safety

IMPACT FACTOR 2.4

CITESCORE 3.9

Article

Editor's Choice

Evaluation of a Biomathematical Modeling Software Tool for the Prediction of Risk in Flight Schedules Compared Against Incidence of Fatigue Reports

Jaime K. Devine, Jake Choynowski and Steven R. Hursh

Special Issue

Aviation Safety—Accident Investigation, Analysis and Prevention

Edited by

Dr. Douglas D. Boyd and Prof. Dr. Nihad E. Daidzic

MDPI

https://doi.org/10.3390/safety11010004

safety

MDPI

Article

Evaluation of a Biomathematical Modeling Software Tool for the Prediction of Risk in Flight Schedules Compared Against Incidence of Fatigue Reports

Jaime K. Devine [1,*], Jake Choynowski [1] and Steven R. Hursh [1,2]

[1] Institutes for Behavior Resources, 2104 Maryland Ave, Baltimore, MD 21218, USA; jchoynowski@ibrinc.org (J.C.); shursh@ibrinc.org (S.R.H.)
[2] Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, 4940 Eastern Ave, Baltimore, MD 21224, USA
* Correspondence: jdevine@ibrinc.org

Abstract: Background: Modeling tools should be tested against real-world outcomes to confirm their predictive ability compared to random chance. Insights is an analytical tool within the biomathematical modeling software SAFTE-FAST that identifies work patterns that consistently result in elevated fatigue risk. This study investigated the ability of Insights to correctly identify duties with an associated fatigue report using previously collected flight schedule and report data. Methods: Planned and completed flight roster schedules were analyzed using SAFTE-FAST Insights after the rosters had been flown. Fatigue reports were independently linked to planned and completed schedules at the duty level. Odds ratio (OR) analysis investigated the ability of Insights to predict which duties would be linked to a fatigue report. Differences in duties were compared using a one-way analysis of variance (ANOVA) and a two-sample *t*-test. Results: There were 157 fatigue reports out of 78,061 planned duties and 235 fatigue reports out of 82,612 completed duties. Insights had 3.04 odds of correctly identifying fatigue reports in planned duties but 0.41 odds for completed duties. Discussion: Insights showed good odds of correctly identifying a fatigue report duty using planned schedules but poor odds of identifying a fatigue report duty from completed schedules. Completed duties started later in the day and were shorter in duration than planned duties. Day-of-operations schedule changes may have reduced the fatigue risk in response to the fatigue reports.

Keywords: fatigue risk management; biomathematical modeling; fatigue reporting; aviation; odds ratio aviation accidents and investigations; aviation safety; human factors; safety management systems

Academic Editor: Raphael Grzebieta

Received: 3 September 2024
Revised: 22 December 2024
Accepted: 27 December 2024
Published: 7 January 2025

Citation: Devine, J.K.; Choynowski, J.; Hursh, S.R. Evaluation of a Biomathematical Modeling Software Tool for the Prediction of Risk in Flight Schedules Compared Against Incidence of Fatigue Reports. *Safety* **2025**, *11*, 4. https://doi.org/10.3390/safety11010004

1. Introduction

Fatigue is a serious concern in commercial aviation. Fatigue has been identified as either the cause of or a contributing factor in several major aircraft accidents, with 23% of all major aviation accidents between 2001 and 2012 attributed to fatigue [1,2]. The prevalence of fatigue as a factor in accidents like the 2009 crash of Colgan Air Flight 3407 in Buffalo, New York, which led the Federal Aviation Administration (FAA) to establish flight and duty limitations and rest requirements for pilots operating passenger flights under the Federal Aviation Regulations Part 117 [3]. Part 117 also outlines the necessary components of a fatigue risk management system (FRMS) that commercial airlines must employ to protect against safety hazards due to fatigue. An FRMS is a data-driven management system used to mitigate the effects of fatigue and the safety risks associated with fatigue-related errors in operations [3]. Among other components, an FRMS must include a method for fatigue monitoring and a fatigue reporting system [3].

A fatigue report should be submitted by a crew member when they feel that they cannot complete their duty as scheduled due to fatigue, similar to "calling in sick". Pilots and crew members are encouraged to submit a voluntary fatigue report as part of their organization's FRMS. The International Federation of Airline Pilots' Associations (IFALPA) Fatigue Reporting Guidance names the identification of risk, establishment of accumulated fatigue issues over time, and safety system information as the main reasons an operator should submit a report [4]. Many airlines allow pilots to submit a fatigue report in real time using an electronic submission system. Reports generally include the time of the occurrence and the flight details in addition to the information about fatigue.

The IFALPA Fatigue Reporting Guidance document also proposes that a sufficient set of fatigue reports can be used to conduct a trend analysis to identify recurring sources of fatigue. However, the formatting and process of submitting fatigue reports differ between organizations, and the information included in the report can be qualitative rather than quantifiable. These features complicate an analyst's ability to determine trends in the dataset. Analysts may not be able to observe how the entire roster (i.e., the arrangement of flight time, standby periods, and days off) led to an accumulation of fatigue over the course of the schedule if the report is only linked to a specific flight. They may not be able to determine whether either acute or chronic sleep debt impacted alertness or if sleep opportunities were curtailed because of insufficient opportunity, a personal choice of how to use a sleep opportunity, or issues with sleep quality beyond any self-report provided by the operator who submits the report. In cases where the report cannot be linked back to the schedule roster for which it was submitted, it may be impossible to determine any trends beyond content review. Another issue is that analyses conducted within an organization rarely get disseminated by academic publications, meaning that the methodology used to evaluate trends in fatigue is neither standardized nor readily reproducible across organizations. Commercial organizations may be hesitant to publicize any data that could reflect negatively on the airline, even when disseminating the information would help improve safety outcomes in the long run.

Not only is it difficult to conduct a trend analysis of fatigue reports, it is also not always clear how the results can be applied to reducing fatigue risk in future operations. One method that airlines use to prospectively evaluate fatigue risk is biomathematical modeling. The FAA advises airlines to use commercially available biomathematical computer models to assess average performance capability from sleep/wake history, circadian factors, and duty schedule information as part of their FRMS [5]. Modeling uses scientific principles about fatigue to identify specific operational factors that may contribute to performance impairment [5,6]. Models can be used in the planning stage to avoid creating a fatiguing schedule in the first place. Models can also be used to retrospectively evaluate fatigue risk during operations, such as during an incident or an accident investigation [7–9]. Continued data analysis and adjustments in response to new information help to improve a model's utility over time. However, changes to how a model predicts fatigue risk should be evaluated against an objective measure of performance, fatigue, or other real-world outcome to establish its accuracy.

Fatigue reports are an underutilized real-world measure of fatigue risk that can be used as a means to evaluate the accuracy of predictive modeling tools. In turn, modeling may help identify the root causes of fatigue during a trend analysis of fatigue reports and be configured to flag similar situations during a prospective analysis of future schedules so that the airline can impose a mitigation. SAFTE-FAST is biomathematical modeling software that uses the Sleep, Fatigue, and Task Effectiveness (SAFTE) model and the Fatigue

Avoidance Scheduling Tool (FAST) to predict fatigue and optimize schedules in commercial aviation [3,10]. SAFTE-FAST is used by airlines to evaluate fatigue risk during the planning of pilot rosters for narrow-body and wide-body aircraft. Narrow-body aircraft refers to airplanes with a single aisle between rows of seats in the main cabin, while wide-body aircraft refers to airplanes with up to two aisles in the main cabin.

