

Does One Model Fit All?

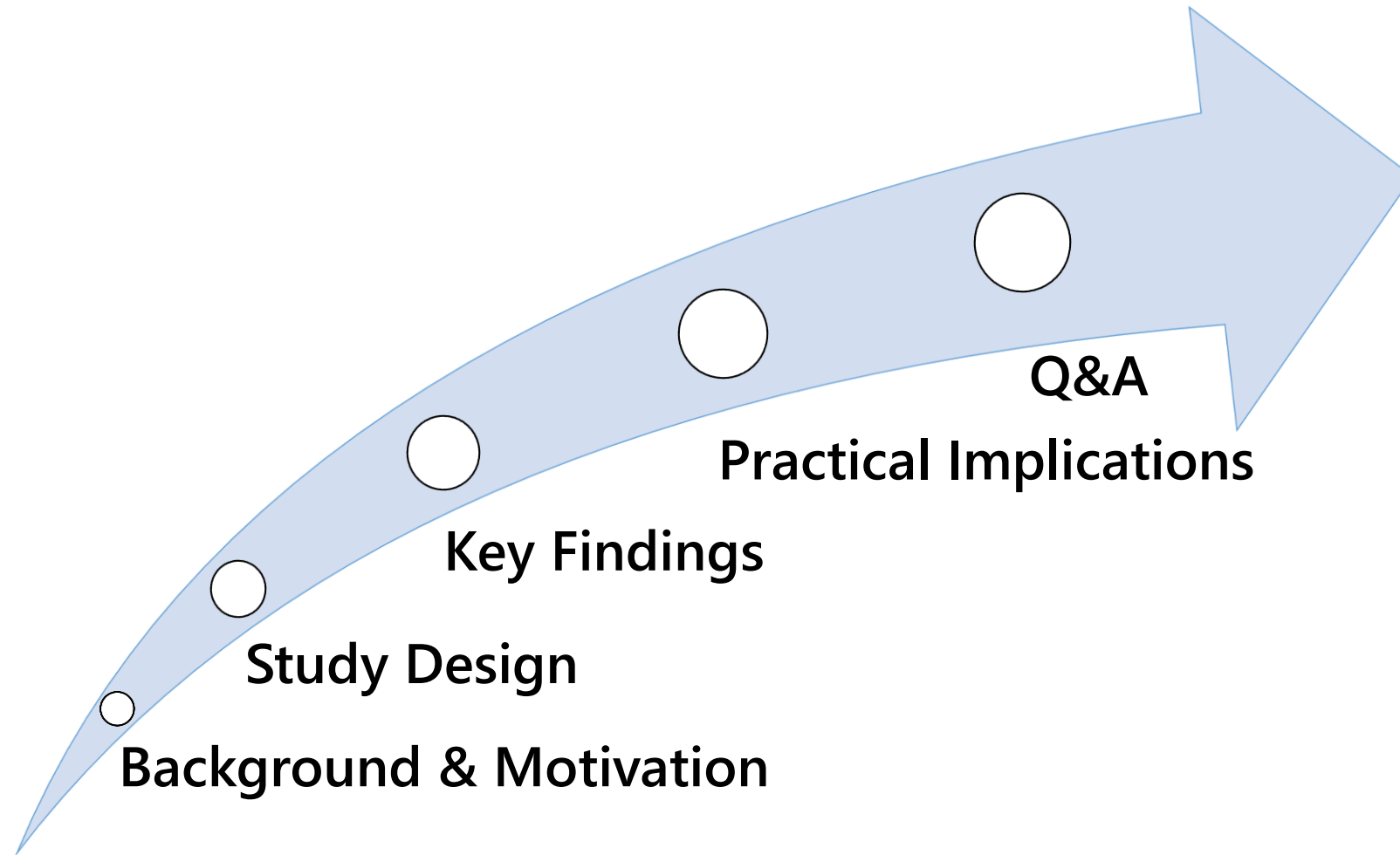
Personalized vs. Generalized Machine Learning Models for
Workload Prediction in Human-in-the-Loop Systems

Md Doulotuzzaman Xames* & Taylan G. Topcu, PhD

Sociotechnical Systems Engineering Lab (STSELab)
Grado Department of Industrial & Systems Engineering
Virginia Tech

