Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 1407–1416
https://doi.org/10.54941/ahfe1008369

Decoding the Language of Team Performance: A Machine Learning Framework for Sentiment-Based Performance Modeling and Team Situational Awareness

Gracie Gray[1], Vineetha Menon[1], Kristin Weger[2], Bryan Mesmer[2], and Nicholas Loyd[2]

[1]Department of Computer Science, University of Alabama in Huntsville, Huntsville 35899, USA

[2]Department of Industrial Systems Engineering and Engineering Management, University of Alabama in Huntsville, Huntsville 35899, USA

ABSTRACT

This research introduces a multi-faceted AI-enabled big data sentiment analytics framework for modeling and predicting team performance in high-stakes simulation environments. Communication data from 40 teams participating in a firefighting simulation were collected. Big data and sentiment analysis techniques were used to examine communication dynamics and identify relationships among individual factors, team communication dynamics, and team situational awareness (TSA) and their collective predictors of mission success. By leveraging these insights, the proposed big data analysis framework provides a data-driven approach to enhance TSA, improve team interaction, and optimize human–machine coordination. It also supports the development of adaptive feedback systems, AI/ML model training, and improving performance prediction capabilities across diverse operational contexts. The findings highlight the potential of sentiment-based analytics to uncover subtle behavioral and cognitive indicators of team performance that traditional evaluation methods often overlook. This integration of human communication data with advanced analytics provides predictive indicators for team success and training optimization. This research advances the field of industrial and systems engineering by bridging behavioral analytics and performance modeling. It provides a systematic method for transforming unstructured team communication data into actionable insights that enhance decision-making, training, and human-AI collaboration in complex environments.

Keywords: Artificial intelligence, Machine learning, Human-computer interaction, Human performance modelling, Human-machine communications, Big data analytics, Human-machine interactions, Sentiment analysis modelling

INTRODUCTION

Effective team performance, especially in high-stakes environments, depends heavily on communication, shared situational awareness (SA), coordination among team members, and understanding of evolving

Received June 29, 2026; Revised September 24, 2026; Accepted September 29, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

tasks. Teams in high-stress domains like emergency response and combat face limited information where communication breakdowns can degrade effective decision making and coordination (Martin et al., 2016). Communication is not only the mechanism through which information flows within teams but a reflection of the underlying team cognition and team situational awareness (TSA) (Marlow et al., 2018). TSA reflects the degree to which each team member holds a shared and accurate understanding of team tasks and performance (Endsley & Jones, 1997). Prior research has identified the influence of communication on TSA, on coordination efforts and overall mission success (Weger, 2018). Other recent research reinforces the connection between text-based communication and its association with team performance (Weger et al., 2022, Menon et al., 2022).

As complex simulation platforms increasingly serve as tests for evaluating human performance in mission-critical contexts, there is a growing need for scalable, data-driven methods that can capture subtle behavioral and cognitive indicators influencing team outcomes. Traditional evaluation approaches, often reliant on expert observation or post-hoc assessment, struggle to account for the richness and volume of unstructured communication data generated during such operations (Weger, 2018). Advances in big data analytics, artificial intelligence (AI), machine learning (ML), and sentiment analysis are promising paths for overcoming these limitations, enabling fine-grained modeling of team communication patterns and their relationship to TSA as cognitive indicators of mission success performance. In particular, natural language processing (NLP)-based sentiment analysis enables automated extraction of behavioral and cognitive indicators from team communication that are difficult to detect through manual observation alone. Recent studies in human-AI teaming emphasize the need for data-driven models that support adaptive feedback, training optimization, and improved human–machine coordination (Menon et al., 2022, Menon et al., 2023). Also situational awareness remains highly relevant across domains such as emergency response, aviation, and increasingly human-machine systems according to a recent systematic review that highlights the importance of understanding SA as a performance indicator (Fang et al., 2025).

