Predictive Biomathematical Modeling Compared to Objective Sleep During COVID-19 Humanitarian Flights

Jaime K. Devine; Caio R. Garcia; Audrey S. Simoes; Marina R. Guelere; Bruno de Godoy; Diego S. Silva; Philipe C. Pacheco; Jake Choynowski; Steven R. Hursh

BACKGROUND: Biomathematical modeling software like the Sleep, Activity, Fatigue, and Task Effectiveness (SAFTE) model and Fatigue Avoidance Scheduling Tool (FAST) help carriers predict fatigue risk for planned rosters. The ability of a biomathematical model to accurately estimate fatigue risk during unprecedented operations, such as COVID-19 humanitarian ultra-long-range flights, is unknown. Azul Cargo Express organized and conducted five separate humanitarian missions to China between May and July 2020. Prior to conducting the missions, a sleep-prediction algorithm (AutoSleep) within SAFTE-FAST was used to predict in-flight sleep duration and pilot effectiveness. Here we compare AutoSleep predictions against pilots' sleep diary and a sleep-tracking actigraphy device (Zulu watch, Institutes for Behavior Resources) from Azul's COVID-19 humanitarian missions.

METHODS: Pilots wore Zulu watches throughout the mission period and reported sleep duration for their in-flight rest periods using a sleep diary. Agreement between AutoSleep, diary, and Zulu watch measures was compared using intraclass correlation coefficients (ICC). Goodness-of-fit between predicted effectiveness distribution between scenarios was evaluated using the R^2 statistic.

RESULTS: A total of 20 ($N = 20$) pilots flying across 5 humanitarian missions provided sleep diary and actigraphy data. ICC and R^2 values were >0.90, indicating excellent agreement between sleep measures and predicted effectiveness distribution, respectively.

DISCUSSION: Biomathematical predictions of in-flight sleep during unprecedented humanitarian missions were in agreement with actual sleep patterns during flights. These findings indicate that biomathematical models may retain accuracy even under extreme circumstances. Pilots may overestimate the amount of sleep that they receive during extreme flight-duty periods, which could constitute a fatigue risk.

KEYWORDS: fatigue risk management, biomathematical modeling, COVID-19, ultra-long-range.

Devine JK, Garcia CR, Simoes AS, Guelere MR, de Godoy B, Silva DS, Pacheco PC, Choynowski J, Hursh SR. ***Predictive biomathematical modeling compared to objective sleep during COVID-19 humanitarian flights.*** **Aerosp Med Hum Perform. 2022; 93(1):4–12.**

In January 2020, the World Health Organization declared a public health emergency of international concern due to the spread of the Coronavirus Disease 2019 (COVID-19), caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2).[14] The outbreak of COVID-19 has forced many industries and governments around the world to adapt to an unprecedented global disaster. The commercial aviation industry in particular has been impacted by COVID-19.[2,17] Airlines had to reduce passenger flights due to global travel restrictions and decreased demand but have also been thrust into an essential humanitarian role to facilitate repatriation and cargo transport.[22]

One strategy that has been suggested for a post-COVID-19 era is the adoption of ultra-long-range (ULR) operations. Defined as flights which are longer than 14.5 h in length, ULR operations strive to connect cities in far-reaching corners of the globe.[2] In the context of COVID-19, ULR operations could help limit exposure since passengers would not have layovers at busy

From the Institutes for Behavior Resources, Baltimore, MD, USA; Azul Linhas Aéreas Brasileiras, Sao Paulo, Brazil; and Johns Hopkins University School of Medicine, Baltimore, MD, USA.

This manuscript was received for review in April 2021. It was accepted for publication in November 2021.

Address correspondence to: Jaime K. Devine, Ph.D., Institutes for Behavior Resources, Inc., 2104 Maryland Ave, Baltimore, MD 21218, USA; jdevine@ibrinc.org.

DOI: https://doi.org/10.3357/AMHP.5909.2022

international hubs and could maintain greater physical distance onboard due to the configuration of an aircraft designed for ULR service.[2]

While air operators need to adapt strategically to circumvent the dual threat COVID-19 poses to revenue and public health, operational safety risks which existed before 2020 have not disappeared and must be properly mitigated in any flight plans specific to COVID-19 or beyond. Pilot fatigue constitutes a well-acknowledged threat to aviation safety,[6,8,20] and regulatory bodies across the globe have imposed rules to prevent pilots from flying before they have had sufficient time to recover from fatigue associated with prolonged air travel, including the use of biomathematical models. Biomathematical models, such as the Sleep, Fatigue, and Task Effectiveness (SAFTE) model, are used to predict fatigue and optimize shift schedules in aviation[10,15] and can help forecast fatigue risk in COVID-19 or post-COVID-19 ULR operations.

The benefit of biomathematical models is that they allow for the identification of rotations, schedule patterns, and time periods with high fatigue risk and allow for targeted mitigation in those areas. Many models incorporate a sleep estimator which can predict sleep patterns during a proposed schedule accurately compared to individuals' actual sleep.[3,15] However, sleep estimators are parameterized based on previously collected sleep data and the generalizability of sleep predictions across operations is unknown.[11,19] The ability of a sleep estimator to accurately predict sleep patterns during COVID-19-related ULR flight operations has not been previously investigated.

Airlines were not afforded the luxury of a priori comparison of sleep estimators to actual sleep patterns when executing COVID-19 pandemic-related operations, but could collect sleep data during operations to allow for post hoc analysis. This was the case with Brazil-based Azul Airlines. The first confirmed case of COVID-19 in Brazil occurred in February 2020, but there were over 177,000 cases across the nation by mid-May.[7] Azul Airlines Human Factors team used the SAFTE model's Fatigue Avoidance Scheduling Tool (FAST) to forecast an estimate of pilot fatigue and effectiveness using the software's sleep estimator, known as AutoSleep, prior to conducting five separate humanitarian missions to China between May and July 2020.