Washington, DC
22 September 2026

Outline



Background

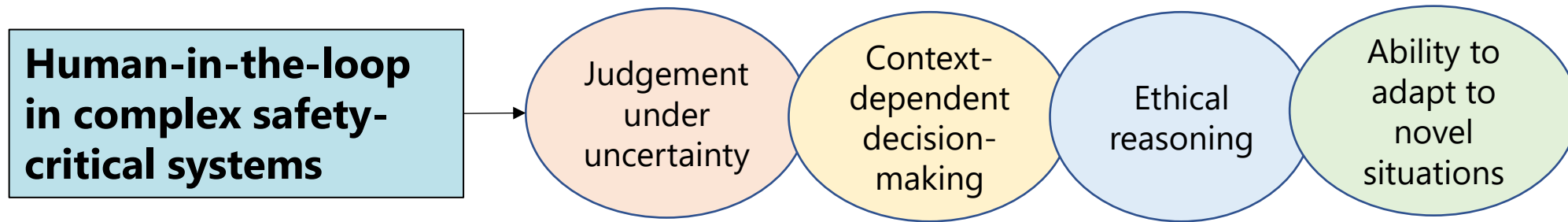
- Human-in-the-loop (HITL) systems are **sociotechnical** systems that achieve their **objectives** through the cooperation of both **human** and **technological** components.



Healthcare delivery system

Defense command-and-control (C2) centers

Motivation



- HITL systems depend on the **HITL** to sense, decide, and act under time pressure (Endsley & Kaber, 1999; Parasuraman et al., 2000).
- When an HITL's **workload** exceeds their capacity, **situational awareness** drops, **errors** climb, and system-level **performance** degrades (Mumford, 2006; Baxter & Sommerville, 2011)

The **next safety incident** may not start with a hardware fault;

it may start with an **overloaded operator** whose excessive workload no model saw coming.

Research Gap & Research Questions

- Most HITL workload prediction models **pool data** across everyone and train a single **global model**, implicitly **assuming** workload is generated the same way for **everyone** (Lou et al., 2022; Jiang et al., 2024; Fagbenle et al., 2025).
- The **same work** configuration may be experienced **differently** by different individuals (Kristof-Brown et al., 2005; Shanafelt et al., 2009).

RQ1

How does personalized, operator-specific machine learning (ML) performance compare with conventional global models trained on pooled data?

RQ2

How much do operational predictors of workload vary across individuals, and what does that reveal about heterogeneity in workload-generation mechanisms?

Why Healthcare System?



Safety-critical HITL system

Clinician workload directly shapes decision quality and patient outcomes (National Academies of Sciences, 2019, West et al., 2018), structurally analogous to cockpits, control rooms, and C2 centers.

A testbed choice, not a scope limit: this is a *proof-of-concept* for personalized workload modeling in HITL systems in general