Airlines track rosters using identifying codes that are unique to individual schedules within the full roster dataset, called schedule IDs. The airline uses SAFTE-FAST to examine planned rosters before they are flown and also tracks the schedules after they are flown. Schedule IDs change between planned rosters and completed rosters depending on the number of changes that occur between the planning process and the completion of the roster. Fatigue reports can be entered into SAFTE-FAST as an event marker and linked to schedules using the schedule ID.

SAFTE-FAST has an analytical tool called Insights that uses signal detection theory [11] to identify patterns of schedule features that consistently result in greater fatigue [12]. Testing the ability of Insights to correctly predict duties that result in real-world outcomes related to fatigue, such as the submission of a fatigue report, is an important field validation of the tool. This study investigated the ability of a biomathematical modeling tool (SAFTE-FAST Insights) to correctly identify flight duty periods that resulted in the submission of a fatigue report based on model estimates of alertness, sleep debt, and workload. This analysis was conducted using previously collected flight schedule data from normal operations originally collected by a major United States-based international commercial airline. The airline shared one month's worth of planned and completed rosters and their associated fatigue reports for the narrow-body fleet. An odds ratio analysis compared the extent to which duties that were identified as high risk by Insights (hits) predicted the submission of a fatigue report for that duty period of planned and as-flown rosters.

Managing fatigue risk in aviation requires a significant amount of planning, data collection, and data analysis. Newly developed analytical tools must be evaluated against objective measures of fatigue to establish their accuracy and should also take advantage of the existing data about fatigue that airlines collect as part of routine safety monitoring. The goal of this study was to evaluate a SAFTE-FAST Insights analysis of flight rosters against real-world fatigue reports using an odds ratio analysis to establish Insight's ability to identify risk from schedule features.

2. Materials and Methods

2.1. Flights Rosters and Fatigue Reports

A major United States-based commercial airline agreed to share data collected as part of their normal operations with the study team for secondary analysis under the condition of anonymity. Secondary use of the data was deemed non-human subjects research by Salus IRB on 18 September 2024 (Study ID: 24595) and was conducted in accordance with the Declaration of Helsinki. Planned and completed flight roster schedule information from September 2023 was provided by the airline's Human Factors team after the completion of all rosters for secondary use by the study team. Planned roster information included the arrangement of flight times and locations where pilots were originally scheduled to fly throughout the month of September. Completed roster information included the arrangement of flight times and locations as flown following any changes to the original schedule for pilots. Schedules were assigned to pilots using a unique schedule identifier (ID). Fatigue reports were submitted through the airline's electronic fatigue reporting system. As part of their normal job requirements, pilots are instructed to submit a fatigue report in instances when they feel that they are impaired or could be in a reduced state of alertness during an imminent flight duty period. Pilots were provided with the schedule ID upon assignment to a roster; fatigue reports that included schedule ID information could be linked to rosters using these IDs. Schedule IDs changed in response to schedule changes but not crew changes (for example, the substitution of one pilot for another). Some schedules were significantly altered between the planned and completed roster datasets, resulting in a functional elimination of the original planned roster (for example, cancellation of all duties within the roster). Functionally eliminated schedules could not be linked to fatigue reports because they did not progress past the planning stage.

2.2. Biomathematical Modeling Software

SAFTE-FAST version 6.10 was used for all modeling in this project. SAFTE-FAST is the brand name of Fatigue Risk Management System (FRMS) software provided by the Institutes for Behavior Resources, Inc. (IBR, Baltimore, MD, USA) to a variety of operational organizations, including commercial aviation. SAFTE-FAST is a two-step, three-process model that estimates sleep patterns around work duties and then estimates performance levels. The three processes involved are circadian function, homeostatic sleep reservoir, and sleep inertia. Auto-Sleep is the sleep estimator in SAFTE-FAST that uses information about work events, time of day, and prior sleep to predict average sleep decisions under operational constraints. The two major output variables used to predict risk in SAFTE-FAST are Effectiveness and Sleep Reservoir, more information can be found in Supplementary Materials [13,14].

Effectiveness is one of the primary output metrics used by SAFTE-FAST. Effectiveness represents the speed of performance on the Psychomotor Vigilance Test (PVT), scaled as a percent (%) of a fully rested person's normal best performance [15]. Minimal effectiveness during critical phases of flight is considered to be 77% [3]. Effectiveness scores range from 0 to 100; scores below 77 indicate a significant risk for fatigue, equivalent to 18.5 h of continued wakefulness for a fully rested person, or a blood alcohol concentration of 0.05 g/dL [16]. Sleep Reservoir is another primary output metric used by SAFTE-FAST. Sleep Reservoir is a measure of the capacity to perform cognitive work as a function of prior sleep. A score of 75 is roughly equivalent to a cumulative sleep debt of eight hours or the loss of one full night of sleep for the average person [10,13]. The workload calculator is comparable to the NASA task load index (TLX) and is used to create weighted triggers for job tasks or operational factors that may impact cognitive demands [17,18]. Workload is then normalized to a 100-point scale, with a score of 0 representing no workload and 100 representing the highest level of workload.

Insights is an analytical tool in SAFTE-FAST that is built on the understanding that schedule-induced fatigue is the product of a pattern of duties and prior rest, not isolated events in time. Insights uses signal detection theory to identify patterns in work-related features in a dataset that consistently results in fatigue risk across the workforce using user-defined criteria. Insights scores for this analysis were based on either (1) minimum Effectiveness, (2) minimum Effectiveness and double-weighted minimum Sleep Reservoir, and (3) Minimum Effectiveness, Sleep Reservoir and maximum Workload (inverted). The results from Insights analysis were exported as a csv file for further analysis at the duty level.

2.3. Statistical Analysis

Two independent SAFTE-FAST project files were created in SAFTE-FAST 6.10 to model (1) planned rosters and (2) completed rosters using the airline's default set-up package for a narrow-body fleet with custom triggers for workload. Fatigue reports were included as markers in each project file using Schedule ID to link the report to the schedule. Incomplete schedules, schedules with input data clerical errors, and schedules that operated on wide-body aircraft were removed from the project file prior to analysis. Figure 1 depicts the data flow chart from the receipt of schedule data from the airline to the completion of statistical analysis.