The purpose of Azul's humanitarian missions was to bring respirators, COVID rapid tests, and medical supplies from mainland China back to Brazil. Azul Airlines had not previously conducted flights to China, and the pilots were unfamiliar with the destination airports within China. The round-trip flights required 30 h of in-flight travel. The flights were designed to be carried out with two relay crews consisting of four pilots each, for a total mission crew of eight pilots. The crews were organized so that all pilots would be available to work during any flight leg, that each pilot was afforded an in-flight sleep opportunity, and that none of the pilots would need to fly extra time.

During missions, pilots wore a validated wrist actigraph, the Zulu watch (Institutes for Behavior Resources, Baltimore, MD, USA),[4] and reported the sleep duration for their in-flight rest periods using a sleep diary. The goal of the current analysis was to compare the in-flight sleep patterns predicted by SAFTE-FAST's AutoSleep feature against pilots' self-report sleep diary and compare AutoSleep-predicted sleep to objective sleep duration as measured by actigraphy during the airline's COVID-19 humanitarian missions.

METHODS

Subjects

Subjects were recruited through Azul Airlines Human Factors Safety Department. Subjects provided written informed consent for their participation. All mission crew were considered eligible for inclusion regardless of gender, ethnicity, age (over 18), sleep habits, or health status. Secondary use of de-identified data for research purposes was approved by the Salus Institutional Review Board and these analyses were conducted in accordance with the Declaration of Helsinki.

Materials

Biomathematical Modeling Software. 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. SAFTE-FAST scenarios are project files which visualize predicted performance metrics such as effectiveness and sleep reservoir against work and sleep events. AutoSleep 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. SAFTE-FAST is the brand name of fatigue risk management system software provided by the Institutes for Behavior Resources to a variety of operational organizations, including Azul Airlines.

Sleep Diary. Pilots were asked to complete a paper diary indicating when and for how long they slept in flight during flight duty periods (FDPs). Pilots were also asked to indicate the quality of their sleep. Paper diaries were returned to airline researchers at the completion of the mission.

Actigraph. The Zulu watch is a commercial actigraph that has been validated against polysomnography and another actigraphy device for sleep-wake determination.[4] The Zulu watch automatically determined sleep on-wrist using a proprietary sleep determination algorithm. Sleep data are scored in real time on the Zulu watch without researcher intervention. The most recent data is stored on the device as sleep events; the Zulu watch can store up to 80 sleep events while on-wrist. The Zulu watch has off-wrist detection to help differentiate between sleep periods and off-wrist periods, can detect multiple sleep episodes per day, and can detect naps as short as 20 min. More information about the Zulu watch device can be found in Devine et al.[4]

Procedure

A separate SAFTE-FAST project file was created for each humanitarian mission (Missions 1–5). Trip pairings were modeled in

SAFTE-FAST 4.0 using Azul's default setup package. Planned work sleep event rules (for example, in-flight sleep) were manually entered for each trip to reflect mission specifics. The flights were designed to be carried out with two relay crews consisting of four pilots each (eight pilots total). The crews were organized so that all pilots would be available to work during any flight leg and that no one pilot would need to fly extra time. Each mission consisted of four flight legs: 1) an outbound flight from Brazil to a layover destination in Europe; 2) a direct flight from Europe to China following the outbound layover; 3) a return flight to Europe from China; 4) a final flight from Europe to Brazil following a return layover. Each flight leg was approximately 12 h; each pilot was afforded a 9-h in-flight sleep opportunity during which they were not required to crew the aircraft. In-flight rest periods were freely chosen by the crew during the mission. Layovers in Europe were between 20 to 41 h long. Pilots disembarked during layovers and slept in hotel rooms. Turnaround in China ranged between 3 to 6 h while supplies were loaded onto the plane. No member of the mission crew deplaned during turnaround in China. Greater detail about pilot sleep patterns across the mission are described in Devine et al.[5]

AutoSleep predicted planned work sleep using an augmentation rule that prohibited work sleep events from occurring within 30 min of beginning an FDP or within 90 min of ending an FDP. One work sleep event per crewmember was assumed to occur during any given FDP. The length of the sleep event was computed by subtracting the time when sleep was prohibited (120 min) from the total flight duration and then dividing that time by two (two crews) and multiplying by 0.75 (75% credit for sleep during that opportunity) to reflect aviation guidelines for onboard rest.[16] No Event Rules were assumed, meaning that each Sleep Rule calculation to the specific flights was entered manually by airline researchers. Planned work sleep quality was set to "Good", assuming two interruptions per hour that each cost 5 min of sleep time, or 50 min of restorative sleep per hour. AutoSleep settings for layover periods were set to base time, per instructions to pilots that they should maintain a home base schedule, i.e., west Brazilian local time (UTC-5). Sleep at hotels during layovers was assumed to be "Excellent", assuming 60 min of restorative sleep per hour.

Pilots were assigned the Zulu watch (Institutes for Behavior Resources, Inc.) in May 2020 prior to COVID-19 support missions and wore the watches continuously until the completion of their mission (between May and July 2020). Crews returned the watch to airline researchers directly upon returning to Brazil from their mission. Data were downloaded by airline researchers using the Zulu Data Extraction application (Institutes for Behavior Resources, Version 2.0) and saved as .CSV files. Files reported all sleep interval start and end times and sleep duration in minutes. CSV files were compiled in Excel 2013 into a .CSV file which could be imported into SAFTE-FAST 4.0 and Stata MP15 (StataCorp, College Station, TX, USA). Files were manually inspected for data corruption prior to inclusion in the dataset. Sleep events were not edited for the purposes of these analyses.

Pilots also completed a sleep diary during FDPs. Pilots were not asked to complete the sleep diary during layovers or ground time in China. All times were reported in Brazilian time. Aircrew were instructed to remain on home base Brazilian time throughout the mission. The pilots reported the flight leg, duty start time, flight time, and the timing, duration, and quality of their preflight and in-flight sleep. Subjective sleep quality was rated on a 4-point scale as either Poor, Fair, Good, or Excellent by pilots. Diary sleep information was manually compiled into a .CSV file which could be imported into SAFTE-FAST 4.0. Sleep and performance metrics are summarized in **Table I**.

