




# Crew Autonomy During Simulated Medical Event Management on Long Duration Space Exploration Missions

**Steven Yule, PhD** , STRATUS Center for Medical Simulation, Brigham and Women's Hospital, Boston, MA, USA; Center for Surgery & Public Health, Brigham & Women's Hospital, Boston, MA, USA; Department of Surgery, Brigham & Women's Hospital/Harvard Medical School, Boston, MA, USA; Department of Clinical Surgery, The University of Edinburgh, Edinburgh, UK, **Jamie M. Robertson, PhD, MPH**, Department of Surgery, Brigham & Women's Hospital/Harvard Medical School, Boston, MA, USA, **Benjamin Mormann, MD**, Department of Emergency Medicine, Harvard Medical School, Boston, MA, USA, **Douglas S. Smink, PhD, MPH**, Center for Surgery & Public Health, Brigham & Women's Hospital, Boston, MA, USA; Department of Surgery, Brigham & Women's Hospital/Harvard Medical School, Boston, MA, USA, **Stuart Lipsitz, ScD**, Center for Surgery & Public Health, Brigham & Women's Hospital, Boston, MA, USA, **Egide Abahuje, MD, MMED**, STRATUS Center for Medical Simulation, Brigham and Women's Hospital, Boston, MA, USA; Department of Surgery, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA, **Lauren Kennedy-Metz, PhD** , Department of Surgery, Brigham & Women's Hospital/Harvard Medical School, Boston, MA, USA; Medical Robotics and Computer Assisted Surgery Laboratory, Division of Cardiac Surgery, U.S. Veterans Affairs Boston Healthcare System, Boston, MA, USA, **Sandra Park, MD** and **Christian Miccile, MSc**, STRATUS Center for Medical Simulation, Brigham and Women's Hospital, Boston, MA, USA, **Charles N. Pozner, MD**, STRATUS Center for Medical Simulation, Brigham and Women's Hospital, Boston, MA, USA; Department of Emergency Medicine, Harvard Medical School, Boston, MA, USA, **Thomas Doyle, MD, PhD**, Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, **David Musson, PhD**, Faculty of Health Science, Northern Ontario School of Medicine, Thunder Bay, ON, Canada, **Roger D. Dias, MD, PhD, MBA** , STRATUS Center for Medical Simulation, Brigham and Women's Hospital, Boston, MA, USA; Department of Emergency Medicine, Harvard Medical School, Boston, MA, USA

---

Address correspondence to Steven Yule, Division of Clinical and Surgical Sciences, University of Edinburgh, Chancellor's Building, Little France, Edinburgh EH16 4SA, UK; e-mail: [steven.yule@ed.ac.uk](mailto:steven.yule@ed.ac.uk)

**CORRECTION (May 2022):** Article originally published with "Research Article in Revision for Human Factors Special Issue" incorrectly preceding the title; this has been removed from the article title.

## **HUMAN FACTORS**

2023, Vol. 65(6) 1221–1234

DOI:10.1177/00187208211067575

Article reuse guidelines: [sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

Copyright © 2022, The Author(s).



**Objective:** Our primary aim was to investigate crew performance during medical emergencies with and without ground-support from a flight surgeon located at mission control.

**Background:** There are gaps in knowledge regarding the potential for unanticipated in-flight medical events to affect crew health and capacity, and potentially compromise mission success. Additionally, ground support may be impaired or periodically absent during long duration missions.

**Method:** We reviewed video recordings of 16 three-person flight crews each managing four unique medical events in a fully immersive spacecraft simulator. Crews were randomized to two conditions: with and without telemedical flight surgeon (FS) support. We assessed differences in technical performance, behavioral skills, and cognitive load between groups.

**Results:** Crews with FS support performed better clinically, were rated higher on technical skills, and completed more clinical

tasks from the medical checklists than crews without FS support. Crews with FS support also had better behavioral/non-technical skills (information exchange) and reported significantly lower cognitive demand during the medical event scenarios on the NASA-TLX scale, particularly in mental demand and temporal demand. There was no significant difference between groups in time to treat or in objective measures of cognitive demand derived from heart rate variability and electroencephalography.

**Conclusion:** Medical checklists are necessary but not sufficient to support high levels of autonomous crew performance in the absence of real-time flight surgeon support

**Application:** Potential applications of this research include developing ground-based and in-flight training countermeasures; informing policy regarding autonomous spaceflight, and design of autonomous clinical decision support systems.

**Keywords:** aerospace medicine, long-term missions, pilot, crew behavior, mental workload, medical simulation/training and assessment, simulation and training

## INTRODUCTION

Spaceflight operations are characterized as a Multi-Team System (MTS), defined as “two or more teams that interface directly and interdependently in response to environmental contingencies toward the accomplishment of collective goals” (p. 290) (Mathieu et al., 2001). The spaceflight MTS is a network that works toward a shared goal of mission success. It comprises a Mission Control Center (MCC) that coordinates communication between diverse specialized ground teams and the flight crew via the Capsule Communicator (CapCom). Operating at a level of analysis above individuals and groups, the spaceflight MTS poses a range of unique challenges. Success requires coordinated efforts between multiple, previously unacquainted teams. Furthermore, the skill sets and expertise of those teams cross boundaries, and require new collaborative ways to address the challenges to mission success. Accordingly, effective coordination, team cohesion (Salas et al., 2015), situational leadership, and shared mental models (Shuffler et al., 2015) seem to be critical for success. Unfortunately, existing models of team performance may not capture the nuances of multi-level function (Hackman, 2003); however Zaccaro et al. proposed a typology suggesting that MTS cover compositional, linkage, and developmental attributes (Zaccaro et al., 2012). Furthermore, in

a comprehensive analysis of the literature comprising 47 existing studies, four essential criteria for successful MTS were identified: attitudes, behaviors, cognitions and performance (Shuffler et al., 2015).