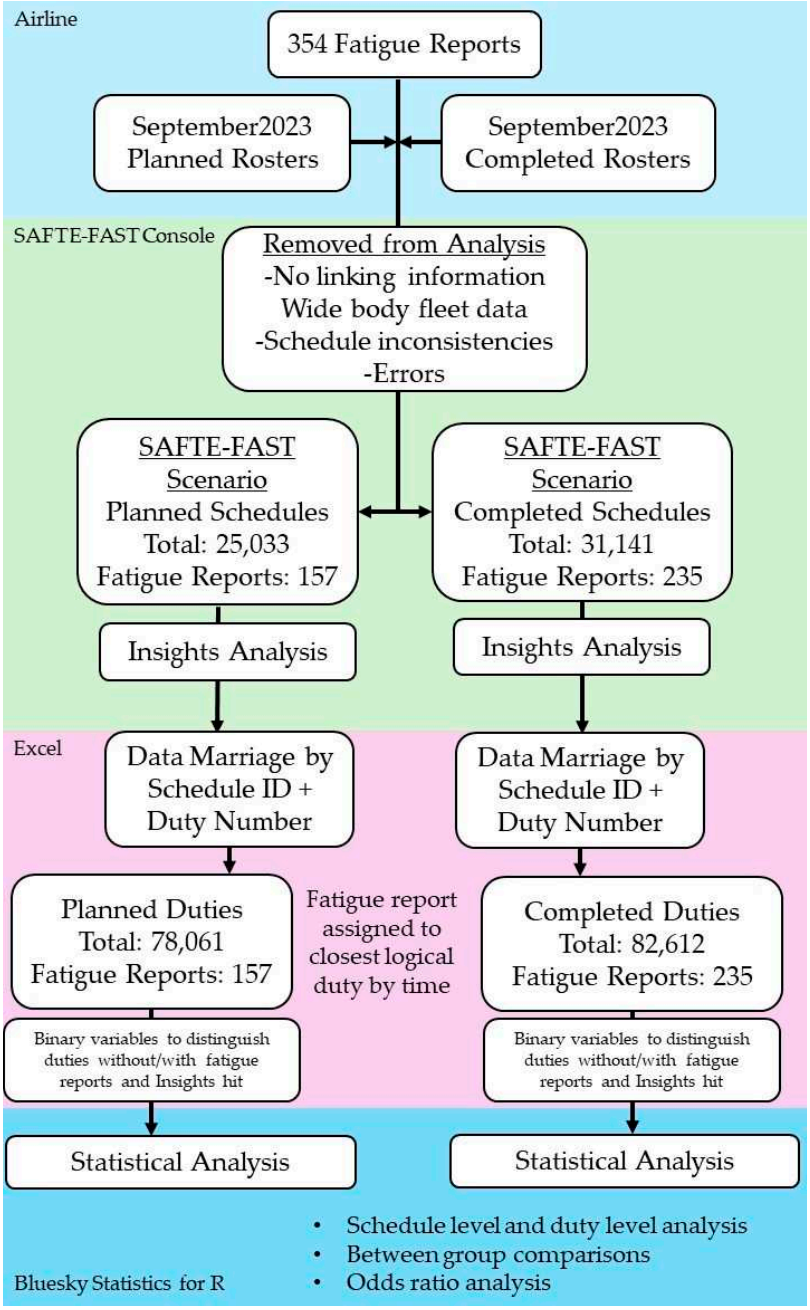

Figure 1. Data analysis flow chart. Flow chart depicting the flow of data across analytical platforms (background). Analyses were conducted independently for planned and completed schedules but followed the same flow across platforms and analytical procedures.

SAFTE-FAST data for all planned and completed schedules were exported as csv files. Data were compiled in Excel MP 15. Each schedule contained between 1 to 7 crewing duties. Unique identifiers were created for each duty within a schedule. Duties were assigned two distinct binary classifications based on (1) whether it had been identified by Insights (1) or not (0) and (2) whether a fatigue report had been submitted (1) or not (0) for each duty. Fatigue reports were assigned to the closest logical duty that occurred within a two-day range (2880 min before and after the duty) of the submission of the report. Fatigue reports were assigned to only one duty per schedule except in cases where multiple fatigue reports had been submitted during the same schedule. Fatigue reports were assigned to crewing duties. If no crewing duties occurred within the logical window, fatigue reports were assigned to non-crewing duties (e.g., relief pilot duty).

Figure 2 illustrates categories and logic for the odds ratio (OR) analysis. An OR analysis was conducted using the chi-square test and the following for categories: (A) true positives: duties that had been identified by Insights and for which a fatigue report has been submitted; (B) false positives: duties that were identified by Insights but did not have an associated fatigue report; (C) false negative: duties with a fatigue report but no Insights identification; and (D) true negative: duties that were not associated with either an Insights identification or a fatigue report. ORs were conducted independently for each Insights. A global OR was also conducted; global ORs considered true positives consist of duties with a fatigue report that had been identified by any Insights analysis criteria. False positives were considered to be any duty that was identified by any Insights analysis criteria but did not have an associated fatigue report.

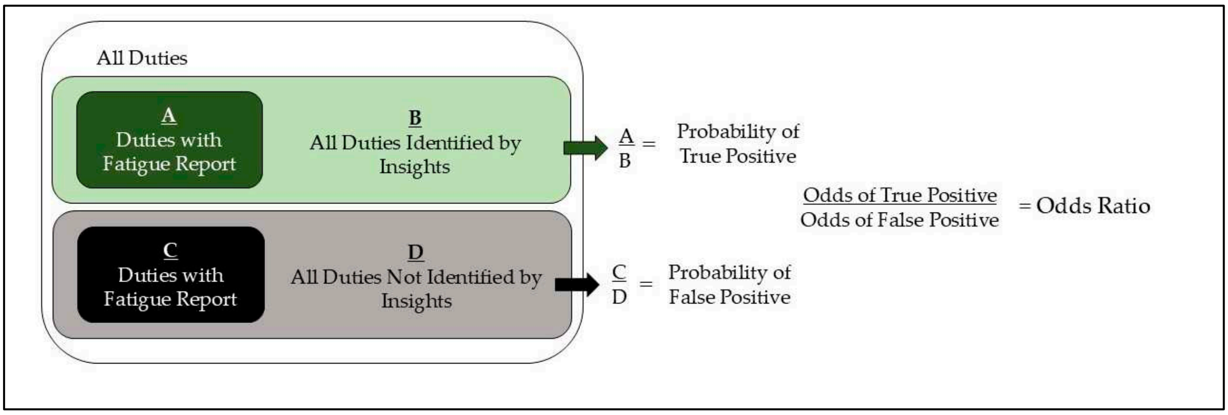

Figure 2. Duty distribution and logic for odds ratio analysis. Flow chart depicting the distribution of duties across the four OR categories: (**A**) Insights and Fatigue Report; (**B**) Insights but No Fatigue Report; (**C**) No Insights but Fatigue Report; and (**D**) No Insights or Fatigue Report (**D**) as well as the logical equation for the OR analysis.

Type II one-way analysis of variance (ANOVA) explored differences in duty start, duty duration, minimum Effectiveness, minimum Sleep Reservoir, and maximum Workload between duties in global OR categories A, B, C, and D for planned and completed duties with Tukey post hoc estimations [9]. Effect sizes are reported as eta squared (η^2), with $\eta^2 \geq 0.01$ indicating a small effect, $\eta^2 \geq 0.06$ indicating a medium effect, and $\eta^2 \geq 0.14$ indicating a large effect [19]. Overall differences in duty start, duty duration, minimum Effectiveness, minimum Sleep Reservoir, and maximum Workload between planned and completed duties were explored using a two-sample t test assuming equal variance. The effect size for t tests was estimated using Cohen's d, with $d \geq 0.2$ indicating a small effect, $d \geq 0.5$ indicating a medium effect, and $d \geq 0.8$ indicating a large effect [19]. All statistical analyses were performed using BlueSky Statistics 10.3.4 for Windows (BlueSky Statistics, Chicago, IL, USA) [20].

