

Submitted by: Steven Ang

Group discussion with: Boram Lee, Juho Immonen

Article Reviewed: Adams, Roy et al. 2022. "Prospective, Multi-Site Study of Patient Outcomes after Implementation of the TREWS Machine Learning-Based Early Warning System for Sepsis." Nature Medicine 28 (7): 1455–60.

- **Summarize background and goals:**

The primary goal of this study was to evaluate the relationship between the healthcare provider interaction with Targeted Real-time Early Warning System (TREWS), a machine learning sepsis detection tool, and the patient outcome. The study focused on whether timely evaluation and confirmation of TREWS alert (within three hours) improved in-patient mortality organ failure progression, and length of hospital stay for sepsis patients, specially focusing on the cases of timely antibiotic administration following TREWS alerts. Additionally, the researchers sought to determine where TREWS could effectively identify high-risk sepsis patient who would benefit most from early intervention, potentially enabling more targeted alert prioritization.

This study extends prior research by addressing critical gaps in real-world implementation and clinical validation of machine learning models for sepsis detection. While the earlier work primarily consisted of retrospective studies that demonstrated predictive accuracy and feasibility, this new research study evaluated the operational effectiveness of TREWS in live hospital environment. By deploying TREWS both academic and community settings, the researcher analyzing data from 590,736 patients across five hospitals. The study highlighted the system's scalability in decentralized workflows. Furthermore, it builds upon prior findings that underscore the importance of timely antibiotic administration in sepsis care, demonstrating how a machine learning tools like TREWS can accelerate clinical response, particularly for the high risk-patients.

The research complements existing knowledge by providing a comprehensive examination of both clinical outcomes and implementation challenges. Beyond simply measuring mortality rates, the study assessed multiple outcomes including organ failure progression and length of stay, while also analyzing these outcomes in both general sepsis populations and high-risk subgroups. This multi-faceted approach, combined with the study's focus on real-world implementation without dedicated staff, provides valuable insights into the practical application of machine learning-based sepsis detection systems in clinical settings. The research thus bridges important gaps between theoretical potential and practical implementation, offering evidence for the clinical utility of such systems while acknowledging and addressing real-world challenges that previous studies had not fully explored.

- **Describe key Attributes of the data:**

The data used for this study was sourced from the electronic health records (EHR) of five hospitals within a single health system, encompassing two academic and three community hospitals in the Maryland and DC areas. These hospitals were selected to represent diverse clinical environments, ensuring the scalability and robustness of the findings. The study period spanned from 2018 through September 2020, providing comprehensive longitudinal data for analysis.

The dataset includes overall of 590,736 adult (≥ 18 years) patient encounters monitored by TREWS, involving patients representing to emergency departments or admitted to inpatient units. Of these, total of 42,089 (7.1%) triggered a TREWS alert and only 13,680 (2.3) met the retrospective criteria for sepsis, and 6,877 were selected in the primary analysis cohort. This cohort comprised patient who triggered a TREWS alert before receiving antibiotics, ensuring that the tool's impact on treatment timing could be assessed.

The EHR data collected for the study encompassed several critical categories. Demographic information included patient age, gender and hospital admission details. Clinical measurements were extracted from vital signs such as blood pressure and heart rate, laboratory data like lactate levels and white blood cell counts, and composite severity scores, including SOFA (Sequential Organ Failure Assessment) and APACHE II. Treatment and intervention data captured the timing of antibiotic administration, use of vasopressors, mechanical ventilation, and other treatments relevant to sepsis management. Comorbidities and medical history were documented using ICD-10 codes to account for baseline health conditions such as diabetes, cancer, liver disease, and chronic obstructive pulmonary disease. Additionally, hospital-related variables, including information on the admitting hospital and care protocols, were incorporated to adjust for differences across institutions. This comprehensive data set allowed for a detailed analysis of TREWS's impact on patient outcomes while controlling for confounding factors.

The structure of the dataset was organized around patient encounters, which served as the primary unit of analysis. Each encounter represented a unique instance of a patient presenting to the emergency department (ED) or being admitted to an inpatient unit across five hospitals. This encounter-based approach allowed for the inclusion of repeated hospital visits by the same patient as separate analytical units, reflecting real-world clinical scenarios.

The data was structured to facilitate comparison between two primary cohorts: a study arm consisting of 4,220 encounters where TREWS alerts were evaluated and confirmed within three hours, and a comparison arm of 2,657 encounters lacking timely alert confirmation. A subset of 2,366 encounters was further identified as a high-risk cohort based on predicted mortality risk without timely antibiotics. Within each encounter, the data included nested elements such as sequential clinical measurements, SOFA scores (measured at alert time and 72 hours post-alert), treatment timing, outcome measures, hospital identifiers, and provider response information.

To maintain analytical rigor given this encounter-based structure, the researchers implemented sensitivity analyses to ensure their findings weren't disproportionately influenced by patients with multiple encounters. This methodological consideration was particularly crucial given that a substantial portion of the study population had multiple hospital visits throughout the study period. The encounter-based approach allowed for a more comprehensive analysis of how the TREWS system performed across different hospital presentations, while still accounting for the potential confounding effects of repeated patient visits.

