Helicopter Simulator Performance Prediction Using the Random Forest Method

Hans Bauer; Dennis Nowak; Britta Herbig

INTRODUCTION: Different aspects of the aviation system, such as pilot's fitness, supervision, and working conditions, interact to produce or protect against flight safety hazards. Machine learning methods such as Random Forests may help identify system characteristics with the potential to affect flight safety from the large number of candidate predictors that results when multiple system levels are considered simultaneously.

METHODS: There were 54 pilot-related and occupational candidate predictors of simulator flight performance in 2 malfunction scenarios completed by 51 male European helicopter emergency medical services pilots derived from pilots' self-report questionnaires and aeromedical examination records. In a cross-sectional explorative analysis, the Random Forest method was used to screen for informative predictors. Predictors scoring above the critical threshold for the conditional permutation variable importance (VI) statistic were selected.

RESULTS: In five predictors, the VI statistic averaged across 2000 Random Forest runs exceeded the selection threshold: higher perceived rewards (VI = 0.0691) and predictability (VI = 0.0501) at work were associated with higher performance scores, and higher physiological dysregulation (VI = 0.0495) and alanine aminotransferase (VI = 0.0224) with lower scores. Performance also differed between the simulators at the two training sites (VI = 0.0298).

DISCUSSION: Random Forests may usefully complement previously applied methods for the identification of human factors safety hazards. The identified performance predictors suggest further areas with potential for safety improvements.

KEYWORDS: flight safety, machine learning, helicopter pilots, human factors, flight simulator.

Bauer H, Nowak D, Herbig B. *Helicopter simulator performance prediction using the Random Forest method*. Aerosp Med Hum Perform. 2018; 89(11):967–975.

Improving aviation safety has become an increasingly challenging task since most easily recognizable hazards, such as technological deficiencies, have been reduced significantly over the last decades and the remaining ones are often of a latent or insidious nature. Current human factors models of accident causation such as the popular "Swiss cheese" model[16] therefore take a systemic perspective acknowledging hazards related to both the aircraft operators themselves and their organizational environment.

Nevertheless, in empirical studies of aircraft accident causes based on such a systemic perspective, there is usually a notable gradient in the frequency of identified contributory factors across the system levels, with unsafe operator acts being identified most often and organizational influences least often.[17] It seems likely that the absence of organizational, supervisory, and working condition-related factors is at least in part due to the reliance on accident investigation board reports, which vary in scope and may investigate more distal conditions surrounding an accident only in the most severe cases.[27] Further complicating the issue, information on the relevant comparison group (i.e., pilots who did not experience an accident or incident) is usually not available or only in aggregated form in report-based retrospective studies.

These issues can be addressed through prospective studies, which, however, introduce different methodological problems. Aviation safety-related prospective studies can only use proxy outcomes such as line check ratings or simulator performance, and are often based on small samples, which is problematic

From the Institute and Clinic for Occupational, Social, and Environmental Medicine, University Hospital of LMU Munich, Munich, Bavaria, Germany.

This manuscript was received for review in February 2018. It was accepted for publication in July 2018.

Address correspondence to: Hans Bauer, M.Sc., Institut und Poliklinik für Arbeits-, Sozial- und Umweltmedizin, Ziemssenstraße 1, 80336 München, Germany; hans.bauer@med.uni-muenchen.de.

DOI: https://doi.org/10.3357/AMHP.5086.2018

given the large number of potentially interacting hazardous and protective factors operating at different system levels. Isolating the independent contribution of individual system characteristics to risk is difficult when samples are small relative to the number of characteristics; conventional techniques such as univariate prescreening of candidate predictors or multiple regression-based stepwise variable selection are problematic since they do not consider the effects of discarded predictors and may result in upwardly biased effect estimates.[7] They furthermore generally assume linear additive relationships between predictors and outcome.

Several alternative procedures have been proposed to solve the problem of selecting influential predictors.[28] One approach which has recently gained popularity is based on Random Forests, a machine learning technique which is appropriate for situations when the number of predictors is large relative to the sample size, avoids overfitting, and automatically takes into account interactions between predictors, while also generating a measure of predictor importance which can be used for variable selection.[23] It is, therefore, well-suited for application to the aviation hazard identification problem outlined above and its capability for predictor identification has, to the best of our knowledge, so far not yet been used in the aviation safety field.

In this cross-sectional explorative study, we therefore apply the Random Forest method in order to identify potential aviation safety hazards by selecting the most powerful predictors of simulator flight performance in a sample of professional helicopter pilots from a deliberately broad set of predictors covering both personal and occupational human factors safety aspects. The predictors are derived from pilot questionnaire self-reports and aeromedical fitness examination records and performance is assessed through flight instructor ratings of the pilots' handling of two naturalistic system malfunction scenarios. Our aim was to evaluate, in an aviation safety context, whether the Random Forest method would produce meaningfully interpretable results when the number of predictors was large relative to the sample size.

METHODS

Subjects

The analysis reported herein is based on data from a study on age and flight safety in helicopter emergency medical services (HEMS). Active HEMS pilots employed by one of five air rescue operators (two based in Germany and one each in Austria, Poland, and the Czech Republic) who completed a simulator flight during the data collection period from September 2015 to October 2016 at training sites in Warsaw, Poland (Polish operator), or in Hangelar, Germany (all other operators), were asked to participate. Although study participation was open to both genders, the operators' HEMS pilot workforces consisted almost exclusively of men at the time of recruitment, resulting in a male-only sample. The study was approved by the Ethics Committee at Munich University's Faculty of Medicine (Project No. 466-15) and written informed consent was obtained from all study subjects prior to data collection. Subjects were able to separately indicate their consent to the collection of simulator performance data and of aeromedical examination record data.