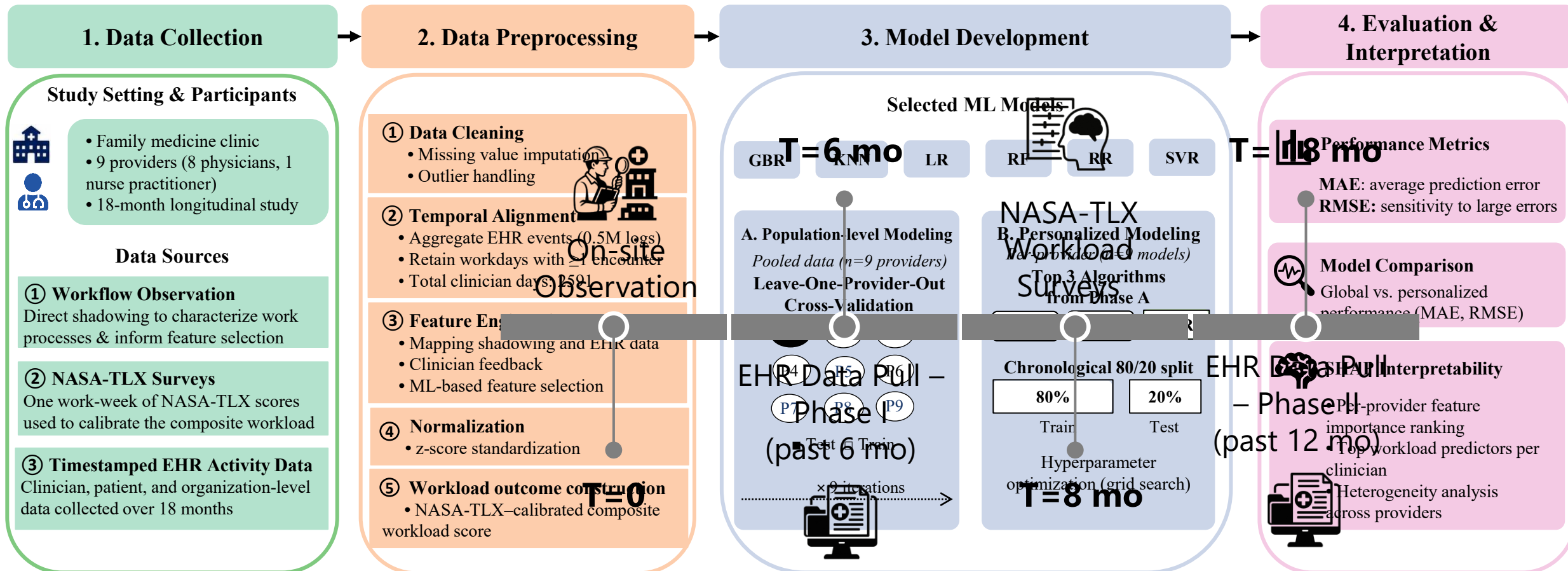


A natural experiment in heterogeneity

Multiple operators performing the same job, in the same organization, under the same policies (Carayon, 2006; Holden et al., 2018), isolating operator-level differences from contextual ones.

*EHR – Electronic health records

Study Design



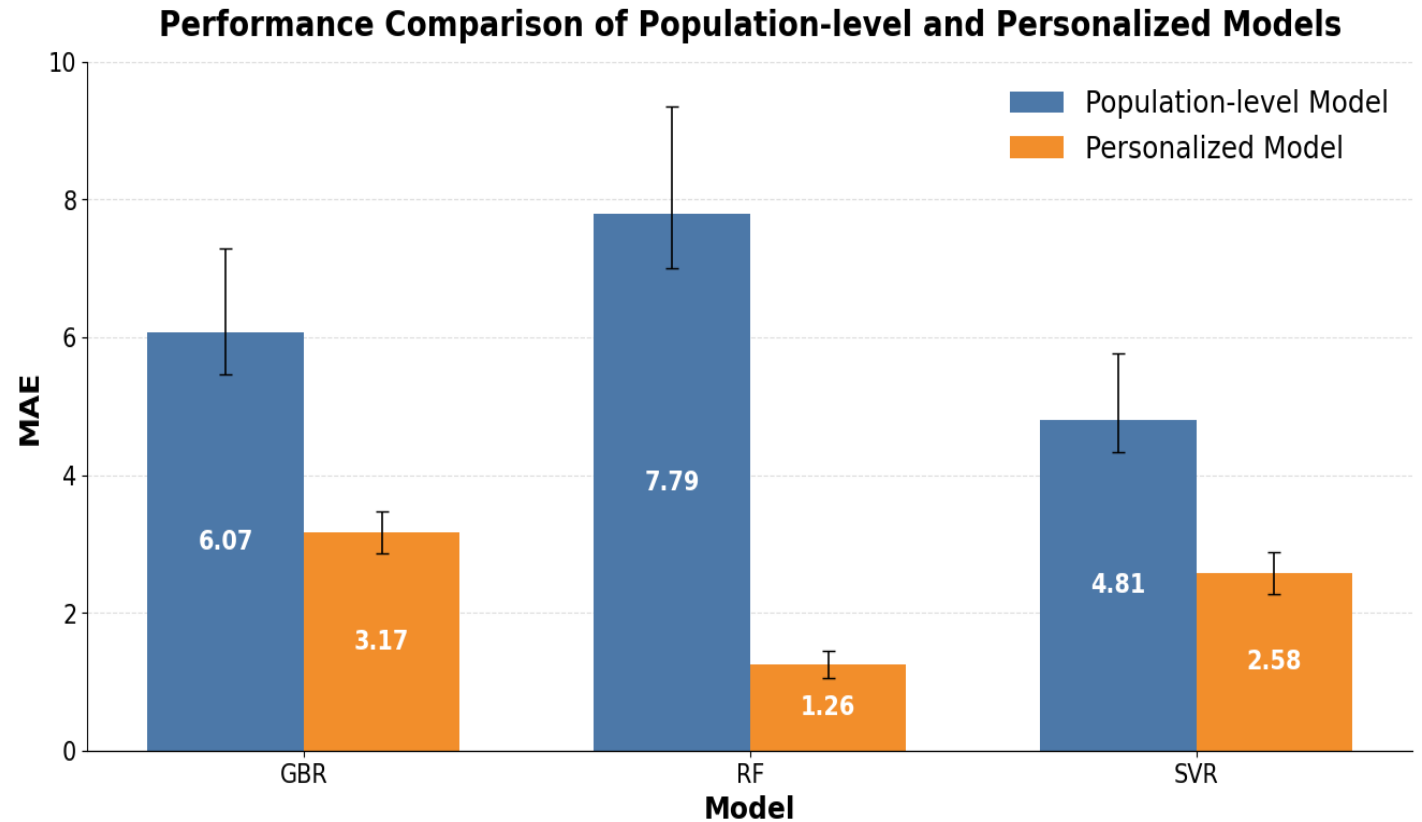
Note: GBR – gradient boosting regression, KNN – k-nearest neighbor, LR – linear regression, RF – random forest, RR – ridge regression, SVR – support vector regression