The dataset collected and analyzed in this study demonstrated strong relevance to the central problem of improving early sepsis detection and treatment outcome through an automated warning system. It was evident in its temporal elements, capturing the critical timing relationships between TREWS alert generation, provider response and antibiotics administration. This timing data was crucial for evaluating system impact on early intervention, which is widely recognized as a key factor in sepsis outcomes. The inclusion of multiple clinical parameters – like vital signs laboratory values, and organ dysfunction scores, are aligned well with establish sepsis detection criteria and enabled through assessment of patient condition and disease progression.

The dataset's breadth also supported robust analysis of potential confounding factors. By including detailed patient characteristics, comorbidities, and hospital-specific variables, researchers could adjust for factors that might influence both provider response to alerts and patient outcomes. The incorporation of provider interaction data with TREWS allowed direct examination of how the system influenced clinical decision-making and subsequent patient care. Furthermore, the inclusion of multiple outcome measures - mortality, organ failure progression, and length of stay - provided a comprehensive view of the system's impact on patient care, making the data highly relevant for evaluating the effectiveness of this early warning approach to sepsis management.

This comprehensive dataset enabled researchers to not only assess the overall impact of TREWS but also identify high-risk patient populations who might benefit most from early intervention, directly addressing a significant gap in existing sepsis care protocols. The multi-site nature of the data also enhanced its relevance, allowing for assessment of the system's effectiveness across different hospital settings and patient populations.

The researcher performed several key data transformations to prepare the dataset for analysis and ensure its usability, accuracy, alignment with the study's objective and ensure meaningful comparisons. These transformations were critical to standardizing raw EHR data, enriching its analytical value, and addressing potential bias or inconsistencies.

Initial steps focused on data cleaning, including imputing missing values for clinical measurements, such as vital signs and lab results, using clinically reasonable defaults or population medians. Outliers, particularly extreme or implausible values like abnormally high lactate levels, were capped or excluded to reduce noise and maintain data validity.

Cohort selection involved applying specific inclusion and exclusion criteria to focus on actionable cases where TREWS could influence treatment timing. Patients who triggered a TREWS alert before antibiotic administration and were not admitted directly to the ICU were included, while those who did not receive antibiotics within 24 hours of the alert were excluded. Additionally, a high-risk cohort was identified using a predictive model that pinpointed patients most likely to benefit from early intervention, allowing for targeted subgroup analysis.

Temporal transformations aligned clinical events relative to the TREWS alert, standardizing the timing of interventions like antibiotic administration and post-alert SOFA score progression. This included defining temporal windows for baseline variables (24 hours pre-alert) and outcomes (72 hours post-alert). Additionally, clinical variables were transformed into composite severity scores, such as SOFA and APACHE II, integrating multiple raw measurements into clinically interpretable metrics. Continuous variables, like lactate levels and blood pressure, were categorized into thresholds reflecting clinical significance.

Derived metrics included primary and secondary outcome measures, such as in-hospital mortality, SOFA progression, and hospital length of stay, along with time-to-antibiotic administration, categorized into intervals (e.g., within 3 hours). Adjustments for confounders were incorporated through standardization across demographic, clinical, and hospital-related variables, and propensity score modeling was used to balance differences between patients who interacted with TREWS and those who did not. These transformations ensured the dataset was clean, standardized, and aligned with the study's objectives, allowing for reliable evaluation of TREWS's effectiveness in improving sepsis outcomes.

- **Describe analysis**

The analysis followed a structured workflow to assess the clinical impact of the TREWS system across 590,736 patient encounters from five hospitals. The process began with cohort selection, identifying patient encounters that met inclusion criteria, such as triggering a TREWS alert before antibiotic administration and excluding cases like ICU admissions or those without antibiotics within 24 hours. High-risk patients were identified using two separate ridge logistic regression models trained on retrospective data, predicting mortality probabilities with and without timely antibiotics. Data preparation involved cleaning, imputing missing values, creating composite scores (APACHE II and SOFA using worst measurements from specified time windows), and aligning clinical events relative to the TREWS alert to standardize temporal relationships. Variables were defined, with TREWS interaction and antibiotic timing as independent variables and outcomes like mortality, SOFA progression, and hospital length of stay as dependent variables.

Statistical modeling employed multiple regression approaches, each chosen for specific outcome types. Logistic regression was applied for the binary outcome of in-hospital mortality, allowing for estimation of adjusted risk differences and relative reductions while controlling for confounders. Linear regression modeled continuous outcomes such as changes in SOFA scores,

while quantile regression addressed skewed distributions like hospital length of stay. All models incorporated piecewise linear terms for continuous variables according to APACHE II thresholds and used heteroskedasticity-robust estimators for standard errors and confidence intervals. The models directly adjusted for numerous confounding factors, including demographics, clinical measurements, comorbidities, and hospital-specific variables.

