject (Note: this is our first investigation that used historical audit data to predict future safety performance), there might be many unexplored limitations or unconsidered factors affecting the prediction range and performance. Further studies from other airlines and regulatory authorities are required to examine the prediction characteristics of prediction models in the aviation maintenance field. From a practical perspective, although a one-month predictive range might be lesser for safety management in long term, it could at least function as an on time warning to MROs, as a preceding caution to responsible managers to evoke situation awareness of specific latent failures and to adjust the resource allocation to correct targeted problems immediately.

For the airline industry, this study showed that a link could be made between human error and safety performance, and provided a potential link to practice. The study used data already collected, a reliable taxonomy, standard weighting factors and a NN methodology with no unusual features. In current practice of aviation maintenance, data is collected independently on audits and incidents, and also in some instances on errors (Taylor, 2000; Reason and Hobbs, 2003). Modern ideas on safety are now being implemented in aviation, for example the move toward safety management system (SMS) being advocated and indeed mandated by the International Civil Aviation Organization (ICAO) in the Annex 6. It is noted that currently the corresponding SMS regulations in North America (FAA and Transport Canada) and Europe (EASA) are still in the rulemaking process.

The idea behind an SMS is that data from various sources need to be integrated to provide a comprehensive view of the safety of the system, leading to actions based on more than mere response to the latest audit or incident (ICAO, 2009). Here we have shown that future safety can be predicted proactively, but only for two airlines, only for maintenance activities and only using existing data from a single regulatory authority. A negative result would have suggested a change in research direction, but the positive result implies that generality be deliberately extended to other authorities, airlines and data collection instruments. Some of this extension is quite straightforward, as much archival data exists, although not always easily available to researchers. Other aspect of extension may not be so obvious, for example testing the predictive power of specific aspect of the HFACS-MA system, or providing better audit systems that give sufficient details of the factors behind the "sharp end" of the person carrying out a task being audited. Part 1 of this study noted that better audit data was needed on supervisory and management factors. Such data can presumably be obtained by better audit instruments and training of auditors. Finally, extensions of the findings to domains beyond aviation maintenance, and indeed beyond aviation, are also appropriate as many industries use audits as the basis for their proactive safety activities.

Reason (1990) proposed that "accuracy of error prediction depends very largely on the extent to which the factors giving rise to the errors are understood", and this was exactly what this study has shown.

6. Conclusion

This is the first validation of audit reports against a widely accepted outcome measure of safety, because most human factors audits have either not been validated or validated against expert opinion. HFACS-MA categorization is the basis for the prediction, and is shown to produce superior prediction performance than either counting audit findings or relying on past safety performance. HFACS-MA model was tested for reliability in Part 1 of this study and, based on adequate reliability, was proven valid for predicting monthly safety performance for two airlines. This predictive ability was shown for existing data collected by a regulatory authority, demonstrating that even current data collection methods can be used. However, improved audit data collection could potentially benefit future research and the integration of disparate safety data into a Safety Management System.

The outcome of this study not only facilitates the application of human factors concepts in aviation maintenance systems, but also demonstrates the potential for audit validation in other safety fields, e.g. chemical and nuclear power industries, which also conduct internal and external safety audits.

References

Azoff, E.M., 1994. Neural Network Time Series Forecasting of Financial Markets. John Wiley & Sons, Ltd.

Bird, F.E., 1975. Control total de pérdidas. Consejo Interamericano de Seguridad, New Jersey.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. Routledge Academic.

Ding, S., Ru, Y., 2008. Flight Risk Assessment to Airlines Using Bayesian Belief Networks and Fuzzy Comprehensive Evaluation. Industrial Engineering and Engineering Management. IEEM 2008. IEEE International Conference on.

Drury, C.G., 1990. In: Lovesey, E.J. (Ed.), The Ergonomics Audit. Contemporary Ergonomics. Taylor & Francis, London, pp. 400–405.

Drury, C.G., 1999. In: Garland, D.J., Wise, J.A., Hopkin, V.D., Mahwah, N.J. (Eds.), Human Factors in Aviation Maintenance. Handbook of Aviation Human Factors. Lawrence Erlbaum Associates, Inc.

Drury, C.G., 2006. Human factors and ergonomics audits. In: Salvendy, G. (Ed.), Handbook of Human Factors and Ergonomics, third ed. John Wiley & Sons, pp. 1106–1132.

Drury, C.G., Corlett, E.N., 1975. Control of performance in multi-element repetitive tasks. Ergonomics 18, 279–298.

Drury, C.G., Dempsey, P.G., 2012. Human factors and ergonomics audits. In: Salvendy, G. (Ed.), Handbook of Human Factors and Ergonomics, fourth ed. John Wiley & Sons.

Fausett, L., 1994. Fundamentals of Neural Networks: Architectures, Algoriths, and Applications. Prentice-Hall, Inc.

Forman, E.H., Gass, S.I., 2001. The Analytic Hierarchy process – an exposition. Operations Research 49 (4), 469–486.

Hashemi, R.R., Le Blanc, L.A., et al., 1995. A neural network for transportation safety modeling. Expert Systems With Applications 9 (3), 247–256.

Heinrich, H.W., 1950. Industrial Accident Prevention A Scientific Approach. McGraw-Hill Book Company, Inc, New York.