Findings

Key Findings 1 of 3

How Does Personalized Model Performance Compare With Population-level Models?

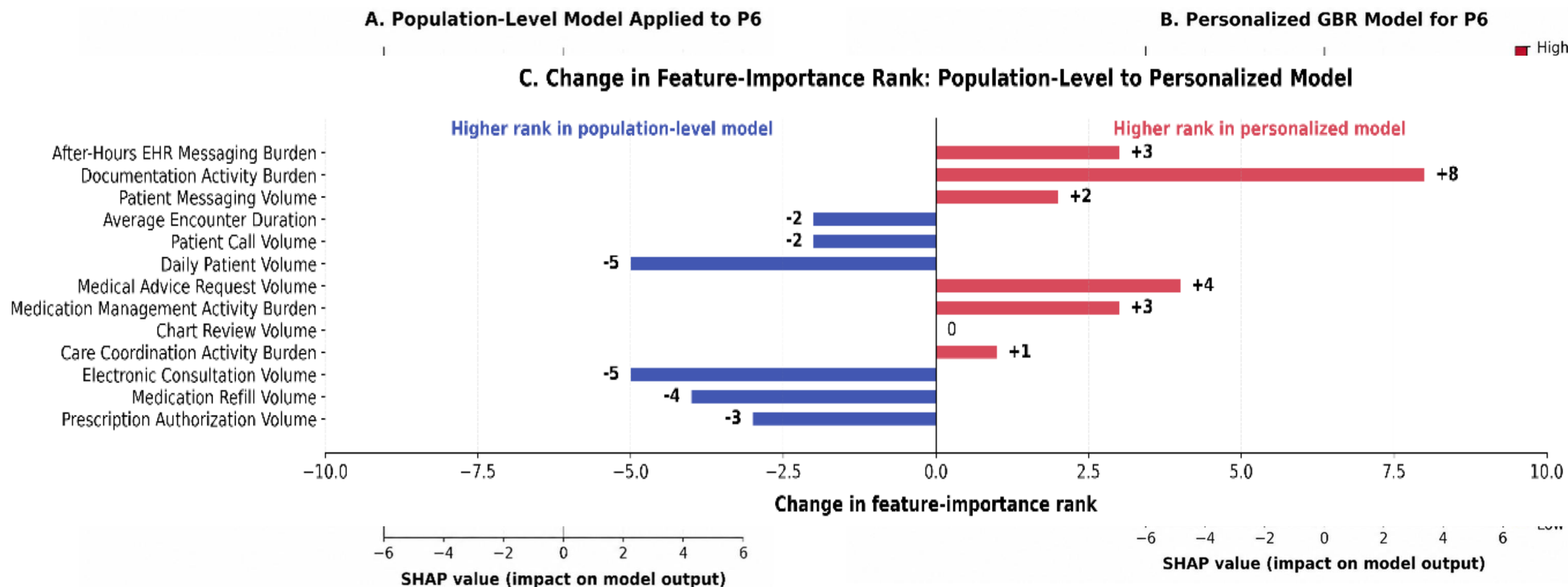
- Personalized models consistently **outperformed** population-level models
- **46-84%** reduction in mean absolute error, across all three algorithms
- Wilcoxon signed-rank gives $p < 0.05$: improvements are **statistically significant**



Note: GBR – gradient boosting regression, RF – random forest, SVR – support vector regression