Complicating coordination of the spaceflight MTS, future long duration exploration missions (LDEMs) will involve delays in signal transmission that will affect communications and data transfer. This will be especially problematic during emergent situations. As crews travel farther and farther from Earth, round-trip communication signal transmission times grow longer and longer. For cis-lunar and lunar missions, that delay is on the order of two to four seconds, though while minor, may still be problematic for verbal communication. However, for missions to Mars, delays are of a greater magnitude, and will significantly disrupt normal communications. These communication delays increase on the outbound leg of the mission as the spacecraft travels away from Earth, are of variable magnitude during the exploration or ground phase of these missions, and shorten during return to Earth. For an extended stay on the surface of Mars, the period of communication delay for a surface mission will vary between 20 and 40 min depending upon the relative positions of Earth and Mars in their respective orbits. The impact of time delay on data and verbal communications will mean that MTS sub-teams must operate in at least semi-autonomous modes during certain phases of Beyond Low Earth Orbit missions.

In-flight medical emergencies under time-delayed communication conditions represent a significant risk for mission failure by compromising crew health and capacity. In future crewed LDEMs, such as those to Mars, the possibility of returning to Earth, emergency medical evacuation to a terrestrial medical facility, or consulting in real time via long distance communications may be impractical (Komorowski et al., 2021). The unique blend of risks, demands, complexity, and novelty of medical emergencies in LDEMs provides an ideal lens through which to study MTS strengths and vulnerabilities as it provides the following constraints:

1. Crew members must coordinate themselves to deal with an ill-defined event

2. The crew cannot operate at full capacity because at least one member is incapacitated
3. The crew may lack comprehensively trained medical personnel
4. Other hazards for mission success such as fire, radiation, and decompression can coexist, and may assume higher priority
5. Medical equipment and supplies are limited due to mass and volume constraints of spaceflight.

The MTS will have to address all potential in-flight medical emergencies, potentially requiring personnel to serve on two teams and work with multiple ground teams. Under conditions of time delay, the crew will have to manage medical emergencies in an autonomous manner, and ground-based teams will need novel methods of keeping track of progress to maintain team situation awareness.

These unique constraints coupled with high consequences support a role for high-fidelity simulation and the use of cognitive aids (e.g. crisis checklists (Arriaga et al., 2013)) as methods to understand human interaction in complex systems. Previous work quantifying the impact of medical emergencies on mission success (Robertson et al., 2020) allows efforts to be focused on developing countermeasures against events that require concurrent management of risks to the spacecraft and to crewmembers (e.g. toxic exposure, fire), contrasting with medical conditions that produce individual-level risk (e.g. seizure, cardiac arrest). Managing these emergencies will require a coordinated response from multiple teams and represent ideal situations to examine inherent strengths and vulnerabilities of the spaceflight MTS, as well as the role of workload management and crew non-technical skills (Flin et al., 2015) in successful performance. A disproportionately high cognitive workload on one individual or team may negatively affect the entire MTS via increases in system delays and errors.

The cognitive and social skills that reinforce knowledge and expertise in high-demand workplaces are often characterized as “behavioral skills” or “non-technical skills” (Wood et al., 2017). They enable team members to exchange task-relevant information and share individual

situation awareness in order to generate team level, shared mental models of understanding that support task execution and early error detection (Musson & Helmreich, 2004). This has particular relevance for LDEMs where a crew of four or six astronauts from different cultures will be together for over a year and need to coordinate to solve problems autonomously, without input from mission control.

Behavior marker systems enable observers to quantify specific behavioral skills. These systems support crew recognition and differentiation between etiologies of medical emergencies, coordination of activities for first response, and creation of an event management plan. Evidence from acute medical emergencies in terrestrial environments show that every minute of delay in identification of the emergency impedes initiation of treatment and increases mortality risk by 14% (Yule & Walls, 2012). Such quantifiable predictions of the role of behavioral skills on outcomes of potential emergencies in space are currently lacking. We can, however, predict that the key factor that detracts from high quality response is not the medical expertise of the team, but rather their inability to organize themselves appropriately using behavioral skills to manage the medical emergency and achieve successful performance outcomes.

How best to cultivate behavioral skills for autonomous astronaut flight crews is not yet established, limiting available countermeasures for medical event management on LDEMs. In line with NASA’s Human Research roadmap and knowledge gaps related to “Risk of Adverse Health Outcomes & Decrements in Performance due to Inflight Medical Conditions” (Human Research Roadmap, n.d.), our long-term goal is to implement spaceflight medical simulation through human factors and cognitive engineering lenses to close knowledge gaps and develop countermeasures against the risks of medical events on LDEMs, to support astronauts, save lives, and promote mission success.

To address these gaps, we developed a fully immersive medical simulation incorporating aspects of the spaceflight MTS to study patterns of crew behavior previously identified as relevant for medical event management in a laboratory environment. We designed a randomized

controlled trial manipulating the role of flight surgeon telemedical support during simulated medical events. Our research question is: “what is the impact of not having support from a flight surgeon (at MCC) on crews’ technical and non-technical performance during in-flight medical event management?” We hypothesized that crews managing medical events autonomously, without support from a flight surgeon, would demonstrate worse performance (technical and non-technical) during simulated inflight clinical scenarios compared to crews with flight surgeon support (current gold standard).