3. Results

A total of 354 fatigue reports were submitted by airline pilots during the month of September 2023. Not every fatigue report could be linked back to a planned or completed roster. The process of linking fatigue reports to planned and completed rosters is outlined in Figure 1. For the planned roster dataset, 157 fatigue reports could be linked to one crewing duty (meaning a work event within their whole schedule when the pilot was on duty to fly the aircraft) within the planned roster dataset by time and date. For the completed roster dataset, 231 fatigue reports could be linked to one crewing duty by time and date, and 4 fatigue reports could be linked to non-crewing duties, for a total of 235 fatigue reports. In total, there were 78,061 duties included for the planned duty OR analysis and 82,612 duties included for the completed duties OR analysis. Table 1 shows the crosstabs of duties across the four possible categories: (A) Insights and Fatigue Report; (B) Insights but No Fatigue Report; (C) No Insights but Fatigue Report; and (D) No Insights or Fatigue Report, with total number of duties and the OR results for each Insights analysis and the global analysis.

Table 1. Odds ratio analysis of planned and completed duties.

Planned Duties				**Completed Duties**			
Effectiveness	Fatigue Report	No Fatigue Report	Total	**Effectiveness**	Fatigue Report	No Fatigue Report	Total
Insights	22	4266	4288	Insights	5	4138	4143
No Insights	135	73,638	73,773	No Insights	230	78,239	78,469
Total	157	77,904	78,061	Total	235	82,373	82,612
OR = 2.81	CI: [1.79, 4.42] $p < 0.001$			OR = 0.41	CI: [0.17, 1.00] $p = 0.04$		
Effectiveness + Sleep Reservoir × 2	Fatigue Report	No Fatigue Report	Total	**Effectiveness + Sleep Reservoir × 2**	Fatigue Report	No Fatigue Report	Total
Insights	25	4777	4802	Insights	2	3677	3679
No Insights	132	73,127	73,259	No Insights	233	78,700	78,933
Total	157	77,904	78,061	Total	235	82,377	82,612
OR = 3.38	CI: [1.88, 4.45] $p < 0.001$			OR = 0.18	CI: [0.05, 0.74] $p = 0.007$		
Effectiveness + Sleep Reservoir + Workload	Fatigue Report	No Fatigue Report	Total	**Effectiveness + Sleep Reservoir + Workload**	Fatigue Report	No Fatigue Report	Total
Insights	21	4281	4302	Insights	1	1788	1789
No Insights	136	73,623	73,759	No Insights	234	80,589	80,823
Total	157	77,904	78,061	Total	235	82,373	82,612
OR = 2.65	CI: [1.68, 4.21] $p < 0.001$			OR = 0.19	CI: [0.03, 1.38] $p = 0.07$		
Global	Fatigue Report	No Fatigue Report	Total	**Global**	Fatigue Report	No Fatigue Report	Total
Insights	33	6264	6301	Insights	6	4911	4917
No Insights	124	71,640	71,760	No Insights	229	77,466	77,695
Total	157	77,904	78,061	Total	235	82,377	82,612
OR = 3.04	CI: [2.07, 4.47] $p < 0.001$			OR = 0.41	CI: [0.18, 0.93] $p = 0.03$		

Insights correctly identified a grand total of 33 out of 157 planned duties that had an associated fatigue report across all Insights analyses. Some fatigue reports were associated with more than one Insights category. Two planned duties with fatigue reports were identified exclusively by minimum Effectiveness alone, and 15 were identified by both minimum Effectiveness and minimum Effectiveness in combination with Sleep Reservoir. Six planned duties with a fatigue report were identified exclusively by Insight's criteria of minimum Effectiveness, Sleep Reservoir, and maximum Workload. Ten planned duties were identified by all three Insights analyses. A planned duty that was identified by one or more Insights analyses was between two and three times more likely to have an associated fatigue report, as shown in the left-hand column of Table 1. The Global OR of 3.04 was based on the 33 fatigue reports that were identified by one or more Insights analyses.

In contrast, Insights correctly identified a grand total of 6 out of 235 completed duties that had an associated fatigue report across all Insights analyses. Three completed duties with fatigue reports were identified exclusively by minimum Effectiveness alone, and two were identified by both minimum Effectiveness and minimum Effectiveness in combination with Sleep Reservoir. One completed duty was identified exclusively by minimum Effectiveness, Sleep Reservoir, and maximum Workload. No duties were identified by all three Insights analyses. A completed duty that was identified by one or more Insights analysis was less than half as likely as a random chance to have an associated fatigue report, as shown in the right-hand column of Table 1. Global OR categories (see Table 1 and Figure 3) were next used to explore differences in duty start, duty duration, minimum Effectiveness, minimum Sleep Reservoir, and maximum Workload. Duty start time in local time and duty duration by global OR category (see Figure 3) with ANOVA statistics are shown in Table 2. Duty timing, duration, and average SAFTE-FAST Effectiveness, Sleep Reservoir, and Workload by Global OR category are depicted in Figure 3 below for planned and completed duties.