Key Findings 2 of 3

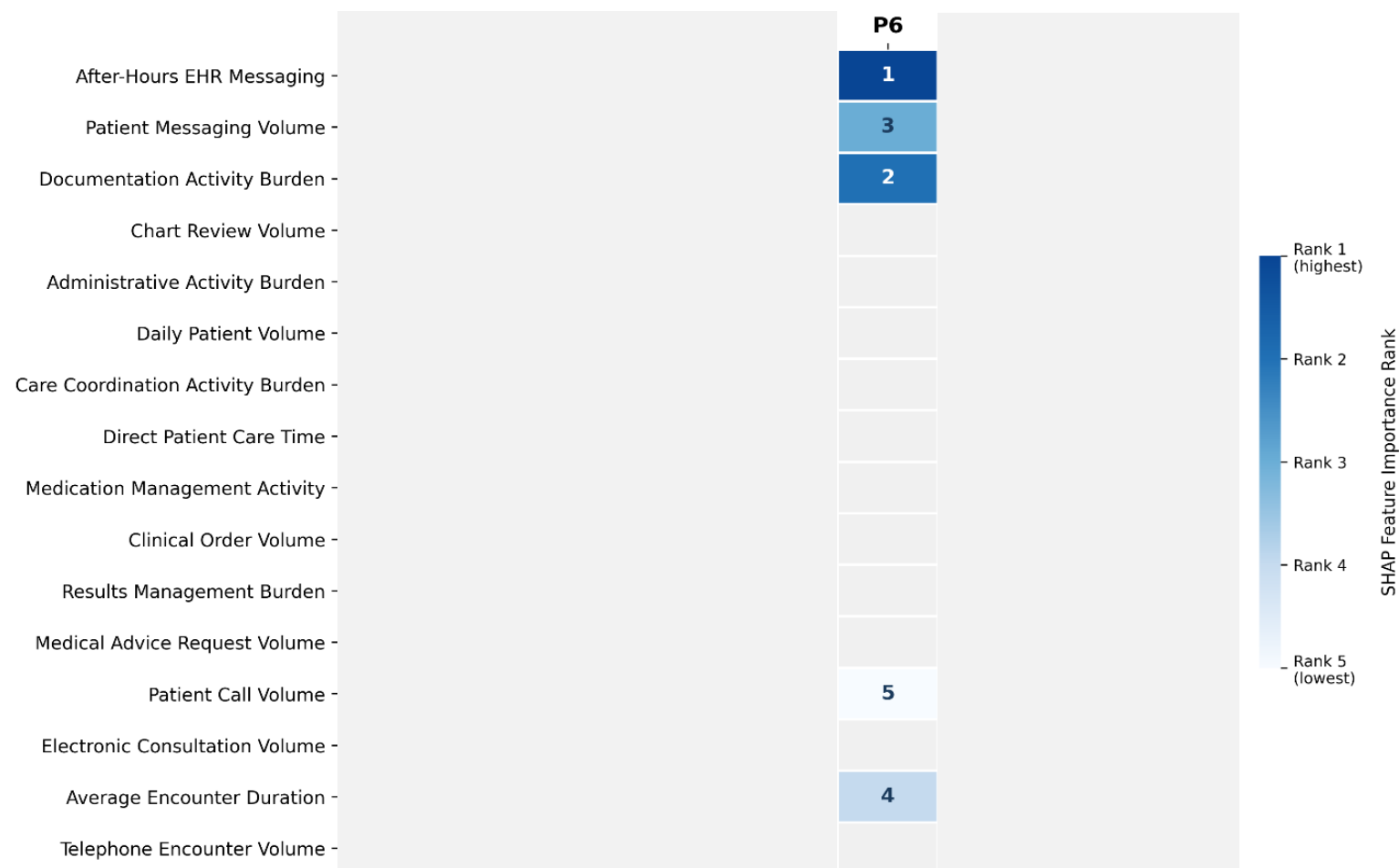
How Workload Predictors Vary in Population vs. Personalized Models?



SHAP value: a feature's contribution to the model prediction relative to the baseline; positive values increase predicted workload, while negative values decrease it.