The original SAFTE-FAST models of COVID-19 humanitarian mission flights using AutoSleep were duplicated to produce two comparison scenarios: 1) explicit sleep scenarios based on Zulu watch objective sleep; and 2) explicit sleep scenarios based on the subjective sleep diaries to model sleep patterns during mission FDP and AutoSleep during layover periods. Both diary and Zulu watch sleep scenarios used sleep start time (indicating sleep onset) and sleep end time (indicating the time of final awakening) to report explicit sleep duration. The subjective sleep diary scenario adjusted Environment settings based on subjective sleep quality ratings, such that Excellent sleep assumed no interruptions, or 60 min of restorative sleep per hour, Good assumed two interruptions per hour or 50 min of restorative sleep per hour, Fair assumed four

Table I. Sleep and Performance Metrics.

METRIC NAME	DEFINITION
Sleep duration	Sleep duration is a measure of the amount of time that crew dedicated to sleep. SAFTE-FAST uses sleep duration from AutoSleep estimations or explicit sleep (diary or Zulu watch) to calculate Effectiveness. Sleep duration is expressed in minutes.
Sleep quality	The sleep diary allowed pilots to report their subjective sleep quality. Sleep quality is a measure of the crewmember's satisfaction with their sleep. Crew selected from four sleep quality categories based on their personal experience following the sleep event. The options were: Excellent, Good, Fair, and Poor.
Sleep environment	Sleep environment is a SAFTE-FAST setting option which refers to the potential for interruptions to sleep due to the quality of the environment. There are four sleep environment categories which can be selected based on knowledge of the sleep environment. Sleep environment quality uses the same labels as sleep quality but may not necessarily represent the crewmember's subjective sleep experience.
Crewing effectiveness	Crewing effectiveness is a performance measure computed by SAFTE-FAST for each minute during crewing events. Effectiveness is expressed as a percentage (%) scaled to a fully rested person's normal best performance. The higher the score, the lower the fatigue risk.
Effectiveness distribution	The distribution of 5% crewing effectiveness bins (e.g., 90–95% effectiveness) across all crewing event time (expressed as a percentage). Effectiveness distribution gives an overview of performance across the whole mission period.

Terms and definitions for measures of sleep and performance for the purposes of subsequent analyses.

interruptions per hour or 40 min of restorative sleep per hour, and Poor assumed six interruptions per hour or 30 min of restorative sleep per hour. Environment was kept at Excellent for all Zulu sleep data, since the watch considers sleep interruptions as awakenings (see **Fig. 1**) and adjusting the Environment setting would result in a falsely low sleep duration.

SAFTE-FAST estimated performance metrics based on the crewmembers' flight schedules and sleep input. Crew schedules were identical across all scenarios within a mission project. Therefore, differences in performance metrics are assumed to be due to differences in the reporting of sleep (i.e., AutoSleep vs. Diary vs. Zulu watch). An example of the three sleep scenarios for an individual pilot is depicted in Fig. 1.

As described in Table I, crewing effectiveness is a SAFTE-FAST output of performance estimation. In SAFTE-FAST, effectiveness represents speed of performance on the Psychomotor Vigilance Test, scaled as a percent (%) of a fully rested person's normal best performance. Effectiveness corresponds to reaction time speed, is highly sensitive to fatigue, and correlated with cognitive performance.[7,9] It is possible for effectiveness scores to be greater than 100%. Minimal effectiveness during critical phases of flight is considered to be 77%.[18,21] Crewing effectiveness refers to the effectiveness score for each work minute over the course of a crewing event. SAFTE-FAST commonly reports the distribution of effectiveness scores by 5% bins (e.g., 90–95%) for scores above 50%. Effectiveness scores below 50% are extremely rare during crewing events; for this reason, effectiveness scores in the range of 0–50% are collapsed into a single bin.

Statistical Analysis

Sleep and performances metrics were exported from SAFTE-FAST 4.0 as .CSV files. All data were subsequently analyzed using Excel 2013 and Stata MP 15. Sleep duration was compared between the three sleep scenarios (AutoSleep, Zulu, and Diary) for all mission flight legs using Stata MP 15.1 statistical analysis software. Sleep duration per 24-h period during layovers was additionally compared between the AutoSleep and