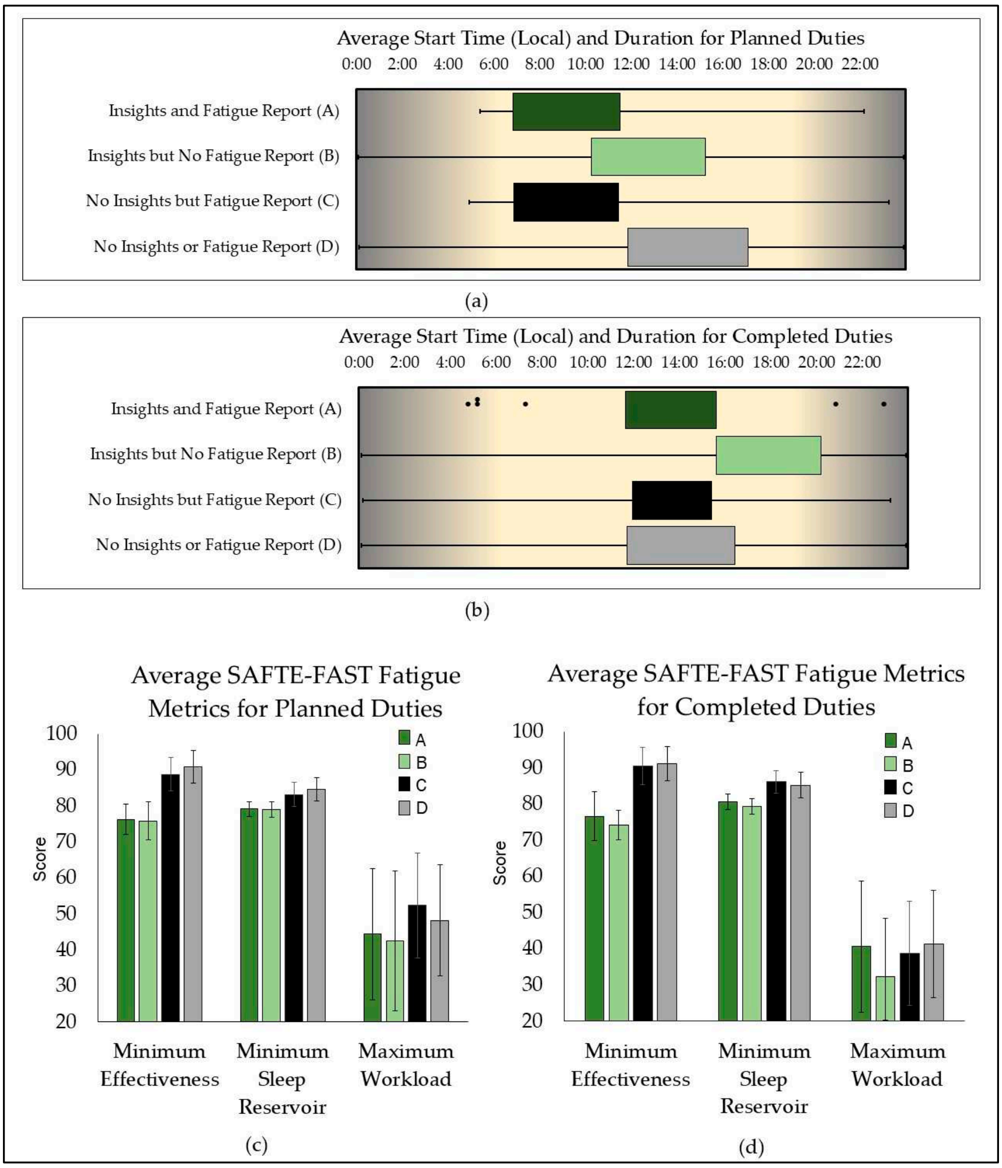

Figure 3. Duty features and SAFTE-FAST fatigue metrics by global OR category for planned and completed duties. (**a**,**b**) Gantt charts visualize the differences in duty timing and duty duration by global OR category across approximate local day for (**a**) planned duties and (**b**) completed duties. Approximate local day is indicated by lighter background shading between 0500 and 1900 h. Standard deviation of start times is indicated by black error bars and occurred across the 24 h day. Completed duties category A: Insights and Fatigue Report shows start times for each of the 6 duties in that category as a black dot rather than standard deviation. Standard deviation for duty duration and statistical significance can be found in Table 2. (**c**,**d**) Bar charts visualize the average minimum Effectiveness, minimum Sleep Reservoir, and maximum Workload by OR category during the duty period for (**c**) planned duties and (**d**) completed duties. Standard deviation is represented by error bars; statistical significance can be found in Table 3.

Table 2. Average start time, duty duration, and ANOVA by global OR category for planned and completed duties.

Planned	n	Duty Start Time (HH:MM)	Duty Duration (min)
Insights and Fatigue Report (A)	33	6:51 ± 1:27	280 ± 78
Insights but No Fatigue Report (B)	6264	10:16 ± 5:08	299 ± 108
No Insights but Fatigue Report (C)	128	6:51 ± 1:44	276 ± 67
No Insights or Fatigue Report (D)	71,640	11:52 ± 5:16	314 ± 97
ANOVA statistics	78,061	$F_{(3,78057)} = 222.16$ $p < 0.001$ $\eta^2 = 0.008$	$F_{(3,78057)} = 64.51$ $p < 0.001$ $\eta^2 = 0.003$
Completed	**n**	**Duty Start Time (HH:MM)**	**Duty Duration (min)**
Insights and Fatigue Report (A)	6	11:38 ± 8:25	238 ± 199
Insights but No Fatigue Report (B)	4911	15:37 ± 9:07	275 ± 69
No Insights but Fatigue Report (C)	229	11:58 ± 5:24	207 ± 105
No Insights or Fatigue Report (D)	77,466	11:43 ± 4:58	289 ± 106
ANOVA statistics	82,612	$F_{(3,82608)} = 827.19$ $p < 0.001$ $\eta^2 = 0.029$	$F_{(3,82608)} = 49.39$ $p < 0.001$ $\eta^2 = 0.002$

Figure 3a visualizes the differences in duty timing and duration by global OR category across approximate local day for planned duties. Duty start times were similar for duties with a fatigue report (A and C: t = 0.01, $p = 1.00$). Planned duties with a fatigue report started earlier in the day relative to other planned duties without a fatigue report (all $p < 0.001$). Duty duration was shorter for planned duties in category D, namely No Insights or Fatigue Report relative to categories identified by Insights (A or C, all $p < 0.001$). There were no other significant between-group differences identified by Tukey's post hoc analysis for planned duty schedule features (all $p > 0.06$).

Figure 3b visualizes the differences in duty timing and duration by global OR category across approximate local day for completed duties. Due to the low number of completed duties in Category A (n = 6), individual duty start times have been plotted instead of standard deviation bars. Start times for completed duties Category A clustered either in the early morning (between 5:23 and 7:58 A.M.) or late evening (21:25, 23:26), resulting in an average start time around 11:38 AM. The average duty start time was later for category B, Insights but No Fatigue Report, compared to other duties ($p < 0.05$). Duty duration tended to be shorter for duties with a fatigue report (A or C) compared to duties without a fatigue report (B or D; all $p < 0.001$). There were no other significant between-group differences identified by Tukey's post hoc analysis for completed duty schedule features (all $p > 0.10$).

SAFTE-FAST predictions of Effectiveness, Sleep Reservoir, and Workload by global OR category with ANOVA statistics are shown in Table 3 and Figure 3c,d. Figure 3c visualizes the average minimum Effectiveness, minimum Sleep Reservoir, and maximum Workload by OR category during the duty period for planned duties. For planned duties, minimum Effectiveness differed significantly between all categories, except those that had been identified by Insights (categories A and B; t = 0.03, $p = 1.00$). Minimum Sleep Reservoir also differed significantly, except between Insights categories (A and B; t = 0.46, $p = 1.00$). Maximum Workload between categories was significantly different, except when comparing category A, Insights and Fatigue Report, to either category B, Insights but No Fatigue Report (t = 0.02, $p = 1.00$) or D, No Insights or Fatigue Report (t = 2.08, $p = 0.16$).

Table 3. Minimum Effectiveness, minimum Sleep Reservoir, and maximum Workload by OR category for planned and completed duties.

Planned	n	Effectiveness	Sleep Reservoir	Workload
Insights and Fatigue Report (A)	33	75.88 ± 4.21	79.24 ± 2.10	42.52 ± 18.34
Insights but No Fatigue Report (B)	6264	75.88 ± 5.35	78.98 ± 2.13	42.57 ± 19.47
No Insights but Fatigue Report (C)	128	88.61 ± 4.89	83.10 ± 3.40	52.51 ± 14.49
No Insights or Fatigue Report (D)	71,636	90.92 ± 4.48	84.62 ± 3.24	48.22 ± 15.38
ANOVA statistics	78,061	$F_{(3,78057)}$ 21,074.91 $p < 0.001$ $\eta^2 = 0.45$	$F_{(3,78057)} = 6129.57$ $p < 0.001$ $\eta^2 = 0.19$	$F_{(3,78057)} = 252.22$ $p < 0.001$ $\eta^2 = 0.01$
Completed	**n**	**Effectiveness**	**Sleep Reservoir**	**Workload**
Insights and Fatigue Report (A)	6	76.67 ± 6.79	80.73 ± 2.13	40.60 ± 18.13
Insights but No Fatigue Report (B)	4911	74.26 ± 4.07	79.41 ± 2.12	32.35 ± 16.10
No Insights but Fatigue Report (C)	229	90.50 ± 5.21	86.18 ± 3.16	38.82 ± 14.37
No Insights or Fatigue Report (D)	77,466	91.18 ± 4.78	85.22 ± 3.53	41.33 ± 14.85
ANOVA statistics	82,612	$F_{(3,82608)} = 19{,}580.52$ $p < 0.001$ $\eta^2 = 0.42$	$F_{(3,82608)} = 4354.17$ $p < 0.001$ $\eta^2 = 0.14$	$F_{(3,82608)} = 558.23$ $p < 0.001$ $\eta^2 = 0.02$

Figure 3d visualizes the average minimum Effectiveness, minimum Sleep Reservoir, and maximum Workload by OR category during the duty period for completed duties. For completed duties, minimum Effectiveness and minimum Sleep Reservoir differed significantly between all categories except those that had been identified by Insights (categories A and B; t = 1.24, $p = 0.21$ and t = 0.94, $p = 0.35$, respectively). The maximum Workload was lower in category B, Insights but No Fatigue Report, compared to categories with no Insights (C and D; all $p < 0.001$). The Workload was also lower in category C, No Insights but Fatigue Report, compared to D, No Insights or Fatigue Report (t = 2.15, $p = 0.004$).

