Predicting the Predictable in the Psychiatric High Risk

Eric V. Strobl, MD, PhD

Department of Biomedical Informatics, University of Pittsburgh

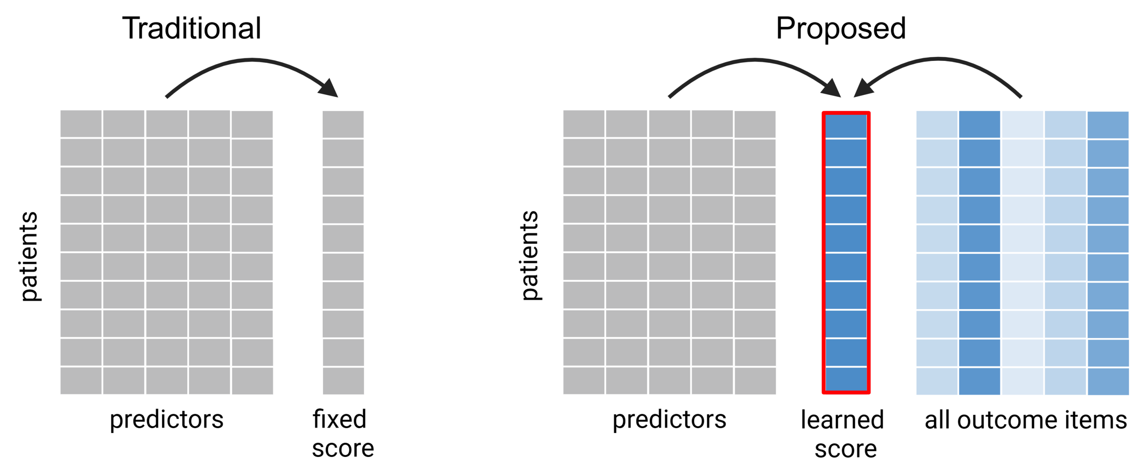
Full paper:

Abstract

- **Problem:** Most investigators in precision psychiatry force models to predict clinically meaningful but ultimately predefined outcomes in a high-risk population, limiting accuracy.
- **Approach:** We propose to *let the data reveal* which symptoms are predictable with high accuracy and then assess whether those predictable symptoms warrant early intervention.
- **Method:** We introduce the *Sparse Canonical Outcome REgression (SCORE)* algorithm, which combines items from clinical rating scales into composite severity scores that are maximally predictable across time.
- **Findings:** This data-driven approach substantially improves prognostic accuracy, identifying interpretable symptom profiles that are both predictable and clinically actionable.
- **Impact:** The resulting scores diverge from conventional metrics, equipping clinicians with memorable, actionable insights for early intervention – even before full diagnostic criteria are met.
- **Open Source:** An R implementation is available: `https://github.com/ericstrobl/SCORE`

Introduction

- **High-risk patients:** Individuals with elevated risk for mental illness due to, e.g., early symptoms, genetic predispositions, or environmental exposures.
- **Prediction challenge:** Accurately forecasting which high-risk patients will develop mental illness remains unreliable, delaying intervention and straining mental health resources.
- **Prevailing optimism:** Many expect that more data and complex models will yield high accuracy – mirroring ML tasks (e.g., object recognition) known to be predictable a priori, so accuracy primarily hinges on sufficient model complexity.
- **Inherent limits:** In psychiatry, complex models and features have yielded limited gains likely because some outcomes may be fundamentally unpredictable, regardless of approach.
- **Reframing the problem:** Instead of forcing prediction of a fixed outcome, models should first identify which outcomes are inherently predictable; clinicians can then interpret the learned targets post hoc to guide actionable interventions.
- **Sparsity and interpretability:** Instead of predicting hundreds of outcomes individually, learn a few sparse, composite scores that leverage multivariate relationships by aggregating statistically dependent outcomes, boosting both predictive power and clinical interpretability.