This research introduces a multi-faceted AI enabled big data sentiment analytics framework designed to identify interpretable communication features associated with team effectiveness and predict team performance using communication data from 40 firefighting simulation teams. These indicators, often too subtle for human raters to detect consistently, can provide valuable insights into team functioning, risk management, and leadership. The goal is to support scalable performance assessment and dynamic feedback for human-machine teaming and human-in-the-loop systems in high stakes environments. These insights advance the field by bridging behavioral analytics with predictive performance modeling to demonstrate how unstructured communication data can be leveraged to enhance decision-making, team training, risk mitigation, and mission resilience. Our findings contribute to future research on human–agent team

evaluation by identifying communication-based indicators of coordination and situational awareness, addressing research needs identified by Iftikhar et al. (2023).

METHODS

Data collection**:** Data for this study were collected from 40 four-person teams performing firefighting tasks in a microworld simulation environment of four simulations. All message data sent was timestamped and recorded for all individuals throughout the course of the timeframe of the experiment. Simulation 0 was a training simulation. Simulations 1 and 3 were independent fire-rescue exercises. Simulation 2 was a team discussion session and held between simulations 1 and 3. Simulation 2 was included as reflexive team behavior has been found to facilitate process shifts in transactional and action phases as well as correlate with improved team performance (Weger et al., 2022). The simulation required teams to protect forest and residential areas and extinguish dynamic raging fires over the course of 15-minute simulation missions. Teams operated from independent workstations, each assigned a fire response vehicle, engine or helicopter, with distinct capabilities. The environment provided full terrain visibility but included hidden-profile design through imbalance information about resources, encouraging explicit communication for coordination. Time stamped text-based communication data from all four simulation phases (Simulation 0-3) were logged and included in the subsequent analysis as illustrated in Figure 1 from (Menon et al., 2022).

Observed Features: The observed features included team member turn taking per team member, simulation number, and situational awareness (SA) features representing perception, comprehension, projection, action, and performance as used in our prior work (Menon et al., 2022, Menon et al., 2023). Team performance was assessed using a combined score derived from individual performance scores, which served as the classification target. Teams were categorized into low, mid, or high performance ranks. A subset of communication was manually coded by domain experts using established TSA classification. These subsets categorized communication segments as reflecting perception (SA level 3), comprehension (SA level 4), projection (SA level 5), or action (SA level 6) categories of situational awareness. These additions contributed to TSA-related features, including role-specific SA averages.

Feature Engineering: New features that capture communication dynamics were extracted from the data to include: player turn taking/participation patterns, player TSA average, sentiment metrics, linguistic subjectivity, pronoun rate (eg. first, second, or third person pronouns), linguistic passivity rate, formality score, hedging rate, average words per message, participation Gini, turn entropy, role-based message metrics, message relevance metrics, role-based relevance metrics, simulation-based relevance metrics, and task completion-based relevance metrics. As a result, we automated extraction of 50 communication based features, which were considered during feature selection prediction modeling. This research is inspired by and is an extension of our prior work in (Menon et al., 2022, Menon et al., 2023).

A	B	C	D	E
Turn Taking	Time	Total	Simulation	SA Team
4	0:01:51	ill take the village in the southeast	0	5
2	0:03:20	ill take the houses in the north	0	5
4	0:03:31	rgr whos the heli?	0	3
1	0:04:01	me	0	
2	0:04:08	it moves so much faster than fire truck	0	
3	0:04:33	the fire truck is annoyingly slow	0	
4	0:04:41	our weakest area seems to be by the northeast	0	
4	0:05:29	how long is the heli water lasting?	0	
2	0:05:40	pretty decent amount	0	
1	0:05:43	not very long	0	
3	0:06:14	is someone keeping an eye on the southeast village	0	

Figure 1: Data snippet showing simulation data, and observed team and new features extracted such as turn taking, simulation number, team SA scores.

Feature Selection

The data resulting from the observation and feature engineering phases were then used in the feature selection phase. For this, a Random Forest model was selected for feature ranking because it can capture non-linear relationships between variables. Figure 1 was generated with Random Forest by computing each engineered feature's importance via Gini importance and then grouping them all into 7 conceptual categories. Each category's Gini importance score was summed to provide a high-level understanding of each category's contribution. Task relevance included task relevance metrics based on a similarity score between chat messages and simulation objectives. Individual TSA was the calculated TSA mean of each team member. Participation measured the metrics calculated via the Gini Coefficient and average message counts. Communication style included metrics such as linguistic passivity, formality, first-person pronoun usage, hedging rate, and subjectivity metrics. Team TSA calculated each team's total TSA. Sentiment included the calculated sentiment score of messages and the delta change of sentiment. Others had the remaining features that did not fit neatly into other categories. Note that the conceptual categories were only used for visualization and understanding. Later feature selection evaluated each engineered feature individually.