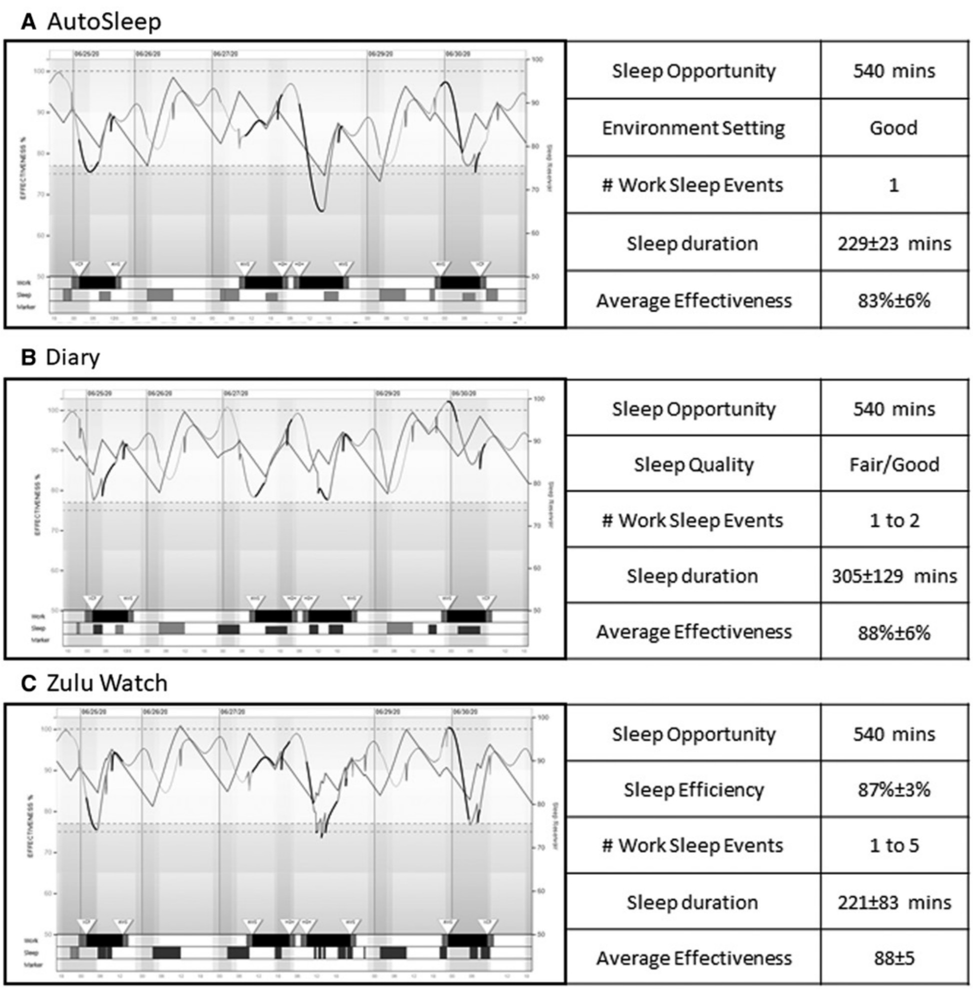

Fig. 1. Example comparison between AutoSleep, diary, and Zulu watch SAFTE-FAST scenarios. Example of SAFTE-FAST modeled performance using: A) AutoSleep; B) sleep diary; or C) Zulu watch data for one pilot subject. Dates are listed at the top of the graph. Each flight leg is depicted as a black bar with white carets on the row marked "Work" which appears beneath the Effectiveness % graph. Sleep events are graphed on the left axis below Effectiveness % and Work events. A) AutoSleep (light gray bars) predicted 1 work sleep event with good sleep quality per flight leg. B) Diary times were imported as explicit sleep (dark gray bars). AutoSleep (light gray) was used to estimate sleep between flight legs since diaries were only completed for in-flight sleep. C) Zulu watch data was imported as explicit sleep (dark gray bars) for the entire mission.

Zulu scenarios. Statistical significance was assumed at $P < 0.05$. Differences in sleep duration between AutoSleep and sleep diary, AutoSleep and Zulu watch, and diary and Zulu watch measurements were explored using paired samples *t*-tests.

Agreement between all three measures of in-flight sleep duration was furthermore evaluated using single rater, two-way random effects intraclass correlation coefficients (ICC) with absolute agreement. ICCs were computed for any flight leg for which all three measures (AutoSleep, diary, and Zulu watch) were collected. ICC values were classified as poor (< 0.50), moderate (0.50–0.75), good (0.75–0.90), or excellent (> 0.90) based on established guidelines.[12]

Minimum, maximum, and average crewing effectiveness scores between AutoSleep and sleep diary, AutoSleep and Zulu watch, and diary and Zulu watch scenarios were compared using paired samples *t*-tests. The distribution of crewing effectiveness by 5% effectiveness bins were combined in Excel 2013 to examine effectiveness distribution across all missions. Linear regressions examined the difference between AutoSleep predictions of effectiveness distribution compared to sleep diary and Zulu watches across all missions. Goodness of fit was evaluated using the R^2 statistic. An R^2 value of 0.5 means that half of the variance in the outcome variable is explained by the model.[23]

RESULTS

Each of Azul's five humanitarian missions had two relay crews of four pilots, for a total of eight pilots per mission. In total, 40 pilots flew between Brazil and China between May and July 2020 for the Azul's humanitarian missions. Out of 40 pilots crewing a COVID-19 humanitarian mission, 32 (80%) completed the sleep diary and 22 out of 40 (55%) wore a Zulu watch between May and June 2021. There were 20 (50%) pilots who completed both the sleep diary and the Zulu watch. Only pilots who both completed the sleep diary and provided Zulu watch data ($N = 20$) have been included in these analyses. A total of $N = 15$ subjects provided Zulu and diary data for all 4 flight legs; $N = 3$ subjects provided Zulu and sleep diary data for 3 out of the 4 flight legs; $N = 1$ subject provided Zulu watch data for all flight legs but only completed the sleep diary for 3 out of 4 flight legs; and $N = 1$ subject completed the sleep diary for all 4 flight legs, but only wore the Zulu watch for 3 out of the 4 flight legs. In total, 77 observations of in-flight sleep opportunities by AutoSleep, sleep diary, and Zulu watch were compiled from all 4 flight legs across all 5 missions and 20 subjects for subsequent analysis.

Each flight leg was approximately 12 h and the planned available rest time for each crewmember per stage was approximately 9 h. In-flight rest periods were freely chosen by the crew during the mission. Crew slept either in crew rest facilities or in the business class section, per their preference. Pilots slept between 0 min and 9 h and 55 min during mission flight legs. There was only one instance in which a pilot did not report any sleep and no sleep event was recorded by the Zulu watch during a flight leg. For flight legs during which sleep occurred, the minimum sleep duration was 22 min.

