Multi-Agent Reinforcement Learning in Value-Based Care

Sophia Martinez
Asst Professor

Abstract

Agentic artificial intelligence (AI) systems, including those realized through multi-agent reinforcement learning (MARL), are emerging in multiple domains, including information filtering, navigation, gaming, and robotics. These novel approaches may also be relevant to healthcare delivery and policy, as MARL is well suited to complex strategic problems that Richard Thaler characterized as "makeshift solutions to impossibly complex problems." Population health management is such an arena, particularly in optimizing risk- and performance-based contracts, payment redistribution across agents and cadre types, and performance-related benchmarks for an actor-network consisting of several competing provider-organizations functioning in a shared ecosystem. Health information systems, including such agentic, MARL-based systems, could help realize "better, faster, cheaper, and 'kinder' healthcare" by improving risk-adjusted health and cost outcomes across multiple resident cohorts, thereby more effectively realizing the goals of modern econometrics in medicine and public health.

Current culture tends to represent healthcare decision-making as a Black Box or an informationally centralized "big brain" where supervision assures correctness and compliance with Thaler's ideal of omniscience for coordinating activities. Yet actual healthcare organizations are not Black Boxes, much less a single Big Brain, corporate or otherwise. Decision makers are often overpromoted, misinformed, late in their decisions, veto the best ideas of the organization, or are unable to manage rich information effectively. Supporting such overloaded Decision Makers-MD, using training and expertise gained trying to wisely harvest large but incomplete information- is a hard problem. Because the decisions should be seen as Agents making active decisions, it is viewed in terms of Agent-based Systems, or the study of multi-oriented organizations as complex adaptive systems.

Keywords: Agentic Artificial Intelligence, Multi-Agent Reinforcement Learning (MARL), Population Health Management, Risk- and Performance-Based Contracts, Healthcare Payment Redistribution, Actor–Network Healthcare Models, Complex Adaptive Healthcare Systems, Agent-Based Decision Support, Health Information Systems Innovation, Risk-Adjusted Outcome

Optimization, Healthcare Econometrics, Decentralized Healthcare Governance, AI-Driven Policy Optimization, Strategic Interaction in Healthcare Ecosystems, Adaptive Multi-Agent Systems in Medicine.

1. Introduction

Success and sustainability of successful Models of Health Care delivery and Health Policy rests on redefining the Gartner Hype Cycle of health IT adoption in Payer and Provider organizations by leveraging recent research advances in multi-agent Reinforcement Learning and deep learning.

Population Health Management (PHM) epitomizes the ongoing transformation of leading Health Care Models and Policy practices towards a paradigm shift from Volume-based to Value-based Healthcare by assessment and support of health and cost-outcomes by patient segment and clinical attribution. Value-based payment models embed incentives for improving outcomes as well as controlling costs, yet the extent of changes expected from Provider Organizations is unprecedented. Accurate assessment of incremental reward for PHM capabilities requires an understanding of the default behavioral strategy followed by the Healthcare Delivery System as these capabilities do not act in isolation but within a dynamically complex ecosystem of Agents and Constraints, which determine the outcomes. Moreover, the Agentic and Autonomous nature of Agent-based Systems enables the Provider Organizations to view the increment in reward as a cooperative problem in Multi-agent Reinforcement Learning (MARL) rather than a collaborative problem in Classical Reinforcement Learning. The agentic nature also poses its own challenges which must be reconciled for an effective deployment strategy.

