Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD30217

# *Multi-Armed Bandits and Clinical Medicine: A Survey of Algorithms, Evaluation, and Applications*

**Xiwen Guo**

*College of Arts and Sciences, The Ohio State University, Columbus, USA*
*guo.2263@osu.edu*

***Abstract.*** Personalized medicine and adaptive clinical trials aim to match optimal treatment plans to individual patients while validating new therapies. Traditional fixed-design trials have limitations, including resource wastage and scarce data. Artificial intelligence has led to the development of dynamic decision algorithms like the Multi-Armed Bandit (MAB) algorithm, which balances exploration and exploitation in treatment allocation. Researchers aim to integrate MAB with clinical needs while adhering to ethical guidelines. This paper discusses the use of Multi-armed Bandit (MAB) algorithms in medicine, highlighting their potential for optimizing treatment allocation and improving patient outcomes, despite ethical constraints and limited data, and their application in contextual and reinforcement learning settings. This research highlight key clinical applications such as adaptive dose-finding, personalized treatment selection, and digital health interventions, supported by both trial-based data and large-scale public datasets like MIMIC-III. Simulation studies are also discussed as a necessary complement to real-world data, facilitating algorithm validation under ethical and logistical constraints. Comparative evaluation of algorithms demonstrates that Bayesian methods, particularly Thompson Sampling and contextual bandits, often provide a more robust balance between efficiency and safety. However, challenges remain in scalability, interpretability, and regulatory acceptance. This research conclude by identifying promising directions for future research, including the integration of deep reinforcement learning and causal inference, which may further enhance the role of MABs in advancing personalized medicine and adaptive clinical trial design.

***Keywords:*** Multi-Armed Bandits, Clinical Medicine, Adaptive Algorithms, Reinforcement Learning, Decision-Making

## 1. Introduction

In recent years, medical researchers have increasingly focused on applying MAB algorithms, aiming to integrate them with clinical needs to address practical challenges like personalized treatment and dose exploration, while adhering to stringent ethical guidelines and data constraints in healthcare settings. Clinical medicine is defined by the need for frequent decision-making under conditions of uncertainty. In this context, physicians and researchers must identify optimal diagnostic strategies, treatment regimens, or intervention timings—often without full knowledge of how individual patients will respond. Traditional approaches, such as randomized controlled trials or rule-based

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD30217

clinical guidelines, provide important evidence but often lack the flexibility to dynamically adapt to new information as it emerges [1]. In practice, medical decision-making requires a delicate balance between exploring untested options and exploiting existing evidence to maximize patient benefit and minimize harm. The multi-armed bandit (MAB) problem offers a mathematical framework for navigating this exploration–exploitation trade-off [2]. Originally studied in statistics and reinforcement learning, the MAB formalism provides adaptive algorithms that can learn from sequential outcomes while maintaining theoretical guarantees on performance. By allocating resources to actions that appear promising yet retaining the capacity to experiment with alternatives, bandit algorithms provide a natural fit for the iterative and uncertain nature of medical practice. MAB methods are increasingly being applied in healthcare, optimizing clinical trials, reducing patient exposure to unsafe treatments, and accelerating intervention identification. They also enable real-time treatment tailoring and dynamic intervention recommendation in digital and mobile health.

This survey seeks to deliver a comprehensive overview of recent advances in multi-armed bandit methodology tailored to clinical contexts. It reviews core algorithmic developments, practical implementations, and domain-specific challenges are reviewed, with emphasis on safety, fairness, interpretability, and ethical considerations. By synthesizing these perspectives, the survey aims to clarify the opportunities and limitations of MABs in clinical research and identify open directions for future work [3].

## 2. Literature review

The literature on multi-armed bandit algorithms in clinical medicine includes early-phase adaptive trials, personalized digital health interventions, and dose-finding studies. Thompson Sampling is the most commonly used algorithm, while UCB and ε-greedy are used in simulation and safety-constrained trials. Contextual bandits are trending towards personalized medicine, integrating patient covariates to support more informed decision-making. Bandit-based designs can reduce treatment exposure, identify optimal interventions, and enhance patient engagement in mobile health platforms. However, limitations include computational complexity and challenges in real-world implementation [4].