## METHODS

### Study Design and Setting

This was a randomized controlled trial in which participants formed teams of three members whose task was to manage a series of four simulated in-flight medical emergencies. Using a random number generator, teams were randomly assigned to one of two groups: with support from a Flight Surgeon (with FS) and without support from a Flight Surgeon (without FS).

The spacecraft vehicle simulator and mission control console used in this study are located at the STRATUS Center for Medical Simulation at Brigham and Women’s Hospital in Boston, MA, USA. The simulation environment was designed in a previous study (Musson et al., n.d.) to reflect distributed communication nodes including MCC, flight surgeon, medical module, science module, and a simulation control unit (Figure 1). This asset is capable of testing ergonomic design, telemedical support systems, and medical event management training.

### Study Population

We recruited undergraduate and graduate students from engineering and science disciplines, paramedics, and emergency medicine technicians (EMTs) from the Boston, MA area. Research subjects were selected to reflect the diversity of the real astronaut crew, described by NASA as ‘astronaut-like participants’. For standardization, the same person (JR) played the role of the MCC Flight Surgeon. In each

scenario, a “sick” crew member was played by either a professional actor with extensive experience working as a standardized patient, or a high-fidelity patient simulator (Laerdal, model SIMMAN 3G). This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at MassGeneral Brigham, protocol number: 2018P002156. Informed consent was obtained from each participant.

### Procedures

At the beginning of each session, teams were oriented to the simulator environment, equipment, and location of resources (e.g. oxygen, checklists, medication). Using the web-based Research Electronic Data Capture (REDCap) platform, participants completed a demographic questionnaire and the International Personality Item Pool (IPIP-120), one of the NASA-identified sets of standard measures to assess the “big five” personality traits. According to the group randomization, participants (teams of three) participated in a series of four medical scenarios, the order of which was also randomized within each group. To reduce the risk of canceling, a simulation session in the event of participant unavailability on the day, we recruited four participants to most sessions and randomly selected teams of three to participate in each simulation, with one member sitting out. Participants who were either EMT or paramedics were assigned the role of crew medical officer (CMO), and the other two participants were not assigned any specific role, besides being part of a space mission crew. The FS was given the same medical checklist that the crew had available, and the FS was instructed to provide medical guidance to the crew using only the questions and recommendations from the checklist. The FS used a script to respond to queries in order to standardize the level of support given to each team.

### Medical Scenarios

We previously developed four medical event scenarios that require crew members to balance competing priorities of healthcare and mission success (Robertson et al., 2020). These scenarios



Figure 1. Study design and measurements.

(Figure 2) were standardized to ensure reproducible measurement of endpoints and outcomes to test study hypotheses; as well as to timing of sound, lighting, and special effects. Each scenario was specifically designed to be conducted with and without support from a FS (See Scenario Scripts – Supplement Material). In the groups that were randomized to receive support from a FS, this support was provided via teleconsultation with real-time audio and video communication on a laptop computer. As part of the space vehicle resources, all crews had

medical checklists from the International Space Station (ISS) Integrated Medical Group available. All medical material, equipment and supplies available in the simulator matched those in the checklists.

**Measurements**

a) *Spaceflight Resource Management Medical Tool (SFRM-Med)*: This is an observation-based crew behavior measurement system that was developed in a previous study (Dias et al., 2019). Two



Figure 2. Simulated medical events.

independent observers (EA and LKM), who are trained in non-technical skills assessment, observed the simulation videos and independently assessed the quality of team performance in four categories (information exchange, communication delivery, supporting behavior, team leadership-followership) using a 1–100 visual analog scale.

- b) *Technical performance*: specific critical processes of care were determined for each medical scenario based on existing clinical guidelines. Independent observers watched the simulation videos and assessed whether or not each critical process was completed. The percentage of completed items was used to indicate adherence to critical processes of care. Additionally, observers provided an overall technical performance score using a 1–100 visual analog scale. Time to scenario completion (in seconds) was also registered.
- c) *Heart rate variability (HRV)*: A chest strap heart rate sensor (Polar® H10) was used to extract intervals between consecutive heartbeats (RR intervals) from electrocardiogram signals. Data were transmitted via Bluetooth to a multisport wrist watch (Polar V800). Several previous studies have validated RR signals from Polar against gold standard electrocardiogram Holter (Giles et al., 2016; Gilgen-Ammann et al., 2019). Spectral analysis was performed in Kubios HRV© software

(v3.02) (Tarvainen et al., 2014), generating multiple HRV parameters as proxies for cognitive workload: low frequency/high-frequency (LF/HF) ratio, root mean square of successive differences (RMSSD), mean heart rate (HR), maximum heart rate, and mean R-R interval.

- d) *Electroencephalography (EEG)*: A 4-channel headband EEG monitor (Muse™) was used to measure participants' brain activity: two on the forehead and two behind the ears. This device produces bipolar readings using Fpz (center of forehead) as the reference for TP9 (top of left ear), Fp1 (left forehead), Fp2 (right forehead), and TP10 (top of right ear). EEG signals are oversampled and then downsampled from 256 Hz to yield an output sampling rate of 220 Hz. Engagement index (EI) was calculated based on previous studies that have validated this measure (Pope et al., 1995). Relative band powers were calculated on each channel, dividing the absolute band power by the sum of the total band powers, resulting in a value between 0 and 1. These values were averaged across the four channels generating one relative band power value for each band (alpha, beta, theta) at 220 Hz (Armanfard et al., 2016; Teo & Chia, 2018). Each band was then averaged over one second and the EI was calculated using the equation:  $\text{beta}/(\text{alpha} + \text{theta})$ . We calculated blinks per minute using the

Muse headband that also exports a Boolean blink detection value every 10 Hz.