Table 4 presents the overall differences between planned and completed duties for average duty start time in local time, duty duration, minimum Effectiveness, minimum Sleep Reservoir, and maximum Workload. A two-sample *t* test indicated a small, significant effect for planned duties on average to have started earlier, have a longer duty duration, lower minimum Effectiveness, lower minimum Sleep Reservoir, and higher maximum Workload compared to completed duties (all $p < 0.001$, all d ≤ 0.45).

Table 4. Differences in duty start time, duty duration, minimum Effectiveness, minimum Sleep Reservoir, and maximum Workload between planned and completed duties.

Variable	Planned Duties (M ± STD)	Completed Duties (M ± STD)	Statistics
n	78,061	82,612	--
Duty Start Time (HH:MM)	11:43 ± 5:17	11:57 ± 5:23	t = 8.68; $p < 0.001$; $d = 0.04$
Duty Duration (min)	313 ± 98	282 ± 104	t = 60.25; $p < 0.001$; $d = 0.30$
Minimum Effectiveness (score)	89.71 ± 6.12	90.17 ± 6.21	t = 15.10; $p < 0.001$; $d = 0.08$
Minimum Sleep Reservoir (score)	84.16 ± 3.52	84.88 ± 3.73	t = 39.46; $p < 0.001$; $d = 0.20$
Maximum Workload (score)	47.77 ± 15.82	40.78 ± 15.08	t = 90.65; $p < 0.001$; $d = 0.45$

4. Discussion

This study evaluated SAFTE-FAST Insights predictions against real-world fatigue reports that have been linked to both duties in a planned schedule and duties from as-flown, completed schedules. The purpose of a predictive tool like Insights is to identify areas of fatigue risk in a planned schedule so that the airline can adjust the schedule to reduce risk before the flights are ever flown. A predictive tool that is no better than chance at identifying real-world fatigue is worse than ineffective; it jeopardizes safety by creating a false sense of security. Poor predictive ability puts pilots at risk of operating a series of duties that will result in fatigue. Therefore, predictive tools need to be tested against real-world objective measures, such as fatigue reports, to evaluate the ability of the tool to correctly identify areas of risk.

Insights effectively predicted fatigue risk in planned duties but not in completed duties. The odds of identifying a planned duty that later resulted in a fatigue report was between 2.65 to 3.38 times more likely than random chance using any of the Insights criteria or global OR described in Table 1. Insights' ability to predict duties in planned schedules bodes well for its use as a proactive tool for fatigue risk management. Insights did a poor job of identifying duties with fatigue reports in completed schedules. This may be due to the fact that the submission of the fatigue report resulted in a real-time change to the schedule. This explanation is supported by the data from Tables 2–4 and Figure 3. Figure 3a shows that the planned duties that were later linked to a fatigue report (categories A or C) started earlier in the morning in local time compared to duties with no fatigue report (categories B or D). However, Figure 3b shows very little difference in start times between categories, indicating that duties that were planned to start earlier did not actually start earlier than schedules with no fatigue report. Table 4 directly compares the duty start times and duration between planned and completed duties. Completed duties started later in the day and were shorter in duration than planned duties.

Day-of changes to flight schedules are commonplace in aviation. Changes may include delays due to weather, congestion, maintenance issues or staffing changes [21]. Any disruption to the day's flight plan can influence subsequent flights' schedules as well [22], making it difficult to determine the root cause of differences between the planned and completed rosters. The submission of a fatigue report may have resulted in a schedule change that reduced predictable fatigue risk when examined in SAFTE-FAST. This hypothesis may explain why Insights did a poor job of predicting duties with fatigue reports in the completed roster dataset. This hypothesis may also explain why duty start times for duties with a fatigue report were earlier in the planned analysis but not in the analysis of completed duties and why duty duration was shorter for completed duties with a fatigue report.

For the purposes of this analysis, fatigue reports were assigned to the closest logical duty within a two-day range, which included duties occurring before or after the submission of the report, since pilots may have submitted a report in anticipation of an upcoming fatiguing duty. Pilots used an electronic submission system, so it is possible that reports were submitted in a timely manner that allowed schedule changes to be made almost immediately. A departure delay would explain the later average start times seen in the completed roster dataset relative to planned duties (Figure 3a,b; Table 3). Duty duration and start time would then affect SAFTE-FAST predictions of fatigue risk. Fatigue reports were linked to schedule IDs, but schedule IDs were modified to reflect schedule changes. Reports submitted with a schedule ID that was later changed, or that lacked schedule ID information, could not be linked to rosters in either dataset. Improving the fatigue reporting system to include a pilot identifier or more information about the prior schedule could improve trend analysis of fatigue reports. However, pilots may not be as willing to submit a report if they were required to identify themselves or provide flight details in a fatigued state. Additionally, the current datasets do not include information about actions taken following the report submission. Further research with more detailed data is required to support or disprove the hypothesis that real-time changes reduce fatigue risk in operations.

Comparing the small number of duties with an associated fatigue report against the large number of duties without one constitutes a limitation to the interpretation of both the planned and completed duties. The *n* is reported for each OR category in Tables 1–3 to highlight the uneven distribution of duties across categories. Insights only picked up a maximum of 33 out of 157 planned duties with a fatigue report by combining the results from all Insights analyses into a global OR. However, there were only 157 instances of fatigue reports out of 78,061 planned duties, drawing to mind the idiom "finding a needle in a haystack". The haystack was even larger for completed duties—235 duties with a fatigue report out of a total of 82,612 duties. Insights identified only six completed duties with a fatigue report. This extremely small sample size relative to all other OR categories severely limits the interpretability of any comparisons between completed duties Category A and other groups, as well as the legitimacy of within-group averages. These numbers are not ideal for statistical testing but reflect the issue of using statistics to describe risk during real-world operations where fatigue reports are relatively rare, and crew members may rely on a number of other fatigue countermeasures other than reporting.

The real-world, secondary-use nature of these data should also be considered a limitation to generalizability. These datasets reflect planned and completed schedules only for narrow-body aircraft from a single carrier collected over a single month. Moreover, September 2023 was selected for this analysis because there was a high proportion of fatigue reports. Even though there were still relatively few schedules with a fatigue report compared to the total number of schedules overall, September 2023 should not be considered representative of normal monthly operations. The data used for this analysis focus on the schedule information and provides no information about the individual working the schedule. While it is appropriate for airlines to consider fatigue risk related to the schedules that they impose upon crew members, the lack of information about the individual who submitted the fatigue report limits interpretability, not only about the cause of the fatigue report but also the impact of the work schedule on submission of the report and vice versa.