Key Findings 3 of 3

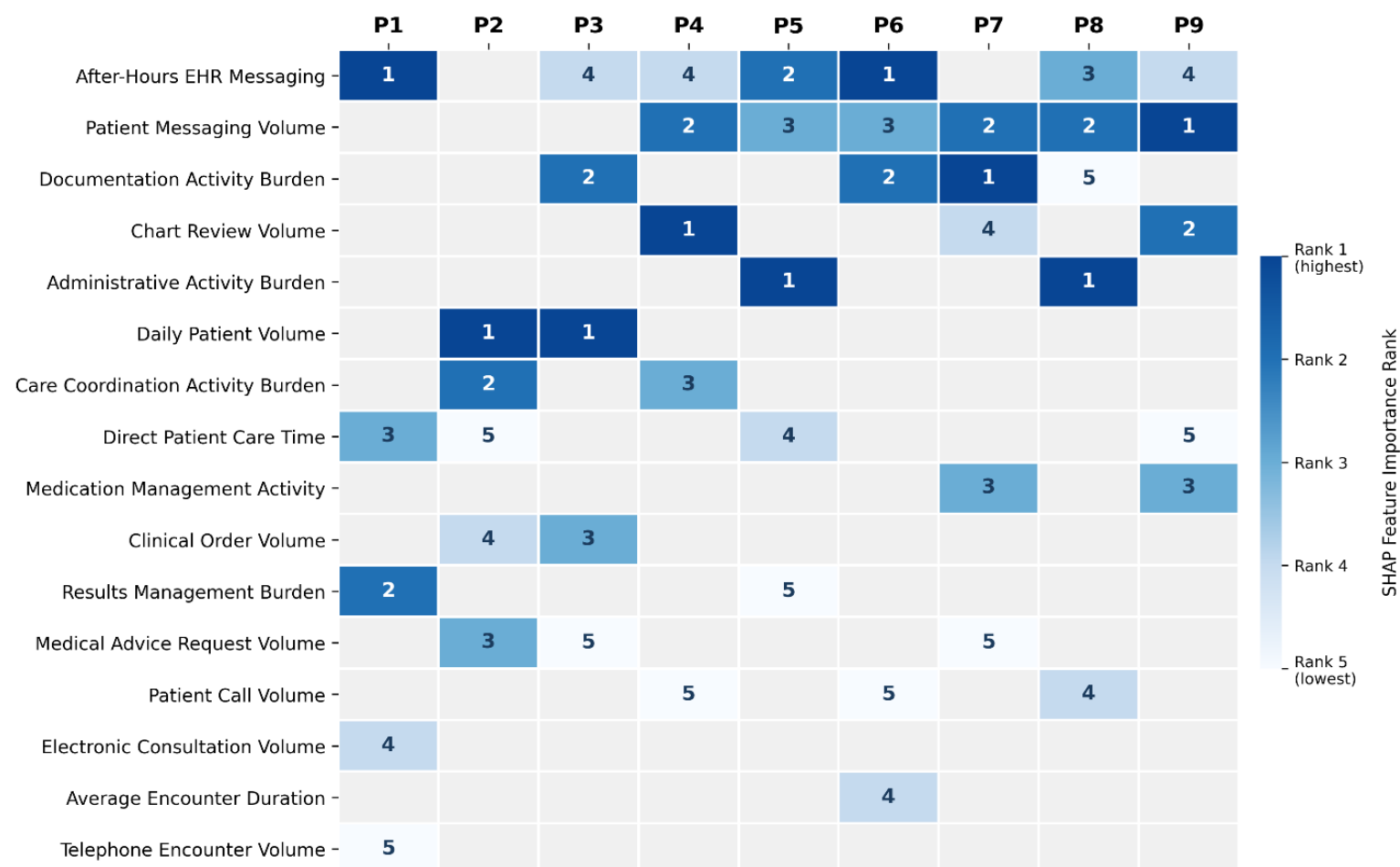
Top Workload Predictors Across Individuals



Key Findings 3 of 3

Top Workload Predictors Across Individuals

- Kendall's $W = 0.178$:
weak concordance
across operators' top-5
predictor rankings.
- Same organization.
Same role. Same EHR.
Different workload
"signature" for every
operator.



Key Takeaway

In human-in-the-loop systems, one model does NOT fit all operators.

Personalized, explainable ML recovers the individual signal that population-level models average away, a necessary foundation for adaptive, operator-aware decision support in safety-critical HITL systems.

Back to the Broader Picture: From the Clinic to Other HITLs



Aviation

Pilot workload during automation hand-offs and high-tempo phases of flight

The core lesson generalizes:

Pooling operator data to build “the” model can average away exactly the individual differences that matter most for safety-critical decision support.