- e) *NASA Task Load Index (NASA-TLX)*: The NASA-TLX is a multidimensional workload assessment tool developed by NASA that subjectively measures six dimensions of workload based on a weighted average of six subscales: mental demands, physical demands, temporal demands, performance, effort, and frustration (Hart & Staveland, 1988). A mobile application (iOS) of NASA-TLX was used to capture data from participants immediately after each simulation.

## Outcomes

The primary outcomes in this study were measures of technical performance and non-technical skills (SFRM-Med). Secondary outcomes included HRV parameters, EEG measures (EI and blink rate), and NASA-TLX score.

## Data Analysis

All datasets were integrated in a relational database using the visual analytics software Tableau (version 2019.4). All statistical analyses were performed using the IBM statistical software SPSS (version 24.0). Variables serving as primary and secondary outcomes were tested for normality using the Kolmogorov-Smirnov test which showed that data distribution was not normal. Continuous variables were reported as median (first–third interquartiles), and categorical variables were reported as absolute numbers (percentage). The Mann-Whitney U Test was used to compare continuous variables between the two groups. A  $p$ -value of less than 0.05 was considered statistically significant. Based on a previous preliminary study (Dias RD, Doyle T, Robertson JM, Thorgrimson JL, Gupta A, Mormann B, Pozner C, Smink DS, Lipsitz S, Musson D, Yule S., 2019) using SFRM-Med to assess space crews during simulated space simulations we found a mean SFRM-Med score of 68.25 (SD= 16.7) without FS support. We hypothesized a 20% increase in the SFRM-Med score with FS. To detect this effect size, we calculated that 64 measures (32 each group) would be needed to achieve 85% power and an alpha of 0.05.

## RESULTS

A total of 59 participants were included in this study. Their demographic characteristics, distributed by group (with FS vs. without FS), are described in Table 1. Both groups presented similar demographics, as expected due to the randomization procedures.

### Inter-Reliability Analysis

The inter-rater reliability analysis of the SFRM-Med tool yielded a moderate intraclass correlation coefficient of 0.62 with a 95% confidence interval of 0.560–0.66, and a  $p$ -value < 0.001. The average score between the 2 raters was used as a measure of a crews' non-technical performance. Forty percent of the videos were rated for technical performance by two independent observers, an emergency medicine resident (BM) and a general surgery resident (SP). We found a good inter-rater reliability with an intraclass correlation coefficient = 0.77, 95% CI: 0.60–0.87, and  $p$  < 0.001. A single rater conducted the assessment of the remaining videos.

### Primary Outcomes

In support of the hypothesis, teams that received support from a flight surgeon via teleconsultation had better clinical performance scores than teams without support in both domains: technical (Table 2) and non-technical (Table 3). There was no statistically significant difference between groups in time to scenario completion (in seconds). Analysis of the SFRM-Med scores by category showed that the information exchange component was the only parameter that was better in teams with support from a FS compared with those without support.

### Secondary Outcomes

Teams that received support from a FS reported a lower NASA-TLX compared with those without support (Table 4). Analysis of the different NASA-TLX domains showed that *mental demand* and *temporal demand* were the

**TABLE 1:** Demographics of study participants

	Without Flight Surgeon (FS) (N = 30)	With Flight Surgeon (FS) (N = 29)
<b>Age (years)</b>	26.0 (23.0–29.0)	26.0 (23.0–30.0)
<b>Sex – female, N (%)</b>	14 (48.3%)	14 (46.7%)
<b>Profession, N (%)</b>		
Graduate student	21 (70.0%)	20 (69.0%)
EMT/Paramedic	5 (16.7%)	7 (24.1%)
Undergraduate student	4 (13.3%)	2 (6.9%)
<b>Personality trait (IPI-NEO-120), 1–100 scale</b>		
Agreeableness	73.1 (63.3–78.1)	73.3 (67.5–80.5)
Conscientiousness	67.5 (62.0–81.4)	76.0 (62.7–83.1)
Openness	66.5 (61.9–70.7)	62.4 (57.3 - 67.80)
Extraversion	63.0 (53.3–74.2)	58.5 (52.9–66.2)
Neuroticism	38.4 (33.5–45.8)	37.5 (30.6–45.5)

Continuous variables were described as median (1st - 3rd interquartiles).

**TABLE 2:** Comparison of technical performance between groups

	Without Flight Surgeon (FS)8 teams 32 scenarios	With Flight Surgeon (FS)8 teams 32 scenarios	p-value
<b>Overall performance, 1–100 scale</b>	60.5 (50.0–70.0)	82.5 (75.0–93.8)	< <b>0.001</b>
<b>Adherence to critical processes, % completed</b>	90.0 (80.0–100.0)	100.0 (100.0–100.0)	< <b>0.006</b>
<b>Time to scenario completion, seconds</b>	438.0 (387.3–631.0)	460.0 (405.5–622.5)	0.619

Continuous variables were described as median (1st - 3rd interquartiles).