AutoSleep predicted 235 ± 20 min of sleep per flight leg, compared to the 325 ± 128 min per flight leg reported by sleep diary, or the 246 ± 132 min per flight leg recorded by Zulu watches. In-flight rest opportunities were determined ad libitum during flights. This planned flexibility in sleep timing precluded any comparisons between AutoSleep estimations of sleep timing and observations of actual sleep onset or awakening. In-depth descriptions of sleep patterns during mission flights are reported elsewhere.[5]

Comparison of average in-flight sleep duration per flight leg is depicted in **Fig. 2A**. Paired samples *t*-tests showed that

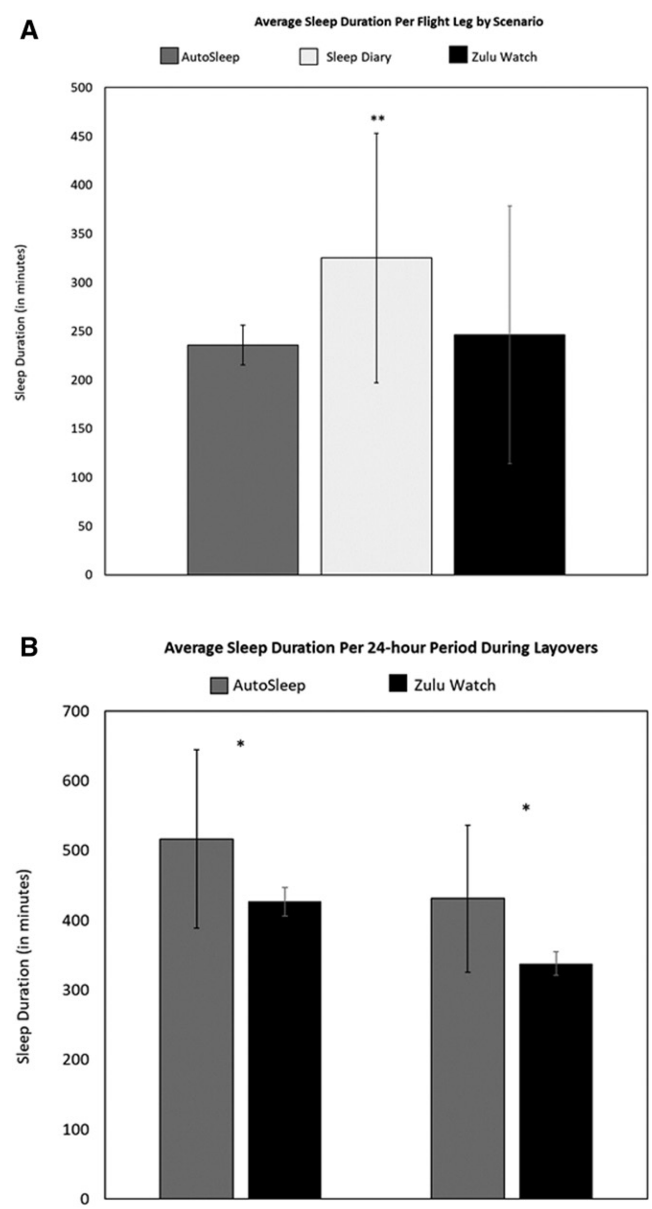

Fig. 2. In-flight sleep duration during COVID-19 humanitarian missions. A) Average sleep duration across all mission flight legs as predicted by AutoSleep (gray), or measured by sleep diary (light gray), and Zulu watch (black). B) Average sleep duration per 24-h period during layover periods as predicted by AutoSleep (gray) or measured by Zulu watch (black). Significance is indicated by an asterisk (*) at $P \leq 0.05$ and a double asterisk (**) at $P \leq 0.001$.

diary reports of sleep duration were significantly higher than AutoSleep predictions ($t = 6.05$, df = 151, $P \leq 0.001$) or Zulu watch sleep duration ($t = 3.73$, df = 150, $P \leq 0.001$). AutoSleep predictions of sleep duration were not significantly different from Zulu watch sleep duration during FDPs ($t = 0.69$, df = 151, $P = 0.48$).

There were two layovers in Europe during each mission—an outbound layover and a return layover. Layovers were on average 1853 ± 473 min long, ranging between 1186–2405 min. A comparison of AutoSleep predictions of sleep duration during layovers against Zulu watch measurements of sleep during layovers is depicted in **Fig. 2B**. AutoSleep predicted an average sleep duration of 475 ± 48 min per 24-h period during layovers. Paired samples t-tests showed that AutoSleep predictions of sleep duration per 24-h period during layovers was significantly higher than Zulu-recorded sleep duration for both outbound (516 ± 18 min vs. 426 ± 1146 min; $t = 2.51$, df = 17, $P = 0.02$) and return layovers (431 ± 15 min vs. 338 ± 112 min; $t = 3.61$, df = 18, $P = 0.002$). Similar analyses including sleep diary information could not be performed since pilots were not required to report their sleep during layover periods. AutoSleep was used to fill in gaps about sleep duration for the Sleep Diary scenario.

ICCs examining agreement between biomathematical prediction of sleep (AutoSleep), subjective report of sleep (diary), and objective measurement of sleep (Zulu watch) during COVID-19 mission flights indicated excellent agreement between all measures for sleep duration (ICC = 0.94, 95% CI = 0.79–0.99, $P < 0.001$). AutoSleep predicted an average crewing effectiveness of 84% ± 6% (range: 66–102%) across all mission flight legs. Sleep diary and Zulu watch estimated average crewing effectiveness to be average: 86% ± 6%, range: 58–104%, and average: 84% ± 7%, range: 50–101%, respectively. There were no significant differences between AutoSleep and diary for minimum ($t = 1.54$, df = 152, $P = 0.13$), maximum ($t = 1.29$, df = 152, $P = 0.20$), or average ($t = 1.18$, df = 152, $P = 0.24$) effectiveness scores during crewing events. There were no significant differences for minimum ($t = 0.99$, df = 152, $P = 0.32$), maximum ($t = 0.27$, df = 152, $P = 0.79$), or average ($t = 2.02$, df = 152, $P = 0.05$) effectiveness scores during crewing events between sleep diary and Zulu watch. Maximum ($t = 1.59$, df = 152, $P = 0.11$) and average ($t = 0.85$, df = 152, $P = 0.40$) effectiveness scores were not significantly different between AutoSleep and Zulu watch, but minimum crewing effectiveness for Zulu watch scenarios (average minimum crewing effectiveness score: 74% ± 8%) were statistically lower than AutoSleep predictions (77% ± 6%; $t = 2.44$, df = 152, $P = 0.02$). The linear regressions for effectiveness distribution by 5% bins across all missions are shown in **Fig. 3A**. R^2 values were greater than 0.9 when compared against sleep diary (**Fig. 3B**; y = 0.93 + 0.67, $R^2 = 0.94$) or Zulu watch (**Fig. 3C**; y = 0.84 + 1.44, $R^2 = 0.90$).

