Reward the Reward Designer: Making Reinforcement Learning Useful for Clinical Decision Making

Reinforcement Learning (RL) is a promising framework for sequential decision-making in healthcare, but clinical adoption depends as much on reward design as on algorithms. Unlike games, healthcare lacks natural reward signals, so rewards must be carefully crafted, validated, and interpreted. Reward functions ultimately determine whether RL policies align with clinical objectives, and mis-specified rewards, already problematic in domains like autonomous driving [Knox et al., 2023], carry far greater risks in healthcare. Yet without shared practices, the knowledge of reward design remains fragmented across tasks. In this position paper, we argue that reward engineering is the key bottleneck for applying RL in clinical time series, making consolidated guidelines essential for safe deployment. Drawing on case studies, we propose practical reward design principles and a cultural shift: reward functions should be benchmarked, published, and recognized as research contributions in their own right to accelerate safe and effective use of RL in healthcare.

Guidelines. Reward design in healthcare is challenging due to physiological complexity, critical outcomes, and ethical limits on exploration. Still, many clinical tasks share structural patterns that can guide reward construction. Unfortunately, this knowledge is fragmented and often buried in broader methodological details. We consolidate these insights, outline common approaches, and propose concrete guidelines on when each should be applied.

1. **Ground Truth.** Use labeled data as the primary reward signal, whether sparse or dense.
2. **Clinical Outcome and Zones.** Terminal rewards can reflect outcomes like survival or discharge; intermediate zone-based rewards keep physiological variables within safe ranges.
3. **Change-Based Signals.** Define intermediate rewards based on biomarker improvements or deterioration.
4. **Patient Satisfaction and Engagement.** Incorporate patient-reported feedback (e.g., comfort, communication, wait times) and engagement signals in mobile health.
5. **Personalization.** Adapt rewards to individual baselines and risks rather than one-size-fits-all targets, since biomarkers vary between healthy individuals and patients with comorbidities.
6. **Treatment Efficiency.** Penalize unnecessary costs, delays, and resource waste to reflect system efficiency.
7. **Safety.** Discourage actions that violate safety constraints or increase risk, emphasizing long-term safety.
8. **Long-term Implications.** Encode not only immediate success but also downstream clinical risks that may not surface during training.
9. **Normalization.** Scale rewards (e.g., [0, 1] or [-1, 1]) to improve stability and convergence.
10. **Evaluation.** Unlike traditional RL, cumulative return alone is inadequate as an evaluation metric in healthcare. Policies must also be judged against diverse clinical objectives, combining biomarkers with qualitative assessments to detect misalignment or reward hacking. Fair evaluation should include surrogate metrics aligned with clinical goals but independent of the training reward.

Large Language Models (LLM) in Reward Design. LLMs can assist by summarizing guidelines, suggesting metrics, or flagging loopholes, but their outputs are prompt-sensitive, costly, and lack domain expertise. Hence, they should support, not replace, human designers, who remain responsible for defining objectives and ensuring safety.

Benchmarks. Benchmarks have long advanced science by enabling fair model comparisons, from ImageNet to MMLU, LiveBench, and Humanity's Last Exam. Reward modeling for aligning LLMs underscored the need for robust evaluation, leading to dedicated benchmarks such as RewardBench, M-RewardBench, and VL-RewardBench, which standardized evaluation while exposing vulnerabilities like reward hacking.

Position. Healthcare urgently needs standardized evaluation of reward functions, as the current lack slows progress, fragments efforts, and undermines clinical trust. Early steps, such as HealthBench, focus on LLMs, but the same rigor must also extend to clinical decision-making tasks. Reward functions should be published, critiqued, and benchmarked transparently, and their design recognized as a primary research contribution. In short, safe and scalable RL in healthcare begins by rewarding the reward designer.

References

W. Bradley Knox, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone. Reward (mis)design for autonomous driving. *Artificial Intelligence*, 316, 2023.