These analytical methods aligned well with the study's objectives of evaluating TREWS's impact on sepsis care. The approach enabled detailed examination of both general and high-risk populations, while allowing comparison with previous studies on antibiotic timing through the novel use of TREWS alert as a reference point. Sensitivity analyses tested the robustness of findings under different scenarios, including alternative statistical methods (inverse probability of treatment weighting), different data sources for comorbidities, first-encounter-only analysis, and pre-COVID-19 data analysis. This comprehensive analytical strategy ensured reliable evaluation of TREWS's effectiveness in improving sepsis detection and treatment, while accounting for potential confounding factors and establishing the system's impact across different patient populations and clinical settings.

- **Critique:**

The authors interpret their findings as evidence that TREWS significantly improves sepsis outcomes by facilitating earlier recognition and treatment. They argue that timely provider interaction with TREWS alerts reduces in-hospital mortality, limits organ failure progression, and shortens hospital stays. Furthermore, the authors highlight TREWS's ability to prioritize high-risk patients for timely interventions, showcasing the system's potential to transform sepsis care without requiring dedicated staff, as it integrates seamlessly into existing workflows.

While the findings strongly support the authors' interpretation, there are limitations to their conclusions. The association between TREWS interaction and improved outcomes is compelling but cannot definitively establish causality due to the observational nature of the study. Despite adjustments for confounders and the use of propensity scores, residual confounding from unmeasured variables—such as provider expertise, institutional protocols, or concurrent interventions—could influence the results. Additionally, the exclusion of patients who did not receive antibiotics or were admitted directly to the ICU narrows the study population to those already more likely to benefit from TREWS, potentially inflating its effectiveness. On the other hand, the study's use of a large, multi-site dataset and rigorous statistical methods enhances the credibility of the results, making a strong case for TREWS's potential in improving clinical outcomes.

The study primarily focuses on antibiotic timing as a critical intervention but does not fully consider other key components of sepsis management. Interventions such as fluid resuscitation, vasopressor use, source control (e.g., surgery), and hemodynamic monitoring are essential to sepsis treatment and may confound the observed effects attributed to TREWS. For example, providers interacting with TREWS may also be more proactive in these areas, influencing outcomes.

Additionally, the appropriateness of antibiotics administered following TREWS alerts is not assessed, leaving a gap in understanding its impact on antimicrobial stewardship and resistance. Incorporating these factors in future analyses could provide a more comprehensive view of TREWS's role in sepsis care.

To fully validate the authors' interpretations, further experiments are necessary. A randomized controlled trial (RCT) comparing outcomes with and without TREWS, while accounting for site-level randomization, would provide stronger causal evidence. Additionally, studies evaluating the appropriateness of antibiotics prescribed in response to TREWS alerts could address concerns about potential overuse or resistance. Qualitative studies exploring provider decision-making and interactions with TREWS would shed light on adoption barriers and workflow integration challenges. Lastly, including data on other sepsis-related interventions, such as fluid management and source control, would help isolate TREWS's direct effects.

This study makes a meaningful contribution to advancing machine learning (ML)-based clinical decision support systems for sepsis. By demonstrating improvements in mortality, organ failure, and treatment timeliness, it validates the potential of TREWS as a scalable, real-world solution. Its decentralized approach, which does not require dedicated staff, makes it particularly relevant for resource-constrained healthcare systems. However, the lack of exploration of broader impacts, such as long-term clinical outcomes or system-wide effects on resource utilization, slightly limits its overall significance. Nonetheless, the study provides a strong foundation for future work and implementation efforts.

Several key lessons emerge from this study. First, integrating ML tools like TREWS into existing workflows requires careful consideration of provider interaction, as human-machine collaboration significantly impacts effectiveness. Second, prioritizing high-risk patients enhances the clinical relevance of alerts while minimizing alert fatigue. Third, comprehensive evaluations of sepsis care should include a holistic view of all interventions, not just antibiotics. Future studies should focus on refining alert specificity, expanding deployment to diverse healthcare settings, and incorporating qualitative insights from clinicians to improve adoption and usability.

Despite the study's strengths, several questions remain. How would TREWS perform in hospitals with varying infrastructure, staffing, and resources? What are the long-term implications of TREWS on antibiotic stewardship, resistance patterns, and overall clinical efficiency? Can TREWS be adapted to include additional ML features, such as learning from provider feedback, to further improve outcomes? Finally, how does TREWS impact other sepsis-related interventions, such as fluid management or escalation to intensive care? Addressing these questions in future research would deepen our understanding of TREWS's capabilities and guide its broader implementation.

- **Final Decision: Accept with Major Revision**

This paper provides valuable insights into the real-world use of Machine Learning in critical care, addressing a pressing need for better sepsis detection tools. However, revisions are needed to strengthen causal claims, explore unanswered questions, and ensure transparency. By addressing these concerns, the paper would make a meaningful contribution to both clinical and machine learning fields.