## 3. MAB algorithms and clinical relevance

Multi-armed bandit (MAB) algorithms offer a systematic framework for sequential decision-making, striking a deliberate balance between exploring untested options and exploiting interventions with established efficacy. In clinical medicine, this framework allows adaptive allocation of treatments or interventions while minimizing patient risk and improving overall outcomes. The ε-greedy is a simple yet effective algorithm that chooses the currently best-known action most of the time, while exploring a random action with a small probability ε. In clinical settings, ε-greedy could be applied to adaptive treatment trials where a limited number of patients are exposed to novel therapies while maintaining a majority allocation to proven treatments [5]. Upper Confidence Bound (UCB) algorithms select actions based on both estimated reward and an uncertainty term, encouraging exploration of less-tested options. UCB has been specifically proposed for dose-finding studies, where safety and efficacy need to be balanced dynamically, ensuring that promising doses are investigated without exposing excessive numbers of patients to risky levels [6]. Thompson Sampling (TS) is a Bayesian approach used in adaptive clinical trials to make probabilistic action choices, while Contextual Bandits (CB) extend MABs by integrating

DOI: 10.54254/2755-2721/2025.LD30217

patient-specific information, particularly in personalized medicine, and has been successfully applied in mobile health interventions [7].

Reinforcement Learning (RL) extensions generalize bandit algorithms to sequential decision processes with delayed rewards. RL methods are uniquely suited for multi-stage treatment strategies, such as chronic disease management, where decisions at one stage influence outcomes at subsequent stages. These extensions enable modeling of long-term patient benefits beyond immediate treatment effects.

## 4. Applications in clinical medicine

MAB algorithms have demonstrated substantial potential in several areas of clinical medicine. A key application lies in adaptive clinical trials, where bandit-based allocation strategies optimize patient outcomes while upholding rigorous statistical standards. For example, dose-finding trials often use Thompson Sampling or UCB variants to adjust doses in real time, reducing patient exposure to suboptimal or unsafe treatments [5,6]. Building on this, in personalized treatment allocation, contextual bandits enable the integration of patient-level covariates such as age, genetic markers, or comorbidities. This is particularly valuable in precision medicine, where treatments may show heterogeneous effects across subgroups. For instance, in diabetes management, contextual bandits can optimize glucose control strategies by tailoring recommendations for insulin dosage or lifestyle interventions to individual patient profiles. The capacity to adapt treatment in near real-time positions bandits as a natural fit for chronic disease management.

Similarly, dose-finding studies in phase I/II clinical trials represent another critical domain. Conventional 3+3 trial designs are statistically inefficient and expose patients to suboptimal doses. By contrast, bandit-based methods allow continuous learning about toxicity and efficacy while efficiently converging on a maximum tolerated dose. Recent work has shown that such designs can both shorten trial duration and enhance safety outcomes. Digital health interventions offer a promising environment for MAB applications, particularly in mobile health, where Bandit algorithms outperform static randomized testing for real-time evaluation. A notable example is the use of bandit-driven recommendation systems that have been piloted in mental health apps to adapt therapeutic exercises dynamically, resulting in improved adherence compared with one-size-fits-all approaches. Together, these applications illustrate how MABs bridge theoretical efficiency with clinical impact.

## 5. Datasets in clinical applications

The implementation of multi-armed bandit (MAB) algorithms in clinical research is heavily contingent on access to high-quality datasets—resources that must strike a balance between information richness and accessibility.. One of the most widely utilized repositories is MIMIC-III, a large critical care database containing detailed patient demographics, vital signs, and treatment histories. Its comprehensiveness has enabled reproducible experimentation in adaptive treatment allocation and dose-response optimization, making it a foundational resource for methodological research. More recent expansions, such as MIMIC-IV extend coverage to longitudinal hospital data, creating new opportunities to evaluate contextual bandit (CB) frameworks—tools that inherently rely on temporal patient data streams to inform decisions.