Materials and Procedure

We collected data from three distinct sources: 1) standardized ratings of simulator flight performance in two malfunction scenarios made by training instructors; 2) self-report questionnaires from participating pilots; and 3) records of the participating pilots' statutory aeromedical examinations. Subjects completed a simulator flight for training or testing purposes as mandated by their employer or the responsible regulatory authority in order to maintain currency of their helicopter type rating or of operating procedures, including emergency procedures. Each session consisted of a preflight briefing of the pilot by the flight instructor, the actual simulator flight, and a short postflight debriefing. At Hangelar, the flights were conducted in motion-capable full flight simulators. A non-motion-capable flight training device was used in Warsaw. At both sites, the simulators corresponded to the helicopter types flown by the pilots during their actual duty and could simulate different geographies, including urban structures and weather conditions.

Embedded in this regular training/check flight, which was independent of the study, the responsible flight instructor—who had previously been briefed by a member of the research team—deployed two study malfunction scenarios, which together took about 10 min to complete, and rated the pilot's responses according to a standardized rating sheet. Several experienced HEMS pilots had been consulted prior to data collection in order to select relevant scenarios and to develop the corresponding rating scheme.

The first scenario, "transmission oil system malfunction," was designed to assess the pilot's vigilance and situational awareness. It involved the timely detection of a gauge indicating helicopter transmission oil status moving slowly toward a critical value and terminated as soon as the pilot detected the problem, or the critical value was reached. More specifically, the malfunction concerned a gradual decrease in transmission oil pressure in the Hangelar simulators, and a gradual increase in transmission oil temperature in the Warsaw simulator, since the oil pressure decrease scenario could not be implemented in the latter simulator. The respective gauges were located next to each other at similar positions within the flight instrument panel in all simulators, and the corresponding malfunctions are comparable in terms of their implications for flight safety. We therefore assumed the tasks to have similar properties in terms of their demands to situational awareness. However, the time from initiation of malfunction to hitting the critical value, which was defined as the first occurrence of an additional optical or acoustic warning signal by the system, was approximately twice as long in the Hangelar simulator compared to the Warsaw simulator (170 vs. 84 s). Recognition of the malfunction by the pilot ahead of the critical value was rewarded with two points (same weight as an "important" subtask of the second scenario described below).

The "tail rotor drive failure" scenario constituted a complex emergency situation which required the pilot to bring the aircraft safely to the ground via a so-called autorotation procedure. Instructors rated performance of 9 subtasks which involved situational awareness, decision-making, knowledge of procedures, spatial orientation, and psychomotor control, for a maximum of 12 points. Proper completion of subtasks yielded one or two points, based on their importance to the successful completion of the entire procedure. There were 15 simulator flights concurrently rated by two instructors, with good interrater agreement over a total of 141 binary decisions (subtask ratings); Cohen's $\kappa = 0.91$ (94.3% concordant decisions).

In the briefing parts of the session, pilots were asked to fill in short questionnaires inquiring about flight experience and current self-rated health (preflight briefing), and about subjectively experienced strain and simulator sickness during the flight as well as risk-seeking propensity (postflight debriefing). Additionally, the pilots were handed a longer questionnaire covering working conditions, subjective experiences at work, and general sense of well-being, which they were asked to fill in and send back over the course of the next few days. Wherever possible, questionnaire scales and items had been taken from published, psychometrically validated instruments such as the Copenhagen Psychosocial Questionnaire. They were included based on their relevance regarding health, work performance, and safety.[14,20]

Consenting pilots were also asked to provide details of the aeromedical examiners and centers they had visited during the preceding 10 yr and release them from their nondisclosure duty. We mailed requests for transfer of full aeromedical examination documents to all specified physicians/centers and sent a reminder letter after 4 wk. If the reminder also did not elicit a response, we made at least one more contact attempt via telephone or email. We received digitalized or paper documents dated between April 2004 and July 2016 from 23 physicians/centers (61% of 38 contacted). All documents were searched for information concerning quantitative clinical measurements of any kind, as well as smoking and medication. Entry of the corresponding data was conducted according to a coding manual developed after an initial review of the received documents and continuously revised during the data entry process.

Statistical Analysis

Our main outcome variable, simulator performance, was calculated as the sum of instructor ratings for the two study scenarios (ranging theoretically between a minimum of 0 and a maximum of 14 attainable points). The predictor variables were based on the self-report questionnaires and aeromedical examination findings. Questionnaire-based variables were calculated as scale means or sum scores, or (in the case of single-item measures) the untransformed item scores. We selected individual medical risk markers for analysis based on their availability in the aeromedical examination findings data and on their association with health status and incapacitation risk of a pilot. Moreover, we included as predictors two composite indices based on the individual risk markers: risk of a fatal cardiovascular event within 6 mo according to the SCORE algorithm,[5] and a "physiological dysregulation index" based on the gerontological concept of "biological aging,"[4] which can be viewed as a subclinical trajectory toward frailty and disease due to insidious functional decline in multiple organ systems.[13] Both cardiovascular event risk scores such as SCORE and physiological dysregulation indices have been found to be associated with cognitive decline.[2,9]

Our dysregulation index is based on all available measurements of 18 health risk-related biomarkers of cardiovascular, metabolic, liver, kidney, immune, hematologic, and ocular function as well as hearing level. For each biomarker, a subject's probability to have a biomarker reading within an "unhealthy" range (e.g., systolic blood pressure >140 mmHg) was estimated based on the longitudinal individual-specific distribution of biomarker readings using a linear mixed model. The dysregulation score is the sum of these probabilities (thus ranging theoretically between 0 and 18) and can be understood as a pilot's expected number of biomarker readings outside the healthy range (see **Appendix A** online, https://doi.org/10.3357/AMHP.5086sd.2018, for methodological details).