The schedule ID naming system also limits our ability to track changes between planned and completed rosters. Schedule IDs could not be linked between many schedules in the planned and completed rosters because schedule changes resulted in ID changes. Even for schedules that had a matching ID between datasets, crew changes could not be accounted for because they were not documented. Another limitation is that not every fatiguing duty necessarily results in the submission of a fatigue report, and the reports in this study did not include details about the reason for submission. Fatigue can be due to a number of factors, only a few of which can reasonably be predicted from work schedule data alone. The SAFTE-FAST system is not capable of predicting fatigue due to illness, commotion, home life, personal choices in sleep behavior, or the presence of an undiagnosed sleep disorder, for example.

High workload alone appears to have limited predictive value. Maximum Workload was lower on average for duties that had been identified by Insights relative to duties that had not been for both planned and completed duties. However, high Workload combined with low Effectiveness and Sleep Reservoir accounted for 18% of the true positives included in the Global OR (6/33). These cases were predicted only when Workload was included in the Insights criterion. Workload in combination with low Effectiveness and Sleep Reservoir would appear to be useful for identifying more instances of fatigue risk compared to examining alertness alone.

Workload is a known contributor to fatigue in aviation [23,24] but exists outside the three-process model (circadian function, homeostatic sleep reservoir, and sleep inertia) commonly employed by biomathematical models to predict fatigue [25]. Workload, like fatigue, is an imprecise term that is commonly used to describe a number of factors. To complicate matters, either too much work (overload) or too little work (underload) can result in fatigue and an increased risk to safety [26]. Some sources of workload can be predicted from schedule features, such as time on task or the number of segments per day [24]. Workload due to a sudden change in the weather, delays, or psychological stressors [23,24] cannot currently be predicted by the SAFTE-FAST system. SAFTE-FAST contains a workload calculator [18], and operators can conduct a study to measure pilots' workload in association with their schedule features. Additionally, workload predictions for the airline in this study had been adapted from the results of another major U.S.-based carrier and may not have been precise enough to accurately estimate fatigue due to workload or been optimal for estimating workload-related fatigue in these specific operations. Without more information about the reason that pilots submitted a fatigue report, it is impossible to determine whether workload was not an important predictor of fatigue reporting or whether SAFTE-FAST predictions simply did not do a good job of modeling the impact of workload on fatigue. Improving Workload predictions in SAFTE-FAST would be expected to improve Insights' ability to identify fatiguing duties.

Despite these limitations, Insights demonstrated between two and three times the odds of correctly identifying a planned duty that later would receive a fatigue report based on a combination of SAFTE-FAST fatigue metrics. This finding indicates that Insights could be useful for the prospective analysis of schedules in order to eliminate fatigue risk before schedules are flown. Biomathematical models are used to evaluate whether fatigue contributed to past incidents, such as during an accident investigation [7,8], but the ultimate goal of fatigue risk management is to prevent adverse events from happening in the future. The ideal model would be able to both correctly identify risk as it occurred in prior instances and predict when the same level of risk appears in future operations so that the airline can take preventative measures. The utility of Insights can be improved by figuring out what information from previous fatigue-related events is most useful for trend analysis and retraining the model to look for similar patterns. This may be a difficult feat using fatigue reports since the submission of the report (at least in theory) should be addressed in a manner that adequately resolves the unacceptable level of risk [4]. If the fatigue risk is appropriately reduced as a result of the submission of a fatigue report, then studying the as-flown schedule would be ineffective. It is remindful of the classic story about survivorship bias from World War II, when Air Force engineers targeted areas of fighter planes that returned from combat with bullet holes in them for additional armament until mathematician Abraham Wald argued that if the plane could return with bullet holes, then the bullet had not hit a part of the plane that was necessary for a safe return. The real solution, Wald posited, was to add additional armor on the plane chassis in areas that had no bullet holes on the assumption that planes hit with bullets in those parts could not safely return for inspection [27]. This logical error in which attention is paid only to cases that have passed through a selective filter, known as survivorship bias, may help explain the OR findings from the completed schedules, in which Insights was worse than random chance at identifying duties with a fatigue report (OR = 0.41, Table 1). Perhaps instead of focusing on how to improve Insights for a better detection of duties with fatigue reports from as-flown data, a better strategy for effective fatigue risk management would be to evaluate what changes were made in response to the submission of a fatigue report that reduced the predictable risk. This interpretation requires further investigation and better tracking between the planned and completed schedules to confirm or reject.

An important follow-up study will be to evaluate if using Insights as a tool to help design rosters is related to an improvement in the objective measures of fatigue, like PVT response times, or real-time changes to flight schedules compared to schedule design alone. Insights should also be evaluated using data collected as part of a comprehensive scientific research study within a pilot population. Secondary use of operational data is an underutilized strategy for studying the occurrence of real-world fatigue but leaves many variables outside the investigators' control. The secondary use of data cannot replace the knowledge that can be gained by research studies designed specifically to answer a research question. Tools like Insights must be tested against real-world measures of fatigue to evaluate accuracy with the goal of constant improvement but should also be tested using data collected as part of a specific research study.

A preliminary test of the ability of Insights to predict risk based on the incidence of fatigue reports using previously collected operational data shows promise for planned roster information but not from completed roster data. Moreover, predictive ability was not greatly improved by including Workload as a factor for fatigue. Insight's predictive ability can be tested and improved upon in multiple ways. First, the logic underlying Insights' detection of patterns related to fatigue could be improved by training the model on data collected specifically to test the impact of work schedules on fatigue rather than secondary use data. Secondly, working with airline safety professionals to prospectively examine the usefulness of Insights to the planning process may help validate the usefulness of the tool to avoid fatigue risk and help develop a set of best practices for using Insights. Finally, a different analysis technique may be able to link fatigue risk prediction to the submission of a fatigue report within the completed roster dataset. It is difficult to pontificate upon an effective analysis technique without additional data about how the schedules changed between planned rosters and the completion of the roster. This is another area where a predetermined scientific study design would benefit the analysis plan.

5. Conclusions

Biomathematical modeling allows airlines to predict the risk associated with a proposed schedule beforehand, thus providing the opportunity to mitigate fatigue risk *a priori*, and reducing the likelihood of a fatigue-related accident or error occurring during commercial flights. However, predictive tools must be tested to establish their accuracy compared to real-world fatigue outcomes. This study tested SAFTE-FAST Insights' ability to detect fatigue risk relative to real-world fatigue reports to establish the tool's accuracy. Insights performed well only when using planned schedule data. Real-time changes to schedules reduced the overall fatigue risk when looking at the completed schedules relative to the planned ones but also resulted in poor predictive power by Insights. Insights may be useful as a proactive tool for fatigue risk management during the schedule planning process but still needs to be tested using a controlled scientific study design to see how using Insights as part of the planning process results in less fatigue during subsequent operations.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/safety11010004/s1, Detailed Description of SAFTE-FAST Insights Functionality; Supplementary: Example of the Insights Dashboard; References [10,13–15].

Author Contributions: Conceptualization, J.K.D. and S.R.H., methodology, J.K.D., J.C. and S.R.H.; formal analysis, J.K.D., J.C. and S.R.H.; data curation, J.K.D. and J.C.; writing—original draft preparation, J.K.D.; writing—review and editing, J.K.D., J.C. and S.R.H.; supervision, S.R.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and was deemed non-human subjects research by Salus IRB on 18 September 2024 (Study ID: 24595).