Beyond critical care databases, specialized clinical trials have provided structured environments for MAB applications. Villar, Bowden, and Wason demonstrated how adaptive randomization using Thompson Sampling can be assessed within cancer trial datasets, highlighting gains in efficiency

DOI: 10.54254/2755-2721/2025.LD30217

compared to fixed randomization [2]. Similarly, Aziz, Kaufmann, and Riviere employed dose-finding data to validate bandit algorithms under ethical constraints, with a dual focus on safeguarding patient safety and maintaining statistical robustness—two non-negotiable priorities in clinical research [6].

Digital health platforms have also emerged as increasingly important sources of real-world data. Recent work in mobile health, such as the CAREForMe project and adaptive just-in-time interventions, has shown how streaming patient data can support contextual bandits for personalized engagement [8]. These data environments differ from traditional electronic health records, as they allow continuous feedback loops that align naturally with the bandit learning paradigm.

Nevertheless, significant challenges to MAB implementation in clinical settings persist. Access restrictions, privacy considerations, and variability in data quality can compromise the generalizability of findings. Moreover, clinical datasets often contain missing or biased entries, requiring careful preprocessing and robust statistical approaches to avoid algorithmic unfairness [3]. Future progress may depend on combining open-access repositories like MIMIC with federated learning approaches that preserve privacy while enabling large-scale bandit evaluation. In this sense, datasets are not merely testing grounds for algorithms but critical enablers of whether MAB applications can be translated into equitable, reliable clinical practice.

## 6. Evaluation and comparison

The evaluation of multi-armed bandit (MAB) algorithms in clinical contexts is multidimensional, involving statistical performance, ethical considerations, and practical feasibility. Traditional evaluation metrics such as cumulative regret (a measure of missed optimal outcomes) and convergence speed (how quickly an algorithm identifies effective actions) remain foundational for comparing algorithmic performance. In cancer trial simulations, Villar, Bowden, and Wason demonstrated that Thompson Sampling consistently minimized patient exposure to ineffective treatments compared to fixed allocation, achieving lower regret while preserving statistical validity [2]. Building on this work, Aziz, Kaufmann, and Riviere extended this analysis to dose-finding studies, showing that bandit-based allocation strategies can maintain patient safety while improving trial efficiency [6].

Comparative analyses of MAB algorithms underscore context-dependent advantages, with no single method achieving superiority across all clinical scenarios. For instance, ε-greedy strategies are simple and computationally efficient, yet tend to underperform in high-stakes medical trials where early mistakes have ethical implications [6]. UCB algorithms, by contrast, provide explicit confidence bounds that serve as a conservative safeguard, rendering them particularly applicable to trials requiring cautious exploration. Thompson Sampling has emerged as the most versatile, striking a balance between exploration and exploitation across clinical domains. This versatility has been noted not only in adaptive oncology trials but also in digital health interventions, where Huch et al. found it supported robust adaptation in noisy, heterogeneous patient populations [8].

Recent advancements in contextual and robust bandits offer additional evaluative perspectives. The CAREForMe framework illustrated that contextual bandits outperform non-contextual counterparts in mental health recommendation tasks by tailoring decisions to patient-specific features [7]. Similarly, the Robust Multi-armed Bandit (RoME) approach incorporates mixed-effects modeling to account for inter-patient variability, demonstrating that robustness to demographic or clinical heterogeneity is essential for real-world deployment in mobile health (mHealth) settings. These advances underscore that evaluating algorithms solely on regret overlooks critical dimensions such as fairness, personalization, and stability.

DOI: 10.54254/2755-2721/2025.LD30217

Evaluation should prioritize fairness and equity, as traditional performance metrics may overlook disparities across demographic or clinical subgroups. This is especially important in digital health applications where patient engagement varies by socioeconomic and behavioral factors. The most appropriate MAB algorithm depends on clinical context, dataset characteristics, and ethical priorities. Evaluation should evolve beyond regret minimization to ensure safe and equitable healthcare outcomes.