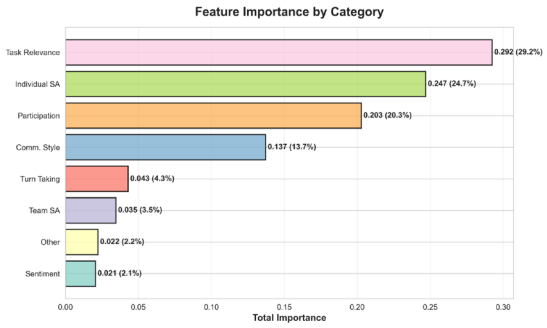


Figure 2: Feature importance scores derived by the random forest classifier.

After feature engineering, highly correlated features (Pearson $r > 0.95$) were removed to reduce multicollinearity. The dataset was then split as 75%

training (n = 30) and 25% test sizes (n = 10) sets using stratified sampling to preserve class distributions. All numeric features were z-score normalized based on the training set.

Feature selection continued in two stages on the training data. First, near-zero variance features (threshold = 0.01) were removed to reduce noise. Second, univariate feature selection via the SelectKBest implementation in scikit-learn was applied to identify the most informative features (Pedregosa et al., 2011). Initial analysis kept five features, but later coefficient stability assessment across cross-validation folds revealed that two features (turn entropy, SA 5 turn-taking) exhibited high variability (CV > 25%). Their removal improved model performance. These features were excluded, yielding a final three-feature model. The top three selected features are as follows:

- ***Hedging rate:*** Frequency of linguistic uncertainty markers (e.g., "maybe," "perhaps," "might"). Higher hedging rates may reflect uncertainty or reduced confidence during decision-making, whereas lower levels may indicate more decisive and coordinated communication.
- ***Relevance variability (SD)***: Standard deviation of task-relevance scores across communication, indicating consistency of task focus. Teams exhibiting more consistent task relevance may maintain greater focus on mission objectives, whereas high variability may indicate shifting attention or inconsistent coordination.
- ***Player 1 SA Average***: Role-specific situational awareness level for the team member in position 1, typically reflecting leadership or coordination. The importance of role-specific situational awareness suggests that maintaining shared understanding within key coordinating roles contributes substantially to overall team performance.

The selected features capture three complementary aspects of effective team performance: communication confidence, consistency of task focus, and role-specific situational awareness. Together, they provide a multidimensional representation of how teams exchange information and coordinate actions under operational demands. Figure 2 illustrates the proposed big data analytics and sentiment analysis based framework to understand the role of communication dynamics and TSA in prediction of mission performance in human-machine teaming scenarios.

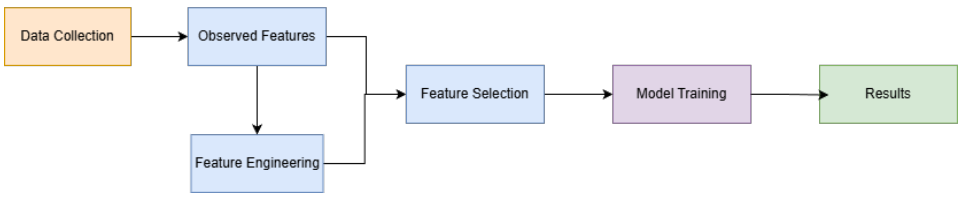


Figure 3: The proposed big data sentiment analytics based framework for prediction of team performance.