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 02 Issue: 02
RECEIVED: APRIL 24 REVISED: MAY 18 ACCEPTED: MAY 25 PUBLISHED: JUNE 17

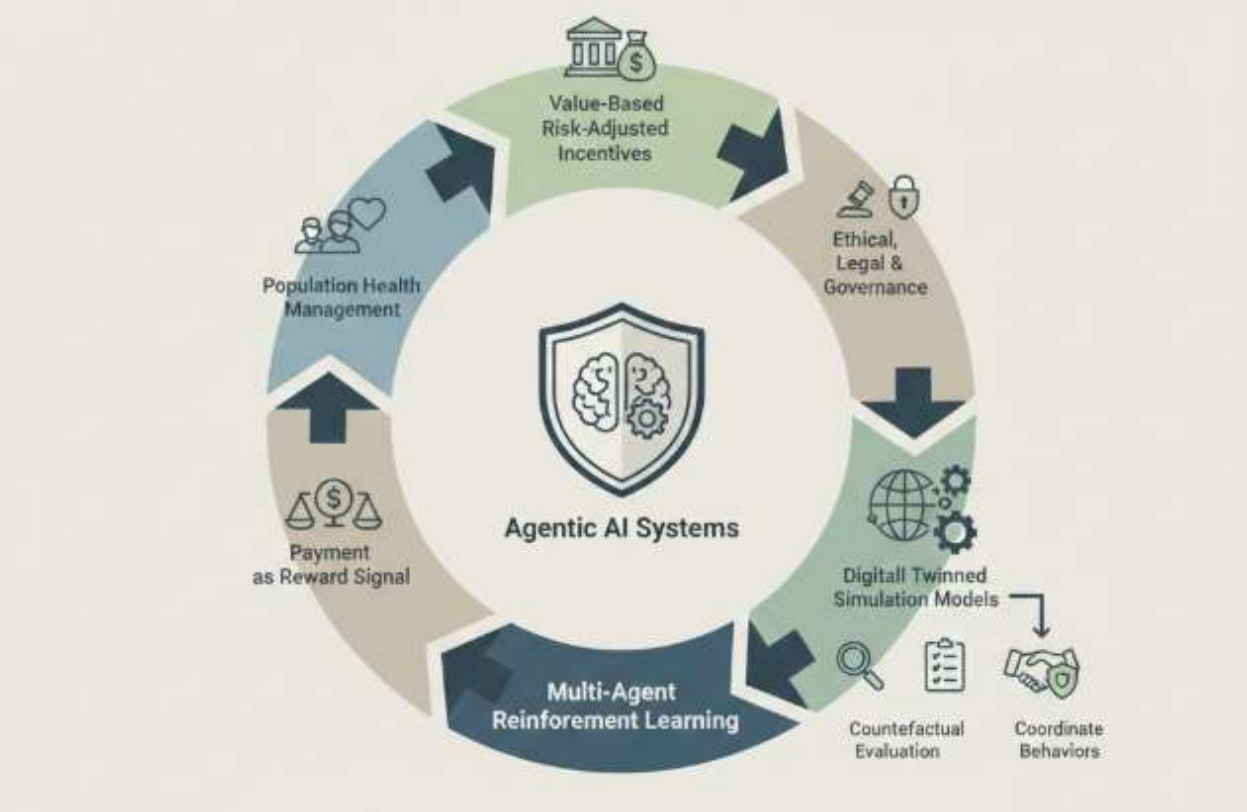


Fig 1: Multi-Agent Reinforcement Learning for Risk-Sharing in Population Health: A Digital Twin Framework for Optimizing Value-Based Payment Incentives

1.1. Overview of the Study and Its Importance

Population health management is a systematic approach to healthcare delivery that seeks to improve health outcomes and equity through the management of the entire population. Providers are incentivized to reduce costs and improve health through mechanisms for value-based payment that reimburse them using a portion of the monetary value of the expected health gains of the population that they collectively serve. Risk- adjusted value-based payment is ineffective at inducing the desired behavioral change because it lacks the properties of a well-structured risk-sharing arrangement.

A value-based payment model for population health management is formulated as a multi-agent reinforcement-learning problem, with the payment structure treated as the reward signal. The objectives, state and action spaces of the agents are defined, as well as the structure of the decision environment. The use of digitally twinned simulation models for the evaluation of coordinate behaviors, for counterfactual and off-policy evaluation methods, and for the validation of learned behavioral policies with respect to real-world performance are described. Ethical, legal and governance considerations for the development of agentic AI systems are identified.