Finally, we also included simulator type as a predictor because of the aforementioned differences in the transmission oil failure scenario and in motion capability. Only variables where less than one-third of subjects had missing data were used for the analysis. In total, this resulted in a set of 54 predictors which can be categorized as psychosocial and physical work stressors (including protective factors such as social support), psychosocial and physical strain symptoms, other aspects of working conditions (e.g., working hours), medical risk markers, subjective experience of the pilot during the simulator training session, and general pilot characteristics (**Table I**; see **Appendix B** online, https://doi.org/10.3357/AMHP.5086sd.2018, for further details). For all medical risk markers except physiological dysregulation (which was based on the complete longitudinal information available), we used only the latest available assessment in the simulator performance prediction.

Table I also shows the number of missing data points per predictor. To impute missing values, we used the R implementation of the missForest algorithm,[21] which iteratively applies the Random Forest method (described below). In order to account for uncertainty in the imputation estimate, we created a total of 20 imputed datasets and compared or aggregated analysis results across these datasets where appropriate.

After the imputation step, we applied Random Forests for simulator performance prediction and variable selection. Random Forest is a supervised machine learning method consisting of an ensemble of decision trees which attempt to predict the outcome by defining, based on predictor values, groups that are homogeneous with respect to the outcome. In Random Forests, many such trees are "grown" on random subsamples of the study subjects and predictor variables, and the single trees' predictions are averaged to produce the Forest's prediction. Since each component tree is trained only on a random subsample, the remaining ("out-of-bag") cases can be conveniently used as

Table I. Analysis, Variable Sources, and Descriptives (Mean/SD for Quantitative Variables, *N*/% for Binary Yes/No Variables in Italics).

VARIABLE CATEGORY AND VARIABLE	# OF ITEMS/SCALE RANGE	SOURCE	# MISSING	MEAN (SD)/ *N (%)*	ASSESSMENT-SIMULATOR LAG[*]
Outcome					
Simulator performance	10/0–14	Rating by flight instructor	0	11.4 (2.1)	--
Psychosocial work stressors					
Emotional demands	2/1–5	Pilot self-report (questionnaire)	0	2.3 (0.6)	--
Social support	4/1–5	Pilot self-report (questionnaire)	0	3.4 (0.9)	--
Work pace	2/1–5	Pilot self-report (questionnaire)	0	3.3 (0.7)	--
Work predictability	3/1–5	Pilot self-report (questionnaire)	0	3.0 (0.8)	--
Role clarity	3/1–5	Pilot self-report (questionnaire)	0	4.7 (0.4)	--
Role conflict	4/1–5	Pilot self-report (questionnaire)	0	1.8 (0.6)	--
Autonomy	3/1–5	Pilot self-report (questionnaire)	0	3.1 (0.7)	--
Supervisor Feedback	3/1–5	Pilot self-report (questionnaire)	0	3.0 (1.1)	--
Procedural Justice	4/1–5	Pilot self-report (questionnaire)	0	3.9 (0.8)	--
Effort at work	2/2–8	Pilot self-report (questionnaire)	0	4.7 (1.6)	--
Reward at work	5/5–20	Pilot self-report (questionnaire)	0	13.9 (2.2)	--
Physical work stressors					
Physical demands, general	10/1–5	Pilot self-report (questionnaire)	0	3.2 (0.6)	--
Physical demands, headgear	4/1–5	Pilot self-report (questionnaire)	0	2.5 (0.9)	--
Psychosocial strain					
Irritation	6/1–7	Pilot self-report (questionnaire)	0	2.2 (0.7)	--
Work engagement	9/0–6	Pilot self-report (questionnaire)	0	4.8 (0.8)	--
Detachment from work	4/1–5	Pilot self-report (questionnaire)	0	3.2 (0.8)	--
Subjective well-being	5/0–25	Pilot self-report (questionnaire)	0	19.0 (3.5)	--
Energy / Fatigue	4/4–20	Pilot self-report (questionnaire)	0	17.2 (2.0)	--
Physical strain					
# of body regions with pain[†]	13/0–13	Pilot self-report (questionnaire)	2	1.0 (1.7)	--
Other work-related factors					
Work hours per month	--	Pilot self-report (questionnaire)	3	169 (27)	--
Vacation days per year	--	Pilot self-report (questionnaire)	1	24.8 (7.7)	--
Day shift duty?	--	*Pilot self-report (questionnaire)*	*2*	*48 (0.98)*	--
Night shift duty?	--	*Pilot self-report (questionnaire)*	*2*	*32 (0.65)*	--
24 h stand-by shift duty?	--	*Pilot self-report (questionnaire)*	*4*	*5 (0.11)*	--
Other shift type duty?	--	*Pilot self-report (questionnaire)*	*3*	*3 (0.06)*	--
Any limit on flight time?	--	*Pilot self-report (questionnaire)*	*9*	*37 (0.88)*	--
Medical risk markers[‡]					
Smoking?	--	*Aeromedical records*	*1*	*10 (0.20)*	*94*
Any medication?	--	*Aeromedical records*	*1*	*9 (0.18)*	*89*
Systolic blood pressure (mmHg)	--	Aeromedical records	0	130 (13)	92
Resting heart rate (bpm)	--	Aeromedical records	0	67.8 (8.7)	92
ECG QTc interval (ms)	--	Aeromedical records	15	406 (21)	91
Body mass index ($kg \cdot m^{-2}$)	--	Aeromedical records	1	27.1 (3.3)	89
Total cholesterol ($mmol \cdot L^{-1}$)	--	Aeromedical records	3	5.3 (0.8)	102
HDL cholesterol ($mmol \cdot L^{-1}$)	--	Aeromedical records	8	1.5 (0.3)	349
Triglycerides ($mmol \cdot L^{-1}$)	--	Aeromedical records	11	1.5 (0.7)	302
Fasting glucose ($mmol \cdot L^{-1}$)	--	Aeromedical records	3	5.3 (0.6)	127
Alanine aminotransferase ($U \cdot L^{-1}$)	--	Aeromedical records	14	37.1 (22.7)	712
Aspartate aminotransferase ($U \cdot L^{-1}$)	--	Aeromedical records	14	28.2 (9.9)	712
Serum creatinine ($\mu mol \cdot L^{-1}$)	--	Aeromedical records	8	91.6 (14.3)	886
White blood cell count ($10^3 \cdot \mu l^{-1}$)	--	Aeromedical records	2	6.5 (1.7)	101
Hemoglobin ($g \cdot dl^{-1}$)	--	Aeromedical records	2	15.3 (1.0)	101
Red blood cell distribution width (%)	--	Aeromedical records	4	13.0 (0.6)	101
Intraocular pressure (mmHg)	--	Aeromedical records	9	15.2 (2.4)	213
Hearing level at 3000 Hz (dB HL)	--	Aeromedical records	4	16.6 (9.6)	268
SCORE 6-mo risk (%)	--	Aeromedical records	3	0.07 (0.07)	102
Physiological dysregulation[§]	--	Aeromedical records	[¶]	1.4 (0.9)	--
Experience during simulator training session					
Self-rated health	1/1–5	Pilot self-report (questionnaire)	0	3.6 (0.9)	--
Task load during flight	5/0–100	Pilot self-report (questionnaire)	0	49.0 (16.0)	--
Simulator sickness	1/0–100	Pilot self-report (questionnaire)	0	14.4 (22.6)	--