Human-AI collaboration

Operator workload and trust calibration when teaming with autonomous or AI-driven agents in shared-control tasks



Any instrumented HITL system

Wherever operator logs (telemetry, HMI, comms) and periodic ground-truth workload assessments exist

Practical Implications: What This Means for Systems Engineers



Design for heterogeneity, not around it

Treat operator-to-operator variation as a system characteristic to architect for, not noise to average out in validation.



Architect an individual-adaptation layer

Don't stop at a population baseline; design monitoring and decision-support systems with a layer that adapts to each operator.



A building block for operator digital twins

Personalized workload models are a foundational state-estimation layer for real-time monitoring and decision support in HITL digital twins (Xames & Topcu, 2023).

Limitations & What's Next

LIMITATIONS

- Single healthcare system in one geographic area
- $n = 9$ participants (though longitudinal, ~2,591 operator-days)
- Limited NASA-TLX subjective workload labels
- Associations, not causal claims; this is an observational design

FUTURE WORK

- Multi-site, multi-domain validation, including defense and aviation settings
- Multimodal signals: physiological and wearable sensors data
- Adaptive/continual learning as operator behavior drifts over time
- Integration into closed-loop, operator-specific workload digital twins

Acknowledgement



INSTITUTE FOR SOCIETY, CULTURE AND ENVIRONMENT



Dr. John Epling, MD

Thank You!

Questions/Comments?