**TABLE 3:** Comparison of SFRM-Med scores between groups

	Without Flight Surgeon (FS)8 teams 32 scenarios	With Flight Surgeon (FS)8 teams 32 scenarios	p-value
<b>SFRM-med overall, 1–100 scale</b>	74.3 (67.2–78.7)	80.1 (75.9–83.0)	< <b>0.001</b>
<b>Behavioral categories</b>			
Information exchange	63.7 (59.7–69.3)	83.9 (79.7–86.8)	< <b>0.001</b>
Communication delivery	74.3 (65.1–79.1)	75.9 (69.4–84.6)	0.138
Supporting behavior	75.6 (65.2–81.6)	75.1 (72.6–81.5)	0.643
Team leadership-followership	82.9 (75.4–87.0)	84.0 (79.1–86.8)	0.1368

Continuous Variables Were Described as Median (1st - 3rd interquartiles).



TABLE 4: Self-reported cognitive workload

	Without <i>Flight Surgeon</i> (FS) (N = 30)	With <i>Flight Surgeon</i> (FS) (N = 29)	p-value
<b>NASA-TLX</b> , 1–100 overall weighted score	66.6 (57.6–71.3)	62.1 (51.5–67.1)	<b>0.012</b>
<b>NASA-TLX domains</b> , 1–100 raw score			
<i>Mental demand</i>	73.3 (63.3–77.9)	63.3 (55.4–74.6)	<b>0.025</b>
<i>Temporal demand</i>	81.7 (71.7–85.0)	73.3 (60.4–81.7)	<b>0.016</b>
<i>Physical demand</i>	32.5 (20.4–49.6)	28.3 (12.1–47.9)	0.279
<i>Effort</i>	60.8 (55.4–68.3)	60.0 (51.2–67.5)	0.264
<i>Frustration level</i>	51.6 (40.0–65.0)	43.3 (33.3–58.3)	0.113
<i>Performance (perceived failure)</i>	38.3 (27.1–50.0)	30.0 (23.7–46.7)	0.224

Continuous variables were described as median (1st - 3rd interquartiles)

TABLE 5: Physiological measures.

	Without <i>Flight Surgeon</i> (FS) (N = 30)	With <i>Flight Surgeon</i> (FS) (N = 29)	p-value
<b>HRV parameters*</b>			
<i>LF/HF ratio</i>	5.7 (4.1–7.5)	6.4 (4.7–8.9)	0.061
<i>RMSSD (ms)</i>	26.9 (17.8–35.0)	22.3 (15.0–32.6)	0.498
<i>Mean heart rate (bpm)</i>	88.9 (82.6–99.0)	94.4 (84.4–105.7)	0.191
<i>Maximum heart rate (bpm)</i>	101.1 (94.0–111.0)	105.6 (97.9–117.9)	0.221
<i>Mean RR (ms)</i>	679.9 (609.0–730.2)	639.1 (569.8–716.1)	0.233
<b>EEG-derived metrics</b>			
<i>Blink Rate (blinks/minute)</i>	35.6 (29.7–42.8)	39.9 (32.4–45.5)	<b>0.044</b>
<i>Engagement Index</i>	.58 (.44–.74)	.60 (.46–.79)	0.622

Continuous variables were described as median (1st - 3rd interquartiles).

only domains that yielded a statistically significant difference between groups.

Of all the physiological metrics measured, including HRV parameters and EEG-derived measures (EI and blink rate), only blink rate yielded a statistically significant difference between groups. Blink rates were higher in the teams with support from FS (Table 5).

## DISCUSSION

In this simulation-based randomized controlled trial, we investigated crew clinical performance and physiological metrics during high-fidelity medical scenarios in a simulated space environment under two different conditions: with and without support from a FS. We found that compared to teams without support

from a FS, teams that received support via teleconsultation performed better clinically in both domains: technical and non-technical skills. Furthermore, teams that received support reported lower perceived workload, although there was no difference in objective biomarkers of cognitive load and engagement between groups.

In support of the hypothesis, we found that groups with access to a flight surgeon via a real-time telecommunications link scored significantly higher on behavioral metrics and displayed more effective team behaviors as assessed by two expert video raters blinded to the intent of the study. Assessing subcomponents of the SFRM-Med tool, teams with support from a flight surgeon scored significantly higher on the category of information exchange. Specific behavioral practices that are enhanced relate to

situation awareness components of team communication such as gathering information, and recognizing that something is wrong. Teams with medical support were also judged to be better at providing situation assessment updates. Although teams with support also scored higher on the other three SFRM-Med categories (communication delivery, supporting behavior, team leadership-followership), there were no significant differences between groups on these variables of interest. It seems that the quality of information exchange was better, and future investigations may measure the style or particular manner of communication between crew members to explore this further.

Overall, technical performance was significantly higher for teams who operated with support from a flight surgeon, and those teams also demonstrated significantly higher adherence to critical processes of care (95.8% vs. 85.8% adherence). This equates to teams without support missing approximately one critical process for each medical event, compared to missing one critical process every fourth event when they had FS support. It is important to note that adherence was not 100% for teams despite using a checklist and having support from the flight surgeon. However, in other clinical contexts, previous studies have shown that checklists are effective in helping to guide teams through technical tasks (Weiser & Haynes, 2018). Teams with support were also faster in providing care and resolving the medical event, but not significantly so. From our observations, it is likely that the added time to coordinate with the flight surgeon via telecommunications resulted in more efficient medical care and coordination, without a net increase in time to treat.