Continued

Table I, *Continued.*

VARIABLE CATEGORY AND VARIABLE	# OF ITEMS/SCALE RANGE	SOURCE	# MISSING	MEAN (SD)/ *N (%)*	ASSESSMENT-SIMULATOR LAG*
General pilot characteristics					
Age (yr)	--	--	0	51.7 (8.2)	--
# of real flight hours	--	Pilot self-report (questionnaire)	0	5340 (3513)	--
# of simulator flight hours	--	Pilot self-report (questionnaire)	1	187 (411)	--
Risk seeking	4/1–5	Pilot self-report (questionnaire)	0	1.4 (0.4)	--
Other					
*Full flight simulator?***	--	--	*0*	*23 (0.45)*	--

* Median time from risk marker variable assessment to simulator session in days. †Number of body regions (e.g., neck, lower back; 13 overall) where subject reported "occasional" or "frequent" pain. ‡As assessed at last available aeromedical examination (except physiological dysregulation). §Index derived from 18 health-risk associated biomarkers (see Auxiliary Appendix A). ¶Individual index component variables had been imputed before computation of index. Overall number of component variable missing values: 139 (= 15.1% of 18*51 data points). **No = flight training device.

an "external" validation set for the prediction quality of this particular tree. Prediction errors for the out-of-bag cases can again be averaged across all trees to yield a measure of prediction accuracy which is less affected by overfitting. Furthermore, and particularly important in the present context, Random Forests are also able to produce a measure of relative importance in the prediction of the outcome for each individual predictor variable even when the number of predictors is larger than the sample size (which is not possible with conventional regression modeling techniques). In this way, relevant predictors can be identified. The importance of a predictor is calculated as the difference in out-of-bag prediction accuracy between a Random Forest grown on the original input data and that of another forest which is identical in parameterization and input data except that the values of the predictor have been randomly permuted, reducing its predictive capacity to chance level.[23] We used a modification of this variable importance measure (termed "conditional permutation importance") which accounts for intercorrelations among predictors (analogous to the mutual adjustment of covariate effects in a multiple regression).[22]

For each of the 20 imputed datasets, we fitted 100 Random Forests and calculated the mean variable importance across the resultant 2000 Forests for each predictor variable to obtain an estimate which is less affected by the random variation inherent in the Random Forest procedure. We then selected those variables for further inspection whose mean variable importance exceeded the absolute value of the minimum mean variable importance among all variables. This selection criterion was suggested as an improvement over the z-score metric commonly used for Random Forest variable selection.[23] Note that conventionally reported statistics such as *P*-values or confidence intervals are not directly applicable to Random Forests, although the described selection criterion can be considered an analog to a *P*-value-based statistical significance threshold such as $P < 0.05$.

We also examined the stability of the importance assigned to a variable by inspecting how strongly the variable's position in the importance ranking varied between the 2000 Random Forest fits. To visualize the relationship between the selected predictors and the outcome, we present partial dependence plots[10] which display the effect of a predictor across its value range averaged over all other predictor value combinations occurring in the sample (i.e., the estimated marginal effect of the predictor). All statistical analyses reported subsequently were done using the R statistical software, version 3.3.1.

RESULTS

Fig. 1 shows the study subject flow. Unavailability of simulator data was mostly due to a mismatch of pilots' training session schedules with the data collection period, whereas nonresponse by physicians/centers from whom examination records had been requested was the main reason for unavailability of aeromedical data. The analysis sample consisted of 51 male pilots with all data types available, 15 (29%) from the Western European countries (Germany, Austria) and 36 (71%) from the Eastern European countries (Poland, Czech Republic); see Table I for descriptives. Simulator performance scores were concentrated at the upper end of the scale and slightly left-skewed (range: 7–14, median: 12, skewness: −0.39). For some of the medical predictors, there was a considerable time lag between their assessment and the simulator training session.