Model Training and Evaluation

For experimental validation, a training size of 75% data was selected, i.e. a sample size of n = 30 training teams as it gave the best results for our use case. To improve our emphasis on interpretability of the selected features, three linear regularized ML models were selected for supervised classification purposes, namely, Ridge Regression (L2 regularization, $\alpha = 10.0$), Logistic Regression (C = 0.001, L2 penalty), and Naïve Bayes classifier. The ML model performance was assessed using a stratified 5-fold cross-validation on the training set, with metrics including accuracy, balanced accuracy, and macro- averaged F1 score. The best-performing ML model was retrained on the complete training set and evaluated once on the held-out test set. Two robustness analyses were conducted: (1) coefficient stability, measuring feature weight variability across CV folds, and (2) feature ablation, quantifying performance impact of individually removing each feature while holding other model factors constant.

RESULTS

The cross-validation (CV) results of all the ML classification models documented in Table 1 shows high consistency across linear classification. Ridge Regression and Logistic Regression produced near-identical performance metrics: mean CV accuracy of 86.7% (±12.5%), balanced accuracy of 88.9%, and macro-averaged F1 score of 86.4%. Both models had zero overfitting gap between training and validation accuracy, indicating that the models captured generalizable patterns rather than just memorizing training data. Naïve Bayes achieved comparable but slightly lower performance (80.0% ±12.5% accuracy) with a small overfitting gap (7.5%). The high predictive performance of these results, especially with the Ridge Regression and Logistic Regression, suggest meaningful aspects of team performance can be predicted using only three communication-derived features. Those features, linguistic hedging rate, relevance variability, and role-specific situational awareness, indicate that the underlying patterns these factors represent may play an important role in mission outcomes. While further validation is needed to assess the generalizability of these findings, the current findings support the value of using communication data to find key interpretable indicators that can support team performance assessment.

Table 1: Cross validation performance across all ML models.

Model	%Accuracy	%Balanced Accuracy	F1 Macro	Overfit Gap
Ridge regression	86.7 ± 12.5	88.9	0.864	0.000
Logistic regression	86.7 ± 12.5	88.9	0.864	0.000
Naïve bayes	80.0 ± 12.5	78.9	0.750	0.075

The cross validation accuracy across all the ML models for 75% training and 25% text data split is as shown in Figure 2. The individual

cross-validation accuracy scores across multiple folds showed moderate variability ranging from 67% to 100% across folds, reflecting the small sample size and inherent prediction difficulty of something as complex as TSA. However, Ridge Classifier and Logistic Regression showed identical fold-level performance (overlapped in plot), further suggesting algorithmic convergence as seen in Figure 2. Thus, 3-fold cross validation gave the best performance for all ML models chosen for this study.

Coefficient Stability and Feature Importance

Ridge Classifier coefficients showed high stability across cross-validation folds, with all three features exhibiting coefficient variation below 10% (Table 2). Hedging rate showed the highest stability (0.2611 ± 0.0185, CV = 7.1%), followed by relevance variability (0.2522 ± 0.0235, CV = 9.3%) and Player 1 SA average (0.2550 ± 0.0247, CV = 9.7%). This low variability indicates that feature contributions are robust across different training subsets and not artifacts of specific train-validation splits. Figure 3 gives the feature importance rankings based on ridge regression's average coefficient results. Further, it shows balanced contributions across top 3 features, with hedging rate (0.277) slightly higher than relevance variability (0.257) and Player 1 SA average of team leadership (0.243). The relatively similar importance distribution suggests that all three features capture unique, complementary aspects of team communication quality for TSA. Hence, demonstrating the relationship between communication dynamics, TSA and team performance prediction.

Collectively, these findings suggest that successful teams are characterized not only by what they communicate, but also by how consistently they communicate, how confidently information is exchanged, and how effectively key team members maintain situational awareness. The relatively balanced contribution of the three features indicates that team performance emerges from the interaction of multiple complementary communication behaviors rather than a single dominant predictor.