Across all scenarios, teams with telemedical support from the flight surgeon at MCC reported significantly less overall cognitive demand. Examining the NASA-TLX subscales revealed that these differences can be attributed to less mental demand (mental and perceptual activity) and less temporal demand (feeling of time pressure). There were no significant differences between groups for the other dimensions (physical demand, effort, frustration, perceived failure) although self-reported cognitive load

was lower for each of these in the group with flight surgeon support. Feelings of time demand can compound high mental activity into a cognitive overload state, so techniques to support astronauts to manage time-pressured situations such as medical event management may be a focus of specific research and countermeasure development. Medical drills involving these psychologically realistic simulations, either in physical simulation or virtual/mixed reality could provide a platform for enhanced management and improved behavioral responses, akin to stress-inoculation training.

An important aim of the present study was to examine the feasibility of gathering objective physiological data reflective of cognitive load via unobtrusive sensors, complementing more established subjective measures (Dias et al., 2018). Unlike for perceived cognitive load (NASA-TLX), there were no significant differences in objective physiological-based measures of cognitive load between groups, although LF/HF ratio trended higher and RMSSD trended lower, both indicative of a high cognitive load for subjects in the groups with flight surgeon support. These findings may indicate that crews had a subjective perception of lower cognitive demand when the flight surgeon was supporting them; however, the actual cognitive demand was higher because it required interaction and coordination with the flight surgeon. Eye metrics (i.e. blink rate) can be used as a proxy for sustained attention. Previous studies have shown that increased attentional demand is associated with a decrease in blink rate (Maffei & Angrilli, 2018), but visual distractions are associated with an increase in blink rate (Annerer-Walcher et al., 2020). The fact that we found a lower blink rate in teams without flight surgeon support compared to those teams with support may reflect a combination of both a higher perceived cognitive workload in the without FS group, and more distraction due to communications via videocall in the with FS group.

As with any applied study, there were a number of limitations. Due to the challenges of recruiting astronauts (there were 48 active astronauts in the US at the time of the study), our subjects were not real astronauts. They did, however, fit NASA's profile of astronaut-like

participants, and we recruited crews with varied background training (life sciences, engineering, technology—along with one crew member with medical training). Our scale replica of the destiny module on ISS, by nature, was not a real space module and did not have the added complexity of weightlessness. The benefit of simulation control meant that we were able to reliably gather sensor data from every participant in each scenario with no missing or corrupted data. Our aim was to engineer and implement high-fidelity simulations that create an authentic experience for participants who are faced with the social reality that they must act as they would in a real emergency to manage the simulated crisis unfolding in front of them in real time. We focused on scenarios that were medically sound, technically feasible and psychologically real, with high ecological validity (Brewer & Crano, 2000) to approximate as many aspects of an experience as is practical. These included the physical, emotional, ergonomic, acoustic, cognitive, and visual aspects of medical emergencies (LeBlanc et al., 2011) during spaceflight.

Consequently, these results have practical applications related to the degree of autonomy under which crews will operate on these extended missions. Cis-lunar communication delays may lead to some challenges, for example, on planned Artemis missions, but are unlikely to have an impact to the degree of future Mars missions. The 20–40-minute time delays will require crews to assume decision making autonomy independent of Mission Control, a significant departure from current and past practices for low earth orbit missions. This will include the management of rapidly evolving medical emergencies during which there is no ability for Mission Control to provide direct supervision via telemedicine or telementoring capabilities.

As with any applied study, there were a number of limitations. Due to the challenges of recruiting astronauts (there were 48 active astronauts in the US at the time of the study), our subjects were not real astronauts. They did, however, fit NASA's profile of astronaut-like participants, and we recruited crews with varied background training (life sciences, engineering, technology—along with one crew member with medical training). Our

scale replica of the Destiny module on ISS, by nature, was not a real space module and did not have the added complexity of weightlessness. The benefit of simulation control meant that we were able to reliably gather sensor data from every participant in each scenario with no missing or corrupted data. Our aim was to engineer and implement high-fidelity simulations that create an authentic experience for participants who are faced with the social reality that they must act as they would in a real emergency to manage the simulated crisis unfolding in front of them in real time. We focused on scenarios that were medically sound, technically feasible, and psychologically real, with high ecological validity (Brewer & Crano, 2000) to approximate as many aspects of an experience as is practical. These included the physical, emotional, ergonomic, acoustic, cognitive, and visual aspects of medical emergencies (LeBlanc et al., 2011) during spaceflight. The study involved a number of human observers and assessors, who rated performance from video. Although inter-rater reliability was good, a single rater assessed technical performance from 60% of the videos which may have introduced unwanted bias in scoring.

A novel research question that arose from this study is how to develop automated monitoring of crew behavioral skills on future lunar (e.g. Artemis) (NASA: Artemis, n.d.) and long duration missions in order to provide remote support and objective performance data under variable autonomous conditions. Observational methods are the current gold standard for measurement of behavioral skills. However, this project lays the groundwork for a machine-learning algorithm to automate analysis of real-time objective measures of crew medical performance, based on biomarkers of cognitive load and team dynamics. By incorporating video observation, outcome measurement, and emerging sensor technologies, this work could be scaled to NASA Countermeasure Readiness Level 7 (Space Analog), and advance development of autonomous medical advisory systems. Future research could also leverage advances in extended reality-based modalities (Andrews et al., 2019) to transform the physical space flight simulator and medical event scenarios into a virtual format. This would bring many benefits for

team-based medical training for space crews, reduce reliance on physical space, and foster interaction for distributed teams on the ground. Critically, extended reality could also facilitate immersive in-flight training and real-time clinical guidance to support crew medical decision making and mission success on future LDEMs. Another potential application of this study findings is related to proving objective measures of performance at the team level, not only within-teams, but also between multiple teams working toward a common goal as part of a multi-team system.