DISCUSSION

The COVID-19 pandemic crisis disrupted almost every facet of modern society, but particularly the aviation industry faced unprecedented changes to daily operations. The purpose of SAFTE-FAST fatigue modeling software is to provide a science-based process for the analysis of on-the-job fatigue risk, including the assessment of suitable pairings and planned rosters to ensure safety during flights. Azul Airlines tested the limits of their pilots' capabilities and the capabilities of SAFTE-FAST to conduct five unprecedented humanitarian missions during an international crisis. Sleep during the 30+ hour flights were prospectively modeled using the SAFTE-FAST 4.0 AutoSleep function. This was the first time AutoSleep was applied to a 30+-h FDP.

While SAFTE-FAST does provide a service through which AutoSleep parameters can be harmonized to previously collected objective measures of sleep, COVID-19 circumstances did not allow for prospective harmonization. As such, flight scenarios were modeled using Azul's default settings for AutoSleep with additional input from the airline's human factor and safety researchers. The aim of this paper was to evaluate the ability of the SAFTE-FAST 4.0 AutoSleep function to predict pilot sleep duration throughout the missions compared against subjective (sleep diary) and objective (Zulu watch) measures of sleep.

Pilots reported longer sleep durations during flights on their sleep diaries compared to AutoSleep predictions or Zulu watch measurements (Fig. 2A). This finding supports previous evidence that individuals overestimate sleep duration by self-report.[1,13] It should be noted that the Zulu watch does not take into account sleep onset latency (i.e., the time it takes to first fall asleep) and treats periods of sustained wakefulness as the termination of a sleep event. As illustrated in Fig. 1C, multiple Zulu sleep events frequently were logged during the time period reported for one diary-reported sleep event. Those periods of wakefulness were not included in the Zulu watch estimation of sleep duration, but could have been reported as sleep by diary, which may contribute to the observed overestimation by self-report.

Interestingly, the overestimation of sleep duration by sleep diary did not result in major downstream effects with regards to effectiveness or goodness-of-fit. One reason why effectiveness distributions were similar despite differences in sleep duration as measured by self-report may be because the Sleep Environment parameter was adjusted based on pilots' diary reports of sleep quality. Sleep events with worse sleep quality ratings (e.g., Poor) were given less credit than sleep events with higher quality ratings (e.g., Excellent). The sleep diary scenario also used AutoSleep to fill in gaps in sleep reporting during layovers, which could be expected to produce similar results. The Zulu watch scenario did not employ the usage of any AutoSleep and Sleep Environment parameters were not adjusted by objective sleep quality. A comparison of sleep quality vs. Zulu watch measures of sleep efficiency or sleep depth can be found in Devine et al.[4] Improving biomathematical modeling of sleep quality is a future goal for the Institute for Behavior Resource's science and software development teams.

AutoSleep predictions of sleep were comparable to objectively measured sleep during in-flight periods, but not during