Hsiao, Y.L., Drury, C., et al., 2013. Predictive models of safety based on audit findings: Part 1: model development and reliability. Applied Ergonomics 44 (2), 261–273.

ICAO, 2009. Safety Management Manual (SMM). ICAO.

Janic, M., 2000. An assessment of risk and safety in civil aviation. Journal of Air Transport Management 6, 43–50.

Jensen, J.R., Qiu, F., et al., 1999. Predictive modelling of coniferous forest age using statistical and artificial neural network approaches applied to remote sensor data. International Journal of Remote Sensing 20 (14), 2805–2822.

Kappert, C.G., Omta, S.W.F., 1997. Neural Networks and Business Modelling – An Application of Neural Modelling Techniques to Prospect Profiling in the Telecommunications Industry. System Sciences (HICSS), Annual Hawaii International Conference on: 465–473.

Khosravi, A., Nahavandi, S., et al., 2011. Lower upper bound estimation method for construction of neural network-based prediction intervals. IEEE Transactions on Neural Networks 22 (3), 337–346.

Kim, B., Park, S., 2001. An optimal neural network plasma model: a case study. Chemometrics and Intelligent Laboratory Systems 56, 39–50.

Koli, S., Drury, C.G., et al., 1995. Ergonomic audit for visual inspection of aircraft. Progress Report (DOT/FAA/AM-95/14). In: Human Factors in Aviation Maintenance–phase IV. Federal Aviation Administration/Office of Aviation Medicine, Washington, DC, pp. 67–83.

Latorella, K.A., Prabhu, P.V., 2000. A review of human error in aviation maintenance and inspection. International Journal of Industrial Ergonomics 26, 133–161.

Liou, J.J.H., Yen, L., et al., 2008. Building an effective safety management system for airlines. Journal of Air Transport Management 14, 20–26.

McFadden, K.L., 2003. Risk models for analyzing pilot-error at US airlines: a comparative safety study. Computers and Industrial Engineering 44 (4), 581–593.

Miyagi, M., 2005. Serious Accidents and Human Factors Breaking the Chain of Events Leading to an Accident: Lessons Learned from the Aviation Industry. American Institute of Aeronautics and Astronautics, Inc, Reston, Virginia.

Nefti, S., Oussalah, M., 2004. A neural network approach for railway safety prediction. 2004 IEEE International Conference on Systems, Man and Cybernetics 4, 3915–3920.

NTSB, 1988. Notification and Reporting of Aircraft Accidents or Incidents and Overdue Aircraft, and Preservation of Aircraft Wreckage, Mail, Cargo, and Records. 49 CFR Part 830.2.

OGP, 2009. Airline Safety Assessment Mechanism.

Ranaweera, D.K., Hubele, N.F., et al., 2002. Application Of Radial Basis Function Neural Network Model For Short-Term Load Forecasting. IEE proceedings. Generation, Transmission, and Distribution.

Raouf, A., Dhillon, B.S., 1993. Safety Assessment: A Quantitative Approach. CRC Press.

Reason, J., 1990. Human Error. Cambridge University Press, New York.

Reason, J., Hobbs, A., 2003. Managing Maintenance Error. Ashgate Publishing Company, Hampshire.

Ricotti, M.E., Zio, E., 1999. Neural network approach to sensitivity and uncertainty analysis. Reliability Engineering and System Safety 64, 59–71.

Robbins, S.P., Coulter, M., 2005. Management. Prentice Hall.

Robson, L.S., Macdonald, S., et al., 2010. Something might be missing from occupational health and safety audits: findings from a content validity analysis of five audit instruments. Journal of Occupational and Environmental Medicine 52, 536–543.

Rose, A. (2008). Understanding Aviation Risk. Information Fusion, 2008 11th International Conference on 1–7.

Saaty, T.L., 2001. Decision Making for Leaders: The Analytic Hierarchy Process for Decisions in a Complex World. Pennsylvania, RWS Publications, Pittsburgh.

Shappell, S., Wiegmann, D., 2003. A Human Error Approach to Aviation Accident Analysis. Ashgate Publishing Limited, Hants.

Shyur, H.J., 2006. The Study of Operation Andairworthiness Risk Assessment Decision Support System for Air Operator. Civil Aviation Administration, Taipei.

Shyur, H.J., 2008. A quantitative model for aviation safety risk assessment. Computers and Industrial Engineering 54 (1), 34–44.

Shyur, H.J., Luxhoj, J.T., et al., 1996. Using neural networks to predict component inspection requirements for aging aircraft. Computers and Industrial Engineering 30 (2), 257–267.

Taylor, J.C., 2000. The evolution and effectiveness of maintenance resource management (MRM). International Journal of Industrial Ergonomics 26, 201–215.

Ung, S.T., Williams, V., et al., 2006. Test case based risk predictions using artificial neural network. Journal of Safety Research 37, 245–260.

Wang, G., Wu, J., et al., 2010. Comparison between BP neural network and multiple linear regression method. Lecture Notes in Computer Science 6377, 365–370.

Werbos, P.J., 1994. The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting. John Wiley & Sons, Inc.

Zhang, G., Hu, M.Y., et al., 1999. Artificial neural networks in bankruptcy prediction: general framework and cross-validation analysis. European Journal of Operational Research 16, 16–32.