## CONCLUSION

Spaceflight poses unique operational characteristics for simulation, and development of simulation platforms to study MTS response to emergency scenarios is an exciting and highly impactful prospect. Including medical emergency simulations (“megacodes”) in the training flow for astronaut crew could have the triple benefit of testing the spaceflight system, MTS response, and drilling specific medical event management skills. Incorporating unannounced medical simulations with metrics described in the present study into established long duration space analogs, such as The Hawai’i Space Exploration Analog and Simulation (HI-SEAS) (Anderson et al., 2016) could test both crew skills and system responsiveness to medical challenges. Time delay introduces an additional complexity to MTS and results from this study demonstrated that lack of real time support can hinder team performance in an emergency setting. This can be mitigated by use of checklists, development of clinical guidance tools, and establishing protocols regarding expectations and boundaries related to shared situation awareness across teams engaged in asynchronous work.

## KEY POINTS

- Future long duration exploration missions (LDEMs) will involve delays in signal transmission of up to 40+ minutes, affecting communications and data transfer. Crews may have to manage emergency situations, including medical events, with a degree of autonomy.

- We found that autonomous crews performed worse than crews with flight surgeon support on clinical, technical and behavioral criteria.
- Flight surgeon support was associated with lowered perceived workload among crew, but we did not see an associated drop in objective biomarkers of cognitive load.
- Aspects of the spaceflight Multi-Team System (MTS) through the lens of psychologically realistic medical scenarios, and is a platform for mixed reality training and clinical decision tools to support autonomous crews during future long duration exploration missions.

## DECLARATION OF CONFLICTING INTERESTS

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Dr. Yule reports consulting fees from Johnson & Johnson Institute, outside the submitted work.

## FUNDING

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded through a research grant from the National Aeronautics and Space Administration, number 80NSSC19K0745.

## AUTHOR’S NOTES

Dias RD, Robertson JM, Mormann B, e A, Thorgrimson JL, Suresh R, Smink DS, Lipsitz S, Doyle T, Musson D, Pozner C, Yule S. Investigating Autonomous Crew Performance and Behavioral Skills During In-Flight Medical Event Management. NASA Human Research Program Investigators’ Workshop. Galveston, TX. January 2020.