## 7. Conclusion

Multi-armed bandit algorithms have emerged as a transformative methodology for clinical decision-making, offering a structured means of adapting treatment allocation in real time. Unlike conventional randomized trials, which often expose patients to suboptimal or unsafe treatments, MABs balance exploration and exploitation to accelerate the learning process while protecting participants. This makes them particularly well-suited for dose-finding studies, rare disease trials, and personalized digital health interventions, where efficiency and ethics are critical. Different MAB algorithms bring distinct strengths to clinical applications, with no universal "one-size-fits-all" solution. Thompson Sampling (TS) stands out for its ability to balance patient safety and trial efficiency, leveraging Bayesian probabilistic modeling to adaptively allocate resources; this versatility has made it increasingly favored in diverse clinical contexts, from oncology trials to mental health interventions. Contextual bandits extend these frameworks by incorporating patient-level covariates, making them especially relevant for precision medicine. Moreover, reinforcement learning generalizations expand applicability to multi-stage, sequential decision processes such as chronic disease management. Challenges in data privacy, ethical considerations, delayed outcomes, and limited sample sizes impede the implementation process of MABs. Future research should combine MABs with deep reinforcement learning, integrate causal inference, and design privacy-preserving approaches.

In sum, MAB algorithms offer a promising pathway to more adaptive, data-driven, and patient-centered medicine. With continued advances in methodology, integration with real-world health data, and attention to ethical safeguards, they have the potential to substantially reshape the design of clinical trials and the delivery of personalized healthcar.

## References

[1] Villar, S. S., & Jacko, P. (2022). Bandit Procedures for Designing Patient-Centric Clinical Trials. In Springer series in supply chain management (pp. 365–389). https: //doi.org/10.1007/978-3-031-01926-5_14

[2] Villar, S. S., Bowden, J., & Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. Statistical Science, 30(2). https: //doi.org/10.1214/14-sts504

[3] Killian, J. A., Jain, M., Jia, Y., Amar, J., Huang, E., & Tambe, M. (2024). New approach to Equitable Intervention planning to improve engagement and outcomes in a digital health Program: Simulation study. JMIR Diabetes, 9, e52688. https: //doi.org/10.2196/52688

[4] Keyvanshokooh, E., Zhalechian, M., Shi, C., Van Oyen, M. P., & Kazemian, P. (2025). Contextual Learning with Online Convex Optimization: Theory and Application to Medical Decision-Making. Management Science. https: //doi.org/10.1287/mnsc.2019.03211

[5] Kojima, M. (2022). Application of Multi-Armed Bandits to Model-assisted designs for Dose-Finding Clinical Trials. arXiv (Cornell University). https: //doi.org/10.48550/arxiv.2201.05268

[6] Aziz, M., Kaufmann, E., & Riviere, M.-K. (2021). On Multi-Armed Bandit Designs for Dose-Finding Clinical Trials. Journal of Machine Learning Research, 22(14), 1–38.

[7] Yu, S., Nourzad, N., Semple, R. J., Zhao, Y., Zhou, E., & Krishnamachari, B. (2024). CAREForMe: Contextual Multi-Armed Bandit Recommendation Framework for Mental Health. In Proceedings of the 11th IEEE/ACM

DOI: 10.54254/2755-2721/2025.LD30217

International Conference on Mobile Software Engineering and Systems (MOBILESoft '24). ACM. https://doi.org/10.1145/3647632.3647993

[8] Huch, E. K., Shi, J., Abbott, M. R., Golbus, J. R., Moreno, A., & Dempsey, W. H. (2024). RoME: A Robust Mixed-Effects Bandit Algorithm for Optimizing Mobile Health Interventions. In Advances in Neural Information Processing Systems (NeurIPS 2024). Available as arXiv preprint arXiv: 2312.06403.