Mean variable importances of the 54 predictor variables ranged from −0.0196 to 0.0691. Five variables had a higher mean importance score than the selection threshold of 0.0196 (see Statistical Analysis above): the reward subscale of the

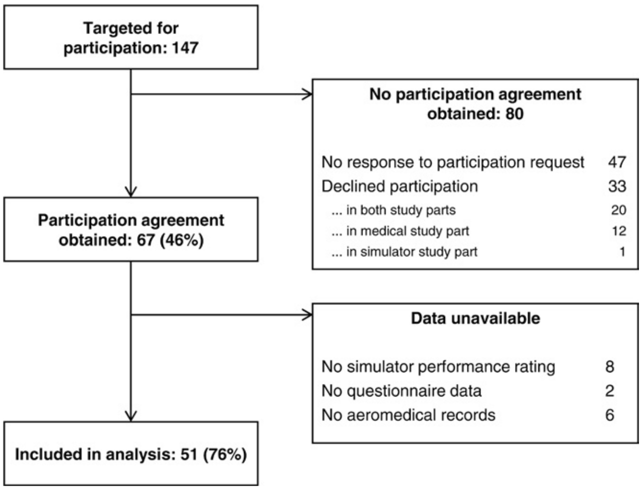
Targeted for participation: 147
No participation agreement obtained: 80
No response to participation request 47
Declined participation 33
... in both study parts 20
... in medical study part 12
... in simulator study part 1
Participation agreement obtained: 67 (46%)
Data unavailable
No simulator performance rating 8
No questionnaire data 2
No aeromedical records 6
Included in analysis: 51 (76%)

Fig. 1. Study subject flow from recruitment to analysis stage.

Effort-Reward-Imbalance inventory[18] measuring perceived rewards at work, the predictability of work demands score,[26] the physiological dysregulation score, simulator type, and the last available measurement of alanine amino-transferase, a marker of liver pathology (**Table II**; see **Appendix C** online, https://doi.org/10.3357/AMHP.5086sd.2018, for details on all predictors). There is a notable drop in mean variable importance from the selected to the unselected variables. Perceived rewards clearly stands out as having the highest predictive power, followed by perceived predictability and dysregulation and, after another distinct drop in variable importance, by simulator type and alanine aminotransferase. The importance rankings were fairly stable across Random Forest fits for the reward, predictability, and physiological dysregulation variables. Rankings were more unstable for simulator type and in particular for alanine aminotransferase.

The estimated marginal effects of the selected variables are illustrated in **Fig. 2**. Even in those variables which have the strongest association to simulator performance in the sample, the effects on performance are all quite small and on the order of 0.1 to 0.2 performance score points (where one performance score point roughly corresponds to one mistake in the simulator scenarios). In other words, simulator performance is, on the whole, poorly predicted by the set of 54 variables considered. Still, the direction of effects is generally plausible in the selected variables: performance was better in those who perceived their work as more rewarding and work demands as more predictable, and worse in those with higher physiological dysregulation scores and alanine aminotransferase levels. Several possible explanations come to mind (e.g., differences in simulator handling characteristics, in the transmission oil malfunction scenario, or in the response tendencies of the involved instructors) regarding the effect of simulator type. However, since this variable was included for technical reasons (i.e., adjustment for simulator idiosyncrasies) rather than theoretical reasons, we will not discuss it any further.

The predicted effects in the quantitative variables were not linear but rather stepwise. Note that for the working conditions variables, the threshold values ($\sim$12 for the Reward Subscale and $\sim$3 for the Predictability scale) correspond to the theoretical scale means; that is, performance was predicted to be better in those who, on the whole, agreed to statements such as "I receive the respect I deserve from my superiors," and/or indicated that all in all, daily job demands could be predicted to a large extent. The threshold effect in the physiological dysregulation score can be interpreted such that pilots who are in an excellent overall state of health according to the biomarkers used for the dysregulation score (i.e., all biomarker levels are well within the healthy range) are predicted to perform better than those who tend to have levels near or beyond the limits of the healthy range in at least one of the biomarkers. Although the alanine aminotransferase effect is harder to interpret, it is noteworthy that the threshold value after which performance decreases ($\sim$40 U $\cdot$ L^{-1}) falls into the upper end of male population reference range limits (e.g., Piton et al.[15]).

DISCUSSION

In this cross-sectional explorative study, we applied the Random Forest machine learning method for the selection of the most influential predictors of HEMS pilot performance in two simulated in-flight failure scenarios from a set of predictors which covered personal and occupational human factors aspects potentially relevant to flight safety and which was larger than the sample size. Although the predictors on the whole explained rather little of the variation in simulator flight performance, five of them (perceived rewards at work, perceived predictability of work demands, physiological dysregulation, alanine aminotransferase, and simulator type) explained more than would be expected by chance alone. Their effects appeared to be stepwise rather than linear and their direction was mostly consistent with theoretical expectations.

To the best of our knowledge, this is the first application of the Random Forest method to identify potential human factors safety issues. Many analyses use conventional bivariate or multiple regression methods.[3,25] These "classical" methods have good statistical properties when their assumptions are met (which notably includes the rather restrictive assumption of linear effects) and are well-suited for confirmatory analyses

Table II. Summary of Variable Importance Characteristics of Variables Selected by Random Forest Procedure (in Italics) and of First Three Variables Below Selection Criterion.

		VARIABLE IMPORTANCE RANK ORDER*			
	MEAN VARIABLE IMPORTANCE	M	SD	LOWEST	HIGHEST
Reward subscale (ERI)	0.0691	1.42	0.72	6	1
Predictability of work demands scale	0.0501	2.55	1.17	16	1
Physiological dysregulation score	0.0495	2.61	1.18	14	1
Simulator type	0.0298	4.49	2.24	44	1
Last available ALT measurement	0.0224	6.40	5.19	54	1
Last available serum creatinine measurement	0.0113	9.66	6.03	47	2
Work hours per month	0.0099	10.66	6.79	48	2
Simulator sickness	0.0097	10.52	6.71	53	3

Statistics calculated across 2000 Random Forest fits (100 in each of 20 imputed datasets). Selection criterion: mean variable importance greater than absolute value of minimum mean variable importance (−0.0196).