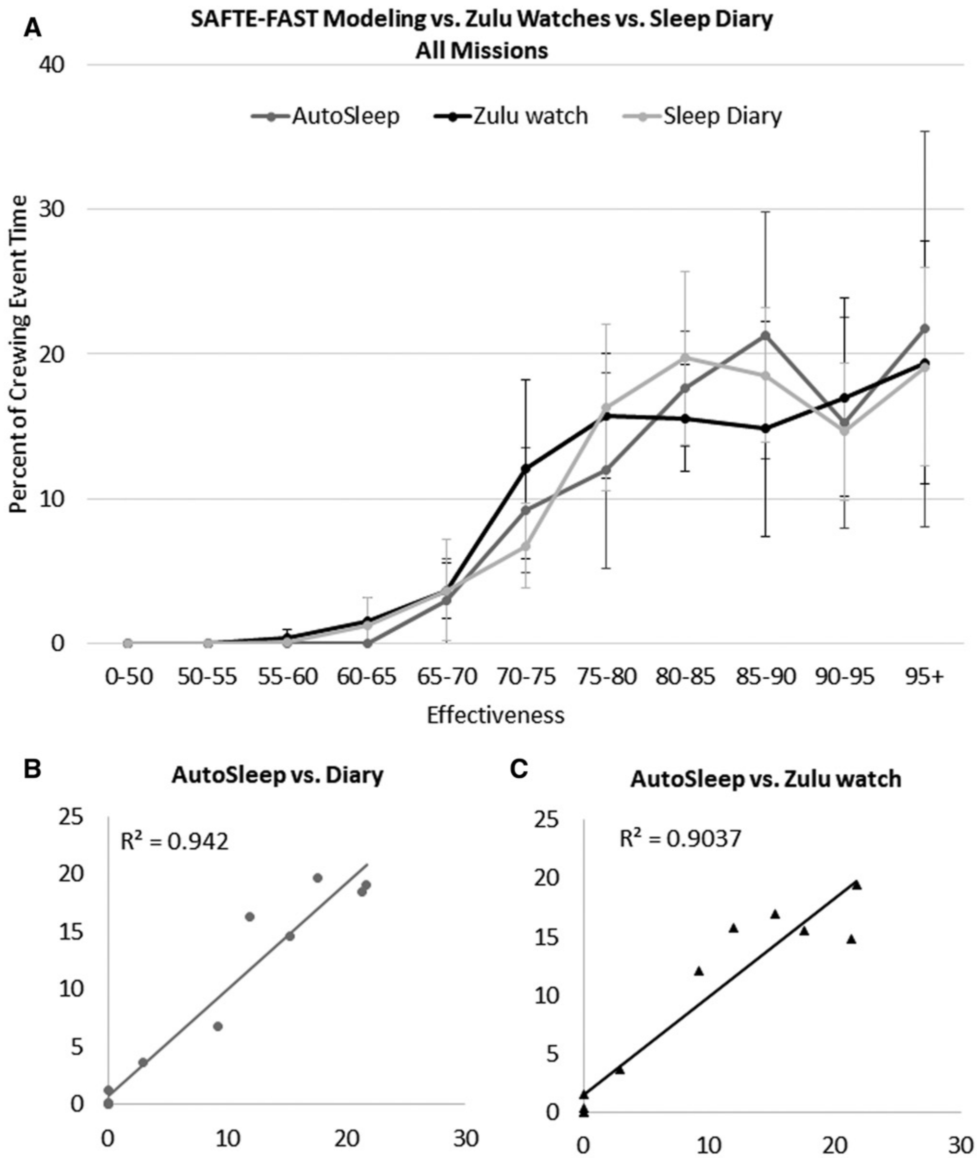

Fig. 3. AutoSleep effectiveness predictions during COVID-19 humanitarian missions compared to sleep diary and Zulu watch. A) AutoSleep predictions (in dark gray) of effectiveness distribution compared to diary (in light gray) or Zulu watch (in black) distributions by percent of crewing time. B & C) Linear regression and R^2 values indicating the goodness of fit for effectiveness distribution between B) AutoSleep and sleep diary and C) AutoSleep and Zulu watch. Higher R^2 values represent smaller differences between the model (in this case, AutoSleep) and the observed data (i.e., diary or Zulu watch). Error bars indicate individual variability in effectiveness distribution across all pilot subjects.

layovers. Pilots slept almost 2 h less during layovers than was predicted by AutoSleep, which may account for why SAFTE-FAST predictions of minimum effectiveness were lower in the Zulu watch scenario compared to the AutoSleep scenario. This finding highlights the importance of obtaining proper rest not only during active-duty periods, but during downtime as well. Foregoing sleep during layovers could constitute a fatigue risk during ULR. One possible explanation for this difference was that AutoSleep was set to predict sleep during layovers under the assumption that pilots retained a home base sleep schedule although pilots were in a different time zone. Even though pilots were instructed to operate on home base (UTC-5) time throughout the missions, being in a new time zone may have negatively impacted their sleep. AutoSleep can be set to predict sleep based on local time rather than home base time and could have yielded a more accurate estimation of sleep duration during layovers.

With the exception of the lowest possible effectiveness during a crewing event, crewing effectiveness was similar whether sleep was estimated by AutoSleep, reported via sleep diary, or collected objectively using the Zulu watch. This finding indicates that a sleep estimator within a biomathematical model can be applied to ULR operations. It is important to note that subjects in this study were not actively piloting during the entire FDP, but that the entire FDP was considered a crewing event. Mission crews decided rest periods and active piloting periods in flight rather than prospectively and, so, crewing periods could not be more acutely predicted. Zulu watch sleep events tended to occur in very close proximity to minimum effectiveness, which could indicate that pilots were resting during periods of reduced effectiveness. However, while pilots reported their sleep periods, they did not report the times during which they were flying the aircraft, so neither this example or other possible scenarios could be investigated. This constitutes an understandable limitation to the present analyses given the operational nature of the data collection. AutoSleep predicted sleep duration well given the limited knowledge of actual flight procedures, but an interesting follow-up study will be to examine the accuracy of the sleep estimator when the information about work constraints is more detailed. An additional follow-up will be to evaluate AutoSleep predictions using local time vs. base time assumptions about sleep.

Variability in actual sleep behavior is to be expected, and no model can perfectly account for this unpredictability. Notwithstanding that limitation, SAFTE-FAST 4.0 did an excellent job of forecasting sleep during ULR humanitarian flights. However, sleep is not the only human factor to consider when determining the applicability of ULR flights. The conditions of Azul Airlines' five humanitarian flights were considerably different (two relay crews of four pilots each; eight pilots total; no cabin crew or passengers) from normal commercial routes. Not only could the absence of passengers or other flight crew allow for better quality in-flight rest, but could impact the amount of workload or psychosocial stress imposed on the pilots. Additionally, events like delays, schedule changes, or flight-duty extensions cannot be anticipated or modeled. No such events occurred during these missions, but could contribute to unforeseen fatigue risk in normal operations. Furthermore, pilot demographics, like age, gender, or years of flight experience, were not collected. These variables may have had an impact on performance or sleep behavior, but could not be investigated in this dataset. Importantly, the humanitarian purpose of Azul's mission motivated these crews to fly 30+-h FDPs. Motivation is an important and often overlooked human factor to consider when assessing the safety of ULR rosters.

In conclusion, Bauer et al. recently suggested that a business model of point-to-point ULR services would allow airlines to successfully adjust to post-COVID-19 operations.[2] However, it is necessary to evaluate ULR flights in terms of human factors as well as economic factors. The human factor considered in these analyses is sleep, but credit must be given to Azul Airline's Human Factors Safety Department for a more wholistic consideration of flight logistics than is reported in this article. Azul Airlines tested the limits of their pilots' capabilities and the capabilities of SAFTE-FAST to conduct five unprecedented ULR humanitarian missions during an international crisis. The ability of the SAFTE-FAST AutoSleep function to accurately predict actual pilot rest patterns during the humanitarian missions is impressive, but not nearly as impressive as the efforts put forth by Azul's human factors team or the dedication of Azul's pilots to plan and execute these missions.

ACKNOWLEDGMENTS

The authors would like to acknowledge and thank Azul's human factors team and the pilots who flew the COVID-19 humanitarian flights.