Md Doulotuzzaman Xames

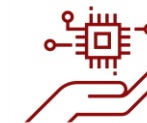
xames@vt.edu
www.mdxames.com



Google Scholar



LinkedIn



STSELab
SOCIO-TECHNICAL SYSTEMS ENGINEERING LABORATORY



Dr. Taylan G. Topcu

ttopcu@vt.edu
www.taylantopcu.com



References

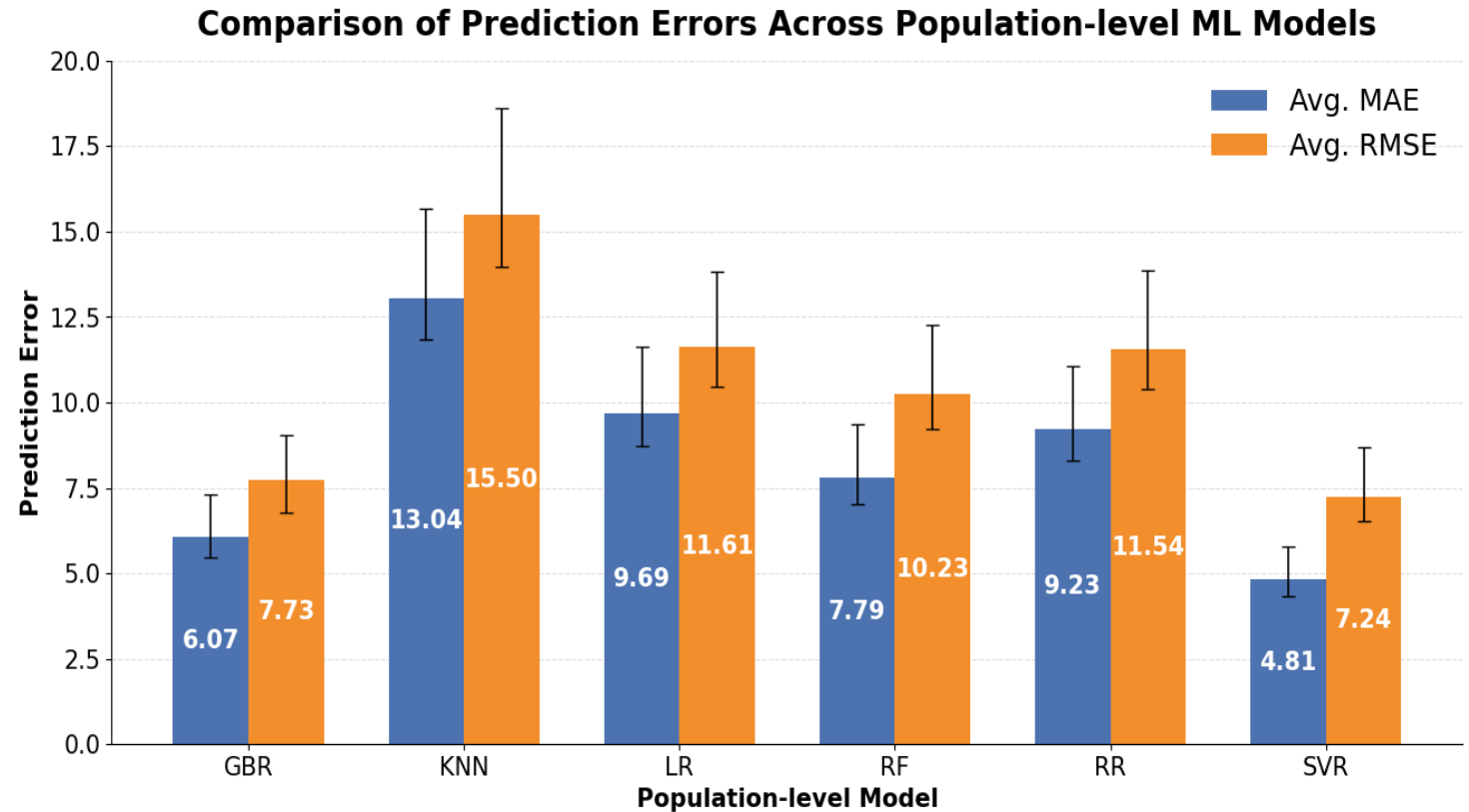
- Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with computers*, 23(1), 4-17.
- Carayon, P. (2006). Human factors of complex sociotechnical systems. *Applied Ergonomics*, 37(4), 525–535.
- Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, 42(3), 462-492.
- Fagbenle, E. O., Solomon, O., Khameneh, J. Z. & Arokodare, O. E. Forecasting Physician Burnout Risk: The Role of Electronic Health Record (EHR) and Operational Data in Predictive Models of Burnout. *Australian Journal of Biomedical Research* 1, aubm004 (2025).
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.
- Holden, R. J., Carayon, P., Gurses, A. P., Hoonakker, P., Hundt, A. S., Ozok, A. A., & Rivera-Rodriguez, A. J. (2013). SEIPS 2.0: A human factors framework for studying and improving the work of healthcare professionals and patients. *Ergonomics*, 56(11), 1669–1686.
- Jiang, Y., Huang, Y.-L., Watral, A., Blocker, R. C. & Rushlow, D. R. Predicting provider workload using predicted patient risk score and social determinants of health in primary care setting. *Applied clinical informatics* 15, 511–527 (2024).
- Lou, S. S., Lew, D., Harford, D. R., Lu, C., Evanoff, B. A., Duncan, J. G., & Kannampallil, T. (2022). Temporal associations between EHR-derived workload, burnout, and errors: a prospective cohort study. *Journal of general internal medicine*, 37(9), 2165-2172.
- Mumford, E. (2006). The story of socio-technical design: Reflections on its successes, failures and potential. *Information Systems Journal*, 16(4), 317–342.
- National Academies of Sciences, E. (2019). Taking action against clinician burnout: A systems approach to professional well-being.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 30(3), 286-297.
- West, C. P., Dyrbye, L. N., & Shanafelt, T. D. (2018). Physician burnout: Contributors, consequences and solutions. *Journal of Internal Medicine*, 283(6), 516–529.
- Xames, M. D., & Topcu, T. G. (2023). Toward digital twins for human-in-the-loop systems: a framework for workload management and burnout prevention in healthcare systems. In *2023 IEEE 3rd International Conference on Digital Twins and Parallel Intelligence (DTPI)* (pp. 1-6). IEEE.

Appendix

Key Findings 4

How Well Do Pooled Population-level Models Perform?

- **SVR** demonstrated the best overall predictive performance.
- Kernel/ensemble methods clearly beat linear models, suggesting workload is driven by **nonlinear** human-task interactions.
- **Top three** performing models: SVR, GBR, RF



Note: GBR – gradient boosting regression, KNN – k-nearest neighbor, LR – linear regression, RF – random forest, RR – ridge regression, SVR – support vector regression

NASA-TLX Workload Survey

Mental Demand How mentally demanding was the task?

Very Low Very High

Physical Demand How physically demanding was the task?

Very Low Very High

Temporal Demand How hurried or rushed was the pace of the task?

Very Low Very High

Performance How successful were you in accomplishing what you were asked to do?

Perfect Failure

Effort How hard did you have to work to accomplish your level of performance?

Very Low Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed were you?

Very Low Very High