Informed Consent Statement: Not applicable.

Data Availability Statement: Data for this analysis are unavailable due to the proprietary nature of the airline's scheduling practices.

Acknowledgments: The authors would like to acknowledge and thank the airline and human factors team for contributing data and facilitating this analysis.

Conflicts of Interest: The Institutes for Behavior Resources provides sales of SAFTE-FAST. Authors J. K. Devine, J. Choynowski, and S. R. Hursh are affiliated with the Institutes for Behavior Resources. Author S. R. Hursh is the inventor of the SAFTE-FAST biomathematical model, and a fraction of his compensation is based on sales of the software.

References

1. Marcus, J.H.; Rosekind, M.R. Fatigue in transportation: NTSB investigations and safety recommendations. *Inj. Prev.* **2017**, *23*, 232–238. [CrossRef] [PubMed]
2. Choi, J.-K. The Improvement of Pilot Fatigue Management. *Korean J. Aerosp. Environ. Med.* **2023**, *33*, 27–31. [CrossRef]
3. *FAA-2009-1093*; Flightcrew Member Duty and Rest Requirements. U.S. Department of Transportation, Federal Aviation Administration: Washington, DC, USA, 2012. Available online: https://www.federalregister.gov/documents/2012/01/04/2011-33078/flightcrew-member-duty-and-rest-requirements (accessed on 30 December 2024).
4. International Federation of Air Line Pilots' Associations. Fatigue Reporting Guidance. Available online: https://www.ifalpa.org/media/3844/22hupbl01-ifalpa-fatigue-reporting-guidance.pdf (accessed on 21 December 2024).
5. *120-103A*; Fatigue Risk Management Systems for Aviation Safety. U.S. Department of Transportation, Federal Aviation Administration: Washington, DC, USA, 2013. Available online: https://www.faa.gov/documentlibrary/media/advisory_circular/ac_120-103a.pdf (accessed on 21 December 2024).
6. Hursh, S.R.; Van Dongen, H.P. Fatigue and Performance Modeling. In *Principles and Practice of Sleep Medicine*, 5th ed.; Belenky, G., Ed.; Elsevier Inc.: Amsterdam, The Netherlands, 2010; pp. 745–753.
7. SAFTE-FAST as a Supporting Tool for Fatigue Investigation. Available online: https://www.saftefast.com/_files/ugd/c8faa9_7f7b509fb18940f5ab04f2cc2c46a6e3.pdf (accessed on 21 December 2024).
8. Pruchnicki, S.A.; Wu, L.J.; Belenky, G. An exploration of the utility of mathematical modeling predicting fatigue from sleep/wake history and circadian phase applied in accident analysis and prevention: The crash of Comair Flight 5191. *Accid. Anal. Prev.* **2011**, *43*, 1056–1061. [CrossRef] [PubMed]
9. Hursh, S.R.; Raslear, T.G.; Kaye, A.S.; Fanzone, J. *Validation and Calibration of a Fatigue Assessment Tool for Railroad Work Schedules, Summary Report*; U.S. Department of Transportation, Federal Railroad Administration: Washington, DC, USA, 2006.
10. Hursh, S.R.; Redmond, D.P.; Johnson, M.L.; Thorne, D.R.; Belenky, G.; Balkin, T.J.; Storm, W.F.; Miller, J.C.; Eddy, D.R. Fatigue models for applied research in warfighting. *Aviat. Space Environ. Med.* **2004**, *75*, A44–A53. [PubMed]
11. Macmillan, N.A. Signal detection theory. In *Stevens' Handbook of Experimental Psychology*; Wiley: Hoboken, NJ, USA, 2002; Volume 4, pp. 43–90.
12. Hursh, S.R. Insights into What Is Driving Fatigue in Your Schedule Construction. Available online: https://www.saftefast.com/safte-fast-blog/insights-into-what-is-driving-fatigue-in-your-schedule-construction#viewer-d90t8 (accessed on 15 September 2024).
13. Paul, M.A.; Hursh, S.R.; Love, R.J. The importance of validating sleep behavior models for fatigue management software in military Aviation. *Mil. Med.* **2020**, *185*, e1986–e1991. [CrossRef] [PubMed]
14. Devine, J.K.; Garcia, C.R.; Simoes, A.S.; Guelere, M.R.; de Godoy, B.; Silva, D.S.; Pacheco, P.C.; Choynowski, J.; Hursh, S.R. Predictive biomathematical modeling compared to objective sleep during COVID-19 humanitarian flights. *Aerosp. Med. Hum. Perform.* **2022**, *93*, 4–12. [CrossRef] [PubMed]
15. Rangan, S.; Van Dongen, H. Quantifying fatigue risk in model-based fatigue risk management. *Aviat. Space Environ. Med.* **2013**, *84*, 155–157. [CrossRef] [PubMed]
16. Dawson, D.; Reid, K. Fatigue, alcohol and performance impairment. *Nature* **1997**, *388*, 235. [CrossRef] [PubMed]
17. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology*; Elsevier: Amsterdam, The Netherlands, 1988; Volume 52, pp. 139–183.
18. SAFTE-FAST Workload Calculator. 2020. Available online: https://www.saftefast.com/_files/ugd/c8faa9_ef6b7cfa65a64c39874778228f3bb73f.pdf (accessed on 15 September 2024).
19. Ordak, M. Multiple comparisons and effect size: Statistical recommendations for authors planning to submit an article to Allergy. *Allergy* **2023**, *78*, 1145. [CrossRef] [PubMed]
20. Ashour, L. A review of user-friendly freely-available statistical analysis software for medical researchers and biostatisticians. *Res. Stat.* **2024**, *2*, 2322630. [CrossRef]

21. Ateş, S.S.; Kafalı, H.; Çeliktaş, M. Analysis of internal/external factors affecting time management and the reasons of delay in aviation. *J. Aviat.* **2018**, *2*, 105–118. [CrossRef]
22. Deshpande, V.; Arıkan, M. The impact of airline flight schedules on flight delays. *Manuf. Serv. Oper. Manag.* **2012**, *14*, 423–440. [CrossRef]
23. Masi, G.; Amprimo, G.; Ferraris, C.; Priano, L. Stress and workload assessment in aviation—A narrative review. *Sensors* **2023**, *23*, 3556. [CrossRef] [PubMed]
24. Arsintescu, L.; Chachad, R.; Gregory, K.B.; Mulligan, J.B.; Flynn-Evans, E.E. The relationship between workload, performance and fatigue in a short-haul airline. *Chronobiol. Int.* **2020**, *37*, 1492–1494. [CrossRef] [PubMed]
25. Mallis, M.M.; Mejdal, S.; Nguyen, T.T.; Dinges, D.F. Summary of the key features of seven biomathematical models of human fatigue and performance. *Aviat. Space Environ. Med.* **2004**, *75*, A4–A14. [PubMed]
26. Mélan, C.; Cascino, N. A multidisciplinary approach of workload assessment in real-job situations: Investigation in the field of aerospace activities. *Front. Psychol.* **2014**, *5*, 964.
27. Baum, N. Solving the Wrong Problem. *J. Med. Pract. Manag.* **2022**, *37*, 183–185.