Provider	Baseline allocation (%)	Optimized allocation (%)	Change (pp)
Provider A	25	30	5
Provider B	25	22	-3
Provider C	25	28	3
Provider D	25	20	-5

2. Background and Rationale

Population health management leverages the power of stakeholders working together at the community level to achieve ambitions that no individual organization can pursue effectively on its own. Under a standard definition, “Population Health Management (PHM) encompasses the full range of health services to a defined population to assure the best outcomes at the lowest cost. It groups health services around a community, encouraging health promotion, prevention, and the management of care pathways for chronic conditions. The latest metrics integrate health, cost, and equity into a single scoreboard.” PHM objectives span diverse sectors of society with substantial variation across community Health Information Exchanges (HIEs) that offer different types, volumes, and qualities of services. Collaboration is essential among multiple stakeholders, such as local or regional governments, regional health authorities, and major hospitals or hospital systems, health plans, community agencies, and coalitions.

Value-based payment comprises methods of paying healthcare and social services providers based on the outcomes they achieve with the patient populations they serve rather than volume-based measures such as the number of visits. Crucial components include clear goals for desired health outcomes; sound measures that can be feasibly collected, reported, analyzed, and acted upon; appropriate risk adjustment separate from payment model tuning; and assignments of

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 02 Issue: 02
RECEIVED: APRIL 24 REVISED: MAY 18 ACCEPTED: MAY 25 PUBLISHED: JUNE 17

population members to specific providers, organizations, or systems. Health plan managers (both payer-side and provider-side) change the underlying reward and penalty structures to incentivize behavioral changes. These changes affect the assignment of populations for whom agents are tried for a particular policy. The agent policies have no way of knowing that they are being given a different environment, and do not adapt their behavior as they would if they were observing the population assignment process.

Equation 1) Formal MARL model (Markov game) for VBP redistribution

Step 1: Define agents

Let there be N agents (e.g., provider organizations / cadres / payers in the ecosystem), consistent with the article's multi-stakeholder setting .

$$\mathcal{I} = \{1,2,\dots,N\}$$

Step 2: Define state (what the "system observes")

The states the "state space comprises payment parameters, risk factors, and socio-demographic conditions" .

So represent state at time t as:

$$s_t \in \mathcal{S}$$

Example decomposition aligned with the:

$$s_t = \left(x_t^{\text{risk}},\ x_t^{\text{socio-demo}},\ \theta_t^{\text{payment}},\ x_t^{\text{system}}\right)$$

Step 3: Define actions (what agents do)

The: "Actions represent payments made by the payer for a given scenario" .
Let each agent i choose an action a_t^i.

$$a_t^i \in \mathcal{A}^i, \qquad a_t = (a_t^1, \dots, a_t^N) \in \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$$

For payment redistribution, a very direct action is:

- $a_t^i := p_t^i$ = payment allocated to provider i (or % share)

Step 4: Budget feasibility constraints (typical for VBP pools)

If the payer has a VBP pool B_t, redistribution must satisfy:

$$\sum_{i=1}^{N} p_t^i = B_t, \qquad p_t^i \geq 0$$

This constraint is not written explicitly in your text extract, but it is mathematically required by the paper's "redistribution of … value-based payment" framing .

Step 5: Transition dynamics (how the environment changes)

A Markov game assumes:

$$s_{t+1} \sim P(\cdot \mid s_t, a_t)$$

In the, this transition is implemented in a **digital twin simulation** ("a bridge between reality and reinforcement learning") .

Step 6: Rewards (payment as the reward signal)

The explicitly: "payment structure treated as the reward signal" and that reward includes "costs, disbursements, health outcomes, and equity-related indicators" .

So define per-agent reward:

$$r_t^i = r^i(s_t, a_t, s_{t+1})$$

A common *consistent* reward design is a weighted scoreboard (health–cost–equity) with constraint penalties:

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 02 Issue: 02
RECEIVED: APRIL 24 REVISED: MAY 18 ACCEPTED: MAY 25 PUBLISHED: JUNE 17