ERI: Effort-Reward-Imbalance Scale. ALT: alanine aminotransferase.

* Ranks (including mean ranks) coded such that lower values denote higher ranks (highest: 1st rank; lowest: 54th rank).

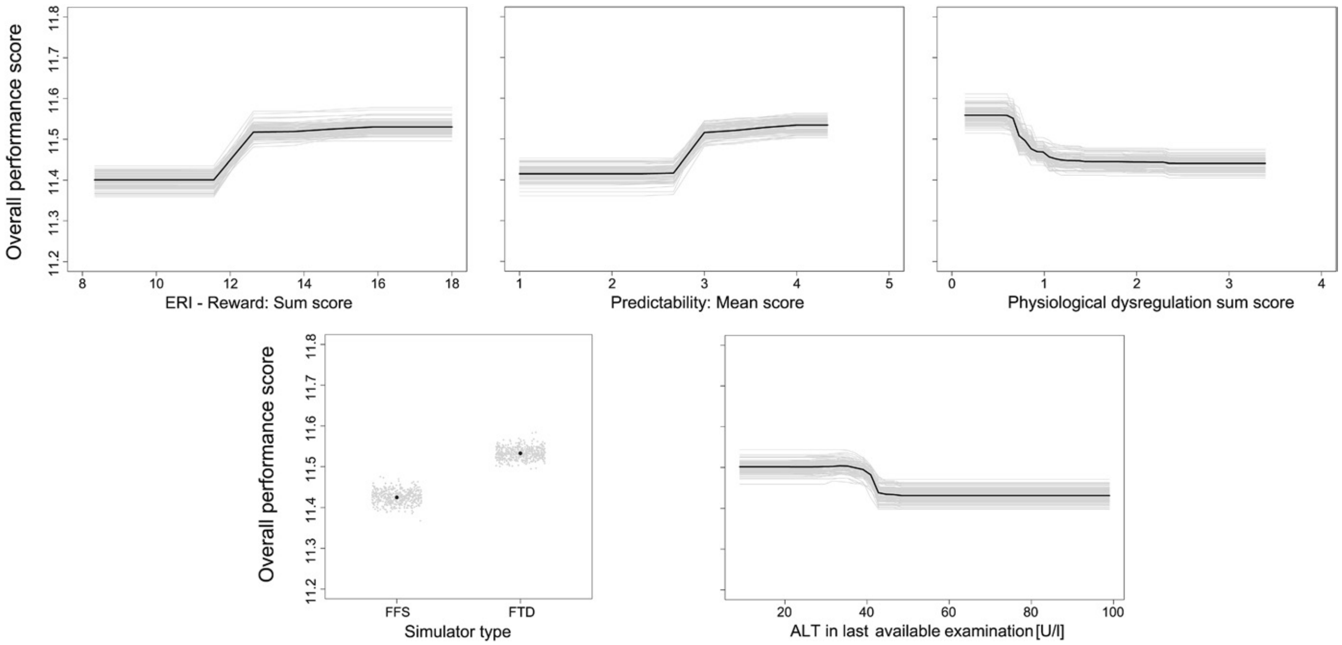

Fig. 2. Partial dependence plots of flight simulator performance vs. predictors selected by Random Forest procedure (predictor value ranges as present in the sample). Black line/dots: Prediction averaged across 2000 Random Forest fits. Gray lines/dots: Prediction of randomly selected individual fits to illustrate variability across fits. Note that the y-axes display only a small fraction of the outcome variable's theoretical range (0–14). ERI-Reward: Effort-Reward-Imbalance Reward Subscale. FFS: Full flight simulator. FTD: Flight training device. ALT: Alanine aminotransferase.

investigating the effects of a smaller number of predictors of interest. In contrast, machine learning methods appear to be better suited for explorative analyses where vast amounts of information on potential predictors is available and little is known about the functional relationship between the predictors and the outcome.

Analysis of natural-language documents are a prototypical example of a high-dimensional input problem; in aviation safety, machine learning methods are becoming increasingly popular for text mining of accident/incident report narratives. Often, unsupervised learning methods are applied to cluster occurrence reports and subsequently identify common underlying themes, such as cigarette smoking by passengers.[24] While these approaches are very flexible and can accommodate data which is otherwise difficult to process, they are able to identify only those factors mentioned in the reports, which tend to be proximal factors.[27] Furthermore, human factors hazards such as "confusion" are often hard to isolate for text mining algorithms since their description mostly lacks highly distinctive signaling words which characterize the more technical issues.[24] Our analysis may be located somewhere in between the two extremes of linear modeling of selected features and indiscriminate text mining, in that it allows a certain preselection of features, including those which are usually not considered in occurrence reports, such as organizational stressors, but does not impose restrictive assumptions on the relationship between predictors and outcome.