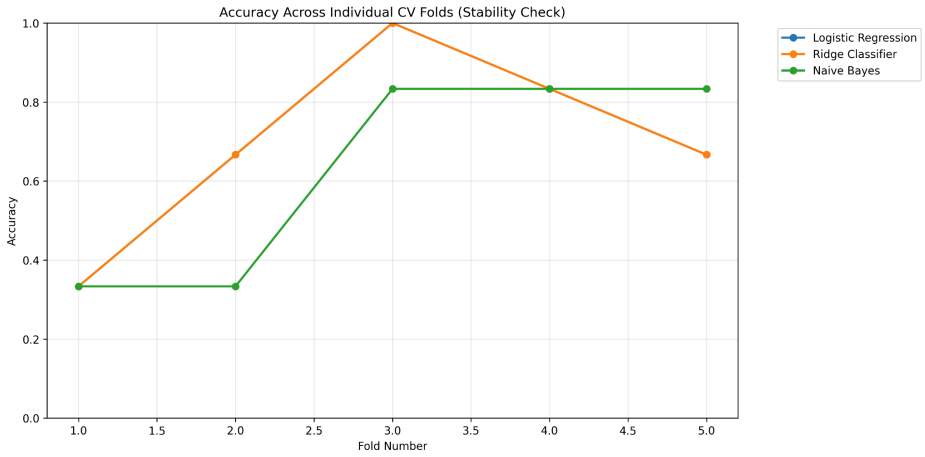


Figure 4: Cross validation accuracy comparison across all the ML models.

While the predictive performance of the ML models provides evidence that communication-based features are effective indicators of team success, the primary contribution of this study lies in identifying interpretable behavioral characteristics associated with effective teamwork. Compared to the traditional complex black-box nature of the ML models, the communication features selected by the interpretable ML models provide insight into how teams coordinate information, maintain shared SA, and communicate uncertainty during dynamic mission execution. These findings suggest that communication behavior itself contains valuable signals for assessing team effectiveness.

Collectively, these findings suggest that successful teams are characterized not only by what they communicate, but also by how consistently they communicate, how confidently information is exchanged, and how effectively key team members maintain situational awareness. The relatively balanced contribution of the three features indicates that team performance emerges from the interaction of multiple complementary communication behaviors rather than a single dominant predictor.

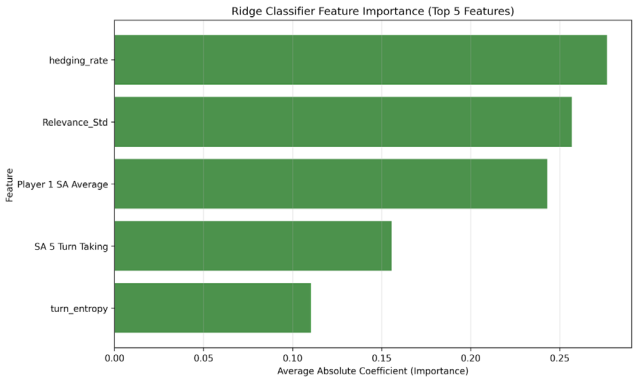


Figure 5: A bar chart of the top five Ridge Classifier features' importance.

Overall, the results demonstrate that interpretable communication features can effectively characterize team behavior while maintaining strong predictive performance. Beyond classification accuracy, the framework identifies communication patterns that may support performance assessment, adaptive feedback, and future intelligent decision-support systems in complex team environments.

CONCLUSION

This research demonstrates the potential of integrating big data methodologies and analytics to capture nuanced collective team behavioral dynamics that complement traditional performance metrics in complex, high-stakes simulation environments. This study analyzed communication data from 40 firefighting simulation teams to reveal how linguistic patterns and communication dynamics are associated with team situational awareness (TSA) and contribute to mission success. The experimental results show that

team performance in complex simulation environments can be predicted with high accuracy (86.7% cross-validation) using a set of three behavioral and linguistic features: linguistic hedging rate, task relevance variability, and role-specific situational awareness. These findings offer interpretable insights into the factors that influence team operational effectiveness and the significance of using ML models to understand team communication dynamics. While these results were derived from a controlled fire-fighting simulation, they offer evidence that fine-grained communication features may be useful indicators of TSA in complex operational settings. The identified communication features provide insight into how teams develop shared understanding, coordinate actions, and adapt to evolving mission demands, offering information beyond traditional performance measures.

This research demonstrates how computational analysis of team communication can uncover collective behavioral patterns that support collaboration, coordination, and adaptive team performance. Furthermore, these findings contribute to the development of human-centered AI by providing interpretable communication indicators that can support explainable AI, adaptive feedback systems, intelligent decision support, and enhanced human–AI collaboration in complex operational environments.