Prioritizing Patient Differentiation via Correlation

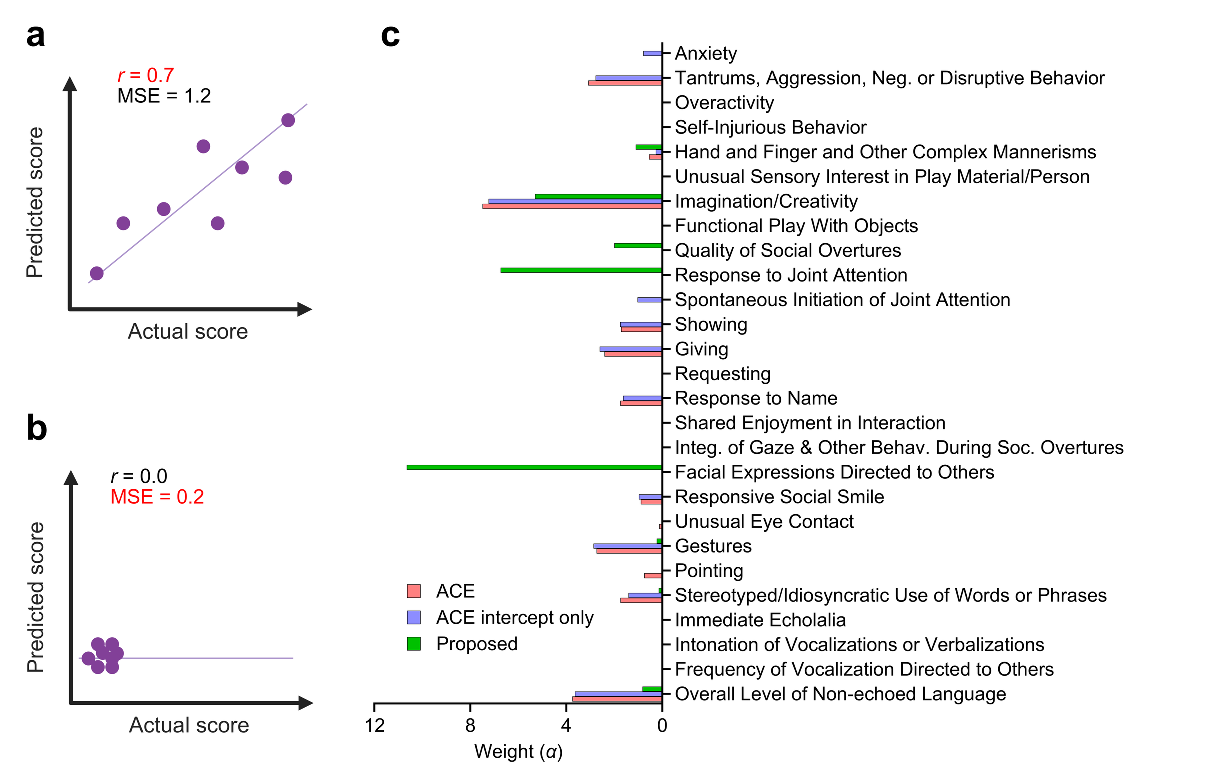


Figure 1. **Correlation vs. MSE.** (a) **Maximizing correlation** ($r = 0.7$) separates patients along the diagonal, allowing effective risk stratification—even with substantial variability in the actual scores (but higher MSE: 1.2). (b) **Minimizing MSE** (to 0.2) reduces overall error but prefers stable scores, failing to differentiate patients ($r = 0$). (c) In autism risk prediction, ACE (minimizing MSE) yields weights nearly identical to an intercept-only model (blue), while maximizing correlation recovers very different weights (green).

Sparse Canonical Outcome REgression (SCORE)

- SCORE constructs sparse, maximally predictable composite scores by learning **non-negative weights** over clinical rating scale items, producing interpretable severity scores that distinguish trajectories in high-risk populations.
- **Objective:** Maximize the **correlation** between baseline predictors X and composite outcome scores across multiple time points:

$$\arg\max_{\alpha \geq 0, \beta} \sum_{i=1}^{m} \text{cor}(Y(t_i)\alpha, X\beta(t_i))$$

- **Interpretability and sparsity:** Non-negativity on α ensures composite scores represent symptom severity and promotes sparsity for clinical interpretability.
- **Optimization:** Blockwise coordinate ascent, where we alternate updates of non-negative outcome weights (α) and time-specific predictor weights (β).
- **Full extraction:** SCORE sequentially identifies all maximally predictable composite scores by removing variance explained by prior predictions, ensuring every composite outcome predictable from X is discovered.
- **Comparison to CCA:** Unlike CCA, SCORE accommodates longitudinal data by maximizing correlation across multiple time points. It also extracts multiple scores through iterative deflation – without conflating mathematical uncorrelatedness with clinical distinctiveness.
- **Efficiency and robustness:** Computationally efficient, at most quadratic in the number of predictors; robust to missing data under a **Missing At Random (MAR)** assumption.