Of the five selected informative predictors, two measured aspects of psychosocial work stress. This kind of stress is known to affect safety at work.[14] Young[29] reviewed the effects of life stress (including work stress) on pilot performance and suggested that life stress might undermine performance by increasing fatigue (through reduced sleep quantity and quality as well as emotional exhaustion), decreasing motivation to perform (e.g., skipping "unimportant" tasks such as checklist procedures), worsening interpersonal relationships and communication with colleagues, and increasing intrusive and distractive worrying. With regards to the most predictive of our variables, perceived rewards at work,[18] reduced motivation may be a plausible contributing factor.[11] On the other hand, the second selected work stressor[26] assesses the degree to which the pilot perceives his work environment to be predictable. This may be related to instances of disrupted action regulation ("hindrance stressors") during work; repeated experience of disrupted action regulation might lead to a more passive style of coping with work demands.[12]

The concept of physiological dysregulation as an overall loss of an organism's capacity to maintain homeostasis has recently received increased interest in gerontology to explain differences in the "healthiness" of aging between individuals.[2,13] Physiological dysregulation was found to be associated with cognitive decline and reduced psychomotor performance already at age 38,[2] but analyses of the relation between dysregulation and work performance or safety are lacking so far. Especially in safety-critical jobs such as piloting, the use of physiological dysregulation indices appears to be an interesting concept for early detection of health-related risks at a subclinical stage.[19] Given the existing framework of aeromedical examinations in professional pilots, more systematic investigations of the effects of physiological dysregulation on flight performance might be implemented with comparatively little effort.

The second selected medical predictor, alanine aminotransferase, is a marker related to liver cell necrosis used in the diagnosis of liver conditions, including alcoholic or nonalcoholic fatty liver disease. This result may evoke associations of a possible role of alcohol use,[6] but clearly the exploratory nature of our findings, especially regarding this predictor, which was quite unstable across Random Forest fits in terms of predictive power, prohibits any such speculations given the sensitive nature of the topic; however, for purposes of replication and confirmation, this marker might be included in future investigations of the relation between pilot health and performance.

Among the limitations to this study, the most immediately apparent is the small sample size, which highlights a drawback of our approach compared to, for example, the use of occurrence reports: collecting a range of data sources, each with its own mechanisms of sample attrition, from an inherently small population (HEMS pilots) will almost inevitably lead to a small sample size. On the other hand, it should be kept in mind that our approach was motivated precisely by the question of whether it is possible to obtain interpretable results when the number of potential predictors is large relative to the sample size, a situation which is not uncommon in human factors aviation safety studies that are not purely based on retrospective review of occurrence reports or administrative records.

Moreover, as is the case with explorative research in general, there is also a threat of false-positive findings due to many simultaneously assessed associations. It should be noted, however, that the directions of the selected variables' effects appear to be generally plausible and that there is a relatively clear separation between the selected variables and unselected variables in terms of the variable importance scores. The logic behind the chosen selection threshold also seems to imply some protection against capitalization on chance since the absolute value of the minimum observed variable importance score should be expected to increase with the number of noise predictor variables involved (whose variable importance should vary randomly around zero). Finally, in the light of the tradeoff between type I and type II errors in statistical decision-making, it has been suggested that there should be a focus on minimizing the latter error in aviation safety research due to the potentially grave implications of false-negative findings.[8]

With regards to the aeromedical data, there was additionally the problem of a time lag between the last available assessment and the simulator session, which was very large (2–2.5 yr) in some of the biomarkers, including alanine aminotransferase, which had been selected as an informative predictor by the Random Forest procedure. According to linear mixed model analyses of time-stability of biomarker levels we conducted earlier, between 44 and 95% of the total variation in biomarker levels across average follow-ups of 5.2 to 8.2 yr were due to differences in pilot averages across time, indicating considerable stability of interindividual differences in the biomarkers (for alanine aminotransferase specifically, the respective figures were 70% and 5.2 yr). One might, therefore, assume that differences at the time of last available assessment carried over to the simulator session to some extent.

Finally, in contrast to occurrence report studies, our outcome of simulator flight performance can be viewed only as a proxy to the actual outcome of interest. Thus, in order to achieve an optimal ecological validity, we consulted extensively with experienced HEMS pilots and flight instructors in the selection of malfunction scenarios as well as in devising the scoring procedure.

To conclude, we identified three well-interpretable predictors of HEMS pilot simulator flight performance (two occupational stressors and an index of physiological dysregulation) from a broad array of candidates by exploiting the capability of the Random Forest machine learning method to select important predictors even when their number is large relative to the sample size. The predictors were taken from different sources (self-report, medical examinations) and covered different aspects of potential relevance to the error chain as outlined in modern systemic human factors safety approaches. Although our study is explorative in nature, which precludes confident statements about concrete measures to improve safety in HEMS, the results do suggest that the effect of working conditions and their perception by the pilots deserve further scrutiny. For example, the role of work stressors on HEMS pilots' subjective well-being might be investigated; well-being and mental health of professional pilots have received considerable attention recently,[1] but data for HEMS pilots are lacking. The physiological dysregulation construct is an intriguing potential tool for early recognition of latent pathology in professional pilots. While our dysregulation index was of a somewhat ad-hoc nature constrained by data availability, the utility of current measures of dysregulation derived from theoretical considerations[2,13] for screening, prevention, and selection purposes in aeromedical examinations might be further investigated. Finally, our study showcases the potential of the Random Forest method in the field of aviation human factors. For example, it could be applied to appropriately quantified information from accident investigation databases to identify factors associated with accident lethality. However, abundant data are also collected in everyday aviation operations and, with some effort invested into database normalization, information about the effect of operative conditions (e.g., timing of missions, weather, geographical location) on mission safety parameters may be quantitatively analyzed by individual HEMS operators using Random Forests. In a more ambitious approach, a framework for a common harmonized database which might include organizational, operational, administrative, and even aeromedical information could be established between operators or also between different aviation sectors. With such a large-scale database, the full potential of machine learning methods, which are designed to handle large amounts of information, could be brought to bear. Careful consideration would need to be given to feasibility (e.g., due to data comparability, data protection, and confidentiality issues) with this approach. In any case, the presented use of the Random Forest method may be a fruitful addition to existing risk analysis tools, helping operators to think "outside the box" in their efforts to identify additional flight safety measures.