## ORCID iDs

Steven Yule  <https://orcid.org/0000-0001-9889-9090>

Lauren Kennedy-Metz  <https://orcid.org/0000-0002-2696-3943>

Roger D. Dias  <https://orcid.org/0000-0003-4959-5052>

## SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

## REFERENCES

- Anderson, A. P., Fellows, A. M., Binsted, K. A., Hegel, M. T., & Buckley, J. C. (2016). Autonomous, computer-based behavioral health countermeasure evaluation at HI-SEAS mars analog. *Aerospace Medicine and Human Performance*, 87(11), 912–920. <https://doi.org/10.3357/AMHP.4676.2016>.
- Andrews, C., Southworth, M. K., Silva, J. N. A., & Silva, J. R. (2019). Extended reality in medical practice. *Current Treatment Options in Cardiovascular Medicine*, 21(4), 18. <https://doi.org/10.1007/s11936-019-0722-7>.
- Annerer-Walcher, S., Körner, C., Beaty, R. E., & Benedek, M. (2020). Eye behavior predicts susceptibility to visual distraction during internally directed cognition. *Attention, Perception & Psychophysics*, 82(7), 3432–3444. <https://doi.org/10.3758/s13414-020-02068-1>
- Armanfard, N., Komeili, M., Reilly, J. P., & Pino, L. (2016). Vigilance lapse identification using sparse EEG electrode arrays. IEEE Canadian conference on electrical and computer engineering (CCECE), Vancouver, BC, Canada, 15-18 May 2016. pp. 1–4.
- Arriaga, A. F., Bader, A. M., Wong, J. M., Lipsitz, S. R., Berry, W. R., Ziewacz, J. E., Hepner, D. L., Booman, D. J., Pozner, C. N., Smink, D. S., & Gawande, A. A. (2013). Simulation-based trial of surgical-crisis checklists. *The New England Journal of Medicine*, 368(3), 246–253. <https://doi.org/10.1056/NEJMsa1204720>
- Brewer, M. B., & Crano, W. D. (2000). Research design and issues of validity. In: *Handbook of Research Methods in Social and Personality Psychology*, Cambridge University Press. pp. 3–16.
- Dias, R. D., Doyle, T., Robertson, J. M., Thorgrimson, J. L., Gupta, A., Mormann, B., Pozner, C., Smink, D. S., Lipsitz, S., Musson, D., & Yule, S. (2019). *Development of a web-based rating platform for measurement of crew behavioral skills during simulated medical emergencies in space*. Aerospace Medical Association (ASMA) Annual Scientific Meeting.
- Dias, R. D., Ngo-Howard, M. C., Boskovski, M. T., Zenati, M. A., & Yule, S. J. (2018). Systematic review of measurement tools to assess surgeons' intraoperative cognitive workload. *The British Journal of Surgery*, 105(5), 491–501. <https://doi.org/10.1002/bjs.10795>
- Flin, R., Youngson, G. G., & Yule, S. (2015). *Enhancing surgical performance: A primer in non-technical skills*. CRC Press.
- Giles, D., Draper, N., & Neil, W. (2016). Validity of the polar V800 heart rate monitor to measure RR intervals at rest. *European Journal of Applied Physiology*, 116(3), 563–571. <https://doi.org/10.1007/s00421-015-3303-9>.
- Gilgen-Ammann, R., Schweizer, T., & Wyss, T. (2019). RR interval signal quality of a heart rate monitor and an ECG Holter at rest and during exercise. *European Journal of Applied Physiology*, 119(7), 1525–1532. <https://doi.org/10.1007/s00421-019-04142-5>
- Hackman, J. R. (2003). Learning more by crossing levels: evidence from airplanes, hospitals, and orchestras. *Journal of Organizational Behavior*, 24(8), 905–922. <https://doi.org/10.1002/job.226>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds) *Advances in psychology*, (Vol. 52, pp. 139–183). [https://doi.org/10.1016/s0166-4115\(08\)62386-9](https://doi.org/10.1016/s0166-4115(08)62386-9)
- Human Research Roadmap (n.d.). <https://humanresearchroadmap.nasa.gov/>
- Komorowski, M., Thierry, S., Stark, C., Sykes, M., & Hinkelbein, J. (2021). *On the challenges of anesthesia and surgery during interplanetary spaceflight*. *Anesthesiology*. <https://doi.org/10.1097/ALN.0000000000003789>
- LeBlanc, V. R., Manser, T., Weinger, M. B., Musson, D., Kutzin, J., & Howard, S. K. (2011). The study of factors affecting human and systems performance in healthcare using simulation. *Simulation in Healthcare: Journal of the Society for Simulation in Healthcare*, 6(Suppl), S24–S29. <https://doi.org/10.1097/SIH.0b013e318229f5c8>
- Maffei, A., & Angrilli, A. (2018). Spontaneous eye blink rate: an index of dopaminergic component of sustained attention and fatigue. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, 123, 58–63. <https://doi.org/10.1016/j.ijpsycho.2017.11.009>.
- Mathieu, J. E., Marks, M. A., & Zaccaro, S. J. (2001). Multi-team systems. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds), *Organizational psychology. Handbook of industrial, work and organizational psychology* (Vol. 2, pp. 289-313). Sage.
- Musson, D. M., Doyle, T. E., Saary, J., Turnock, M., Dias, R., & Yule, S. (n.d.). *Evolution of design concept for a ground-based immersive simulator for spaceflight medical operations*. NASA Human Research Program Investigator's Workshop.
- Musson, D. M., & Helmreich, R. L. (2004). Team training and resource management in health care: current issues and future directions. *Harvard Health Policy Review: A Student Publication of the Harvard Interfaculty Initiative in Health Policy*, 5(1), 25–35.
- NASA: Artemis (n.d.) <https://www.nasa.gov/specials/artemis/>
- Pope, A. T., Bogart, E. H., & Bartolome, D. S. (1995). Bio-cybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology*, 40(1–2), 187–195. [https://doi.org/10.1016/0301-0511\(95\)05116-3](https://doi.org/10.1016/0301-0511(95)05116-3).
- Robertson, J. M., Dias, R. D., Gupta, A., Marshburn, T., Lipsitz, S. R., Pozner, C. N., Doyle, T. E., Smink, D. S., Musson, D. M., & Yule, S. (2020). Medical event management for future deep space exploration missions to mars. *The Journal of Surgical Research*, 246(1), 305–314. <https://doi.org/10.1016/j.jss.2019.09.065>
- Salas, E., Grossman, R., Hughes, A. M., & Coultas, C. W. (2015). Measuring team cohesion: observations from the science. *Human Factors*, 57(3), 365–374. <https://doi.org/10.1177/0018720815578267>.
- Shuffler, M. L., Jiménez-Rodríguez, M., & Kramer, W. S. (2015). The science of multiteam systems: a review and future research Agenda. *Small Group Research*, 46(6), 659–699. <https://doi.org/10.1177/1046496415603455>
- Tarvainen, M. P., Niskanen, J. P., Lipponen, J. A., Ranta-Aho, P. O., & Karjalainen, P. A. (2014). Kubios HRV—heart rate variability analysis software. *Computer Methods and Programs in Biomedicine*, 113(1), 210–220. <https://doi.org/10.1016/j.cmpb.2013.07.024>
- Teo, J., & Chia, J. T. (2018). EEG-based excitement detection in immersive environments: An improved deep learning approach. *AIP Conference Proceedings*, 2016(1), 020145. <https://doi.org/10.1063/1.5055547>

- Weiser, T. G., & Haynes, A. B. (2018). Ten years of the surgical safety checklist. *The British Journal of Surgery*, *105*(8), 927–929. <https://doi.org/10.1002/bjs.10907>
- Wood, T. C., Raison, N., Haldar, S., Brunckhorst, O., McIlhenny, C., Dasgupta, P., & Ahmed, K. (2017). training tools for nontechnical skills for surgeons—a systematic review. *Journal of Surgical Education*, *74*(4), 548–578. <https://doi.org/10.1016/j.jsurg.2016.11.017>
- Yule, S. J., & Walls, R. M. (2012). Advanced life support training: does online learning translate to real-world performance? [Review of advanced life support training: does online learning translate to real-world performance?]. *Annals of Internal Medicine*, *157*(1), 69–70. <https://doi.org/10.7326/0003-4819-157-1-201207030-00013>
- Zaccaro, S. J., Marks, M. A., & DeChurch, L. (2012). *Multiteam systems: An organization form for dynamic and complex environments*. Routledge.

*Date received: May 28, 2021*

*Date accepted: November 16, 2021*