Results

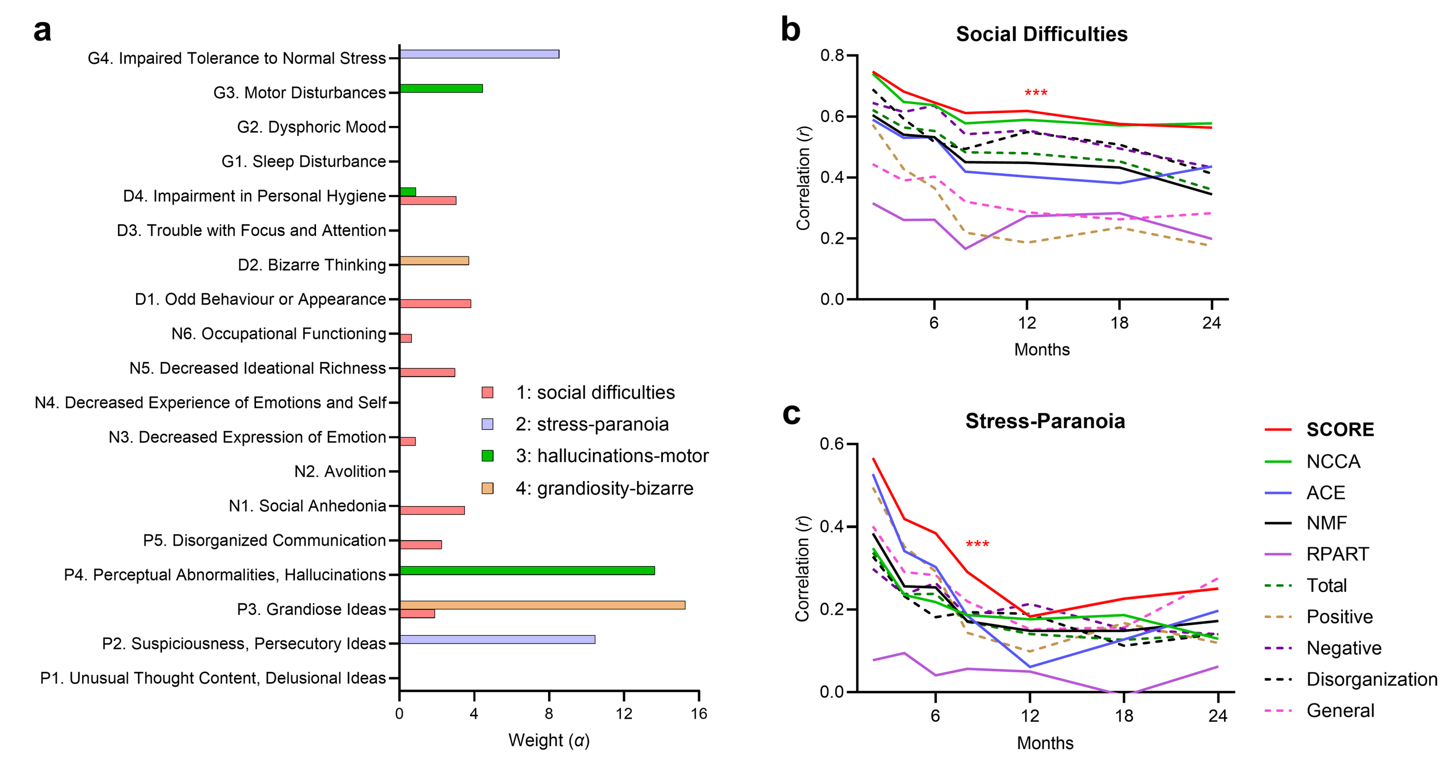


Figure 2. **Predicting psychotic symptoms in high-risk youth over two years.** (a) Sparse weights for the top 4 severity scores discovered by SCORE. (b) SCORE outperforms all comparators for the first score at nearly all time points (except NCCA at 24 months). (c) Similar superiority for the second score. Similar results for the third and fourth scores in the Appendix. Confidence intervals are narrow and not visible due to the large number of bootstrap samples. $^{***}p < 0.0005$ vs. nearest competitor. See the full paper (QR code) for an additional application to autism with similar results.

Conclusions

- **SCORE identifies all severity scores** that best distinguish future trajectories in high-risk patients, maximizing predictability through correlation.
- **Multi-score extraction:** SCORE iteratively extracts multiple clinically meaningful severity scores – enabling nuanced differentiation, even in challenging domains such as psychosis and autism.
- **Novel clinical insights:** The resulting scores frequently highlight interpretable profiles (e.g., social difficulties, stress-paranoia) that differ from conventional clinical total or subscores.
- **Prevention and early intervention:** These interpretable composite scores can guide earlier or preventive strategies (e.g., social skills training, CBT) before diagnostic thresholds are reached.
- **Practical implementation:** Static and interpretable, SCORE's composite scores are precomputable, memorable, and require no real-time computation – making them especially suitable for resource-constrained clinical settings.
- **Limitations and future directions:** Current limitations include reliance on linear combinations and focusing on slowly changing conditions. Future work could add non-linear methods or adaptive scores that better track rapid changes over time.
- **Summary:** SCORE offers a principled, actionable, and fine-grained approach to risk stratification, illuminating subtle symptom profiles and supporting timely, proportionate clinical responses.