Financial Disclosure Statement: The authors declare no direct conflicts of interest. The Institutes for Behavior Resources provides sales of SAFTE-FAST, software that uses a biomathematical model of fatigue, and sells the Zulu watch as a research tool. Authors J. K. Devine, J. Choynowski, and S. R. Hursh are affiliated with the Institutes for Behavior Resources, but do not benefit financially or non-financially from sale of the Zulu watch. 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.

Authors and Affiliations: Jaime K. Devine, Ph.D., M.S., Jake Choynowski, B.S., AA, and Steven R. Hursh, Ph.D., B.A., Institutes for Behavior Resources, Baltimore, MD; Caio R. Garcia, M.B.A., B.A., Audrey S. Simoes, B.A., Marina R. Guelere, B.A., Bruno de Godoy, M.B.A., M.A., Diego S. Silva, B.A., and Philipe C. Pacheco, B.A., Azul Linhas Aéreas Brasileiras, Sao Paulo, Brazil; and Steven R. Hursh, Johns Hopkins University School of Medicine, Baltimore, MD.

REFERENCES

1. Arsintescu L, Kato KH, Hilditch CJ, Gregory KB, Flynn-Evans E. Collecting Sleep, Circadian, Fatigue, and Performance Data in Complex Operational Environments. J Vis Exp. 2019; 2019(150).
2. Bauer LB, Bloch D, Merkert R. Ultra long-haul: an emerging business model accelerated by COVID-19. J Air Transp Manag. 2020; 89:101901.
3. Darwent D, Dawson D, Roach GD. Prediction of probabilistic sleep distributions following travel across multiple time zones. Sleep. 2010; 33(2):185–195.
4. Devine JK, Chinoy ED, Markwald RR, Schwartz LP, Hursh SR. Validation of Zulu Watch against polysomnography and actigraphy for on-wrist sleep-wake determination and sleep-depth estimation. Sensors (Basel). 2020; 21(1):76.
5. Devine JK, Choynowski J, Garcia CR, Simoes AS, Guelere MR, et al. Pilot sleep behavior across time during ultra-long-range flights. Clocks Sleep. 2021; 3(4):515–527.
6. Gander PH, Mulrine HM, van den Berg MJ, Smith AA, Signal TL, et al. Effects of sleep/wake history and circadian phase on proposed pilot fatigue safety performance indicators. J Sleep Res. 2015; 24(1):110–119.
7. Gertler J, Hursh S, Fanzone J, Raslear T, America QN. Validation of FAST model sleep estimates with actigraph measured sleep in locomotive engineers. Washington (DC):Federal Railroad Administration; 2012.
8. Holmes A, Al-Bayat S, Hilditch C, Bourgeois-Bougrine S. Sleep and sleepiness during an ultra long-range flight operation between the Middle East and United States. Accid Anal Prev. 2012; 45(Suppl.):27–31.
9. Hursh SR, Raslear TG, Kaye AS, Fanzone Jr JF.Validation and calibration of a fatigue assessment tool for railroad work schedules—summary report. Washington (DC): Federal Railroad Administration; 2006.
10. Hursh SR, Redmond DP, Johnson ML, Thorne DR, Belenky G, et al. Fatigue models for applied research in warfighting. Aviat Space Environ Med. 2004; 75(3, Suppl.):A44–A53, discussion A54–A60.
11. Hursh S, Waggoner L.Refining sleep predictions using actigraphy in operational environments: studies with pilots. 10th International Conference on Managing Fatigue Conference. [Accessed 10 Nov. 2021]. Available from https://fatigueconference2017.com/materials/wednesday-am/modeling/Abstract_Hursh.pdf.
12. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016; 15(2):155–163. Erratum in: J Chiropr Med. 2017; 16(4):346.
13. Lauderdale DS, Knutson KL, Yan LL, Liu K, Rathouz PJ. Self-reported and measured sleep duration: how similar are they? Epidemiology. 2008; 19(6):838–845.

14. Listings of WHO's response to COVID-19. 29 June 2020. [Accessed February 16, 2021]. Available from: https://www.who.int/news/item/29-06-2020-covidtimeline.
15. Mallis MM, Mejdal S, Nguyen TT, Dinges DF. Summary of the key features of seven biomathematical models of human fatigue and performance. Aviat Space Environ Med. 2004; 75(3, Suppl.):A4–A14.
16. Management of aviation fatigue: guidance evaluation and qualification of onboard flightcrew member rest facilities for Part 117 operations. General Technical Administration, Volume 3; 11/8/2013. [Accessed Aug. 3, 2021]. Available from: https://fsims.faa.gov/WDocs/8900.1/V03%20Tech%20Admin/Chapter%2058/03_058_003_CHG_304A.htm.
17. Pearce B. Challenging outlook for airlines, despite vaccine progress. COVID-19. [Accessed Nov. 12, 2021]. Available from https://www.iata.org/en/iata-repository/publications/economic-reports/challenging-outlook-for-airlines-despite-vaccine-progress/.
18. Rangan S, Van Dongen H. Quantifying fatigue risk in model-based fatigue risk management. Aviat Space Environ Med. 2013; 84(2): 155–157.
19. Riedy SM, Fekedulegn D, Andrew M, Vila B, Dawson D, Violanti J. Generalizability of a biomathematical model of fatigue's sleep predictions. Chronobiol Int. 2020; 37(4):564–572.
20. Roach GD, Darwent D, Dawson D. How well do pilots sleep during long-haul flights? Ergonomics. 2010; 53(9):1072–1075.
21. Rodrigues TE, Fischer FM, Bastos EM, Baia L, Bocces R, et al. Seasonal variation in fatigue indicators in Brazilian civil aviation crew rosters. Rev Bras Med Trab. 2020; 18(1):2–10.
22. Serrano F, Kazda A. The future of airport post COVID-19. J Air Transp Manag. 2020; 89:101900.
23. Xu R. Measuring explained variation in linear mixed effects models. Stat Med. 2003; 22(22):3527–3541.