ACKNOWLEDGMENTS

This work was supported by an unrestricted educational grant from the European HEMS & Air Ambulance Committee e.V. (EHAC) to the Institute and Clinic for Occupational, Social, and Environmental Medicine. Freedom from interference in any stage of the research process was laid down contractually between EHAC and the research project team, represented by the manuscript authors. EHAC had no role in the formulation of the study design, the collection, analysis, and interpretation of the data, the writing of the report, or the decision to submit the paper for publication.

Authors and affiliations: Hans Bauer, Dipl. Psych., M.Sc., Staburo GmbH, Munich, Germany; and Dennis Nowak, Prof., Dr. med., and Britta Herbig, PD, Dr. phil., Institute and Clinic for Occupational, Social, and Environmental Medicine, WHO Collaborating Centre for Occupational Health, University Hospital of LMU Munich, München, Germany.

REFERENCES

1. Aerospace Medical Association Working Group on Pilot Mental Health. Pilot Mental Health: Expert Working Group Recommendations—Revised 2015. Aerosp Med Hum Perform. 2016; 87(5):505–507.
2. Belsky DW, Caspi A, Houts R, Cohen HJ, Corcoran DL, et al. Quantification of biological aging in young adults. Proc Natl Acad Sci USA. 2015; 112(30):E4104–E4110.
3. Boyd DD. Causes and risk factors for fatal accidents in non-commercial twin engine piston general aviation aircraft. Accid Anal Prev. 2015 77: 113–119.
4. Bürkle A, Moreno-Villanueva M, Bernhard J, Blasco M, Zondag G, et al. MARK-AGE biomarkers of ageing. Mech Ageing Dev. 2015; 151:2–12.
5. Conroy RM, Pyörälä K, Fitzgerald AP, Sans S, Menotti A, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003; 24(11):987–1003.
6. Cook CC. Alcohol and aviation. Addiction. 1997; 92(5):539–555.
7. Harrell FE. Regression modeling strategies. With applications to linear models, logistic regression, and survival analysis. 2nd ed. Cham: Springer; 2015:67–72.
8. Harris D. The importance of the Type II error in aviation safety research. In: Farmer E, editor. Stress and error in aviation. Aldershot: Avebury Technical; 1991:151–157.
9. Harrison SL, Ding J, Tang EY, Siervo M, Robinson L, et al. Cardiovascular disease risk models and longitudinal changes in cognition: a systematic review. PLoS One. 2014; 9(12):e114431.
10. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning. Data mining, inference, and prediction. 2nd ed. New York: Springer; 2017:369–370.
11. Judge TA, Thoresen CJ, Bono JE, Patton GK. The job satisfaction-job performance relationship: a qualitative and quantitative review. Psychol Bull. 2001; 127(3):376–407.
12. Lepine JA, Podsakoff NP, Lepine MA. A meta-analytic test of the challenge stressor-hindrance stressor framework. An explanation for inconsistent relationships among stressors and performance. Acad Manage J. 2005; 48(5):764–775.
13. Milot E, Morissette-Thomas V, Li Q, Fried LP, Ferrucci L, Cohen AA. Trajectories of physiological dysregulation predicts mortality and health outcomes in a consistent manner across three populations. Mech Ageing Dev. 2014; 141–142:56–63.
14. Nahrgang JD, Morgeson FP, Hofmann DA. Safety at work: a meta-analytic investigation of the link between job demands, job resources, burnout, engagement, and safety outcomes. J Appl Psychol. 2011; 96(1):71–94.
15. Piton A, Poynard T, Imbert-Bismut F, Khalil L, Delattre J, et al. Factors associated with serum alanine transaminase activity in healthy subjects: consequences for the definition of normal values, for selection of blood donors, and for patients with chronic hepatitis C. MULTIVIRC Group. Hepatology. 1998; 27(5):1213–1219.
16. Reason J. Human error: models and management. BMJ. 2000; 320(7237): 768–770.
17. Shappell S, Detwiler C, Holcomb K, Hackworth C, Boquet A, Wiegmann DA. Human error and commercial aviation accidents: an analysis using the human factors analysis and classification system. Hum Factors. 2007; 49(2):227–242.
18. Siegrist J, Dragano N, Nyberg ST, Lunau T, Alfredsson L, et al. Validating abbreviated measures of effort-reward imbalance at work in European cohort studies: the IPD-Work consortium. Int Arch Occup Environ Health. 2014; 87(3):249–256.
19. Sluiter JK. High-demand jobs: age-related diversity in work ability? Appl Ergon. 2006; 37(4):429–440.
20. Stansfeld S, Candy B. Psychosocial work environment and mental health – a meta-analytic review. Scand J Work Environ Health. 2006; 32(6):443–462.
21. Stekhoven DJ, Bühlmann P. MissForest – non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012; 28(1):112–118.
22. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics. 2008; 9(1):307.
23. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods. 2009; 14(4):323–348.
24. Tanguy L, Tulechki N, Urieli A, Hermann E, Raynal C. Natural language processing for aviation safety reports. From classification to interactive analysis. Comput Ind. 2016; 78:80–95.
25. Taylor JL, O'Hara R, Mumenthaler MS, Yesavage JA. Relationship of CogScreen-AE to flight simulator performance and pilot age. Aviat Space Environ Med. 2000; 71(4):373–380.
26. Tetrick LE, LaRocco JM. Understanding, prediction, and control as moderators of the relationships between perceived stress, satisfaction, and psychological well-being. J Appl Psychol. 1987; 72(4):538–543.
27. von Thaden T, Wiegmann D, Shappell S. Organizational factors in commercial aviation accidents. Int J Aviat Psychol. 2006; 16(3):239–261.
28. Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. Eur J Epidemiol. 2009; 24(12):733–736.
29. Young JA. The effects of life-stress on pilot performance. Moffitt Field (CA): Ames Research Center; 2008.