Although the framework demonstrated promising predictive performance, the findings should be interpreted within the context of several limitations. The study was conducted using text-based communication from a controlled firefighting simulation involving 40 teams and did not incorporate verbal communication, non-verbal behaviors, or other multimodal measures that may also influence team situational awareness and performance. Therefore, additional validation is needed before generalizing the framework to broader operational environments.

Future work should validate this framework with operational teams, automate feature extraction to support real-time communication analytics, and investigate temporal communication dynamics throughout missions. Such capabilities could enable continuous monitoring of team communication, early detection of declining situational awareness, and adaptive feedback to instructors or decision-support systems during mission execution. Future research should also investigate the use of large language models (LLMs) to support automated situational awareness coding and behavioral annotation while preserving the interpretability of communication-derived performance indicators.

ACKNOWLEDGMENT

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF2420158. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

Endsley, M., & Jones, W. (1997). Situation awareness information dominance and information warfare. In researchgate.net (p. 94). Wright-Patterson Air Force Base. https://www.researchgate.net/profile/Mica-Endsley/publication/235167825_Situation_Awareness_Information_Dominance_Information_Warfare/links/542b1add0cf27e39fa917e04/Situation-Awareness-Information-Dominance-Information-Warfare.pdf

Fang, R., Zhou, Q., Zhou, C., Yuan, S., Wang, K., Li, Q., Zhang, Y., & Li, J. (2025). A Systematic Review of Empirical Studies on Situation Awareness: Perspectives From the Interaction Among Humans, Machines, and the Task Environment. PsyCh Journal, 14(5), 718–733. https://doi.org/10.1002/pchj.70027

Iftikhar, R., Chiu, Y.-T., Mohammad Saud Khan, & Caudwell, C. (2023). Human–Agent Team Dynamics: A Review and Future Research Opportunities. IEEE Transactions on Engineering Management, 1–16. https://doi.org/10.1109/tem.2023.3331369

Marlow, S. L., Lacerenza, C. N., Paoletti, J., Burke, C. S., & Salas, E. (2018). Does team communication represent a one-size-fits-all approach?: A meta-analysis of team communication and performance. Organizational Behavior and Human Decision Processes, 144(1),145–170. ScienceDirect. https://doi.org/10.1016/j.obhdp.2017.08.001

Martin, E., Nolte, I., & Vitolo, E. (2016). The Four Cs of disaster partnering: communication, cooperation, coordination and collaboration. Disasters, 40(4), 621–643. https://doi.org/10.1111/disa.12173

Menon, V., Weger, K., Mesmer, B., & Gholston, S. (2022). Using big data analytics for sentiment analysis to explore team communication dynamics in human machine interactions for team situational awareness. Proceedings of the IEEE International Conference on Human-Machine Systems (ICHMS), 5, 1–6. https://doi.org/10.1109/ichms56717.2022.9980604

Menon, V., Weger, K., Mesmer, B., & Gholston, S. (2023). Sentiment analysis based human-machine teaming dynamics modelling for improved situational awareness in simulation environments. Proceedings of the IEEE Conference on Systems, Man, and Cybernetics, 2023.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Louppe, G., Prettenhofer, P., Weiss, R., Weiss, R. J., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res., 12. https://doi.org/10.5555/1953048.2078195

Weger, K. (2018). A classification of behavior indicators exhibited in team communication: The intricate role of team reflexivity with team cognition, regulatory processes and performance (p. 305) [Dissertation].

Weger, K., Leder, S., Mesmer, B., Menon, V., & Schaub, H. (2022). How Effectively Do We Communicate? An Analysis of Team Reflexivity in Transition and Action Phases of Team Collaboration. IEEE Transactions on Professional Communication, 65(3), 392–410. https://doi.org/10.1109/tpc.2022.3186773

Weger, K., Menon, V., Hollingsworth, A., & Schaub, H. (2022). Information sharing behaviors in text-based communication during emergency response. Journal of Emergency Management, 20(6), 499–516. https://doi.org/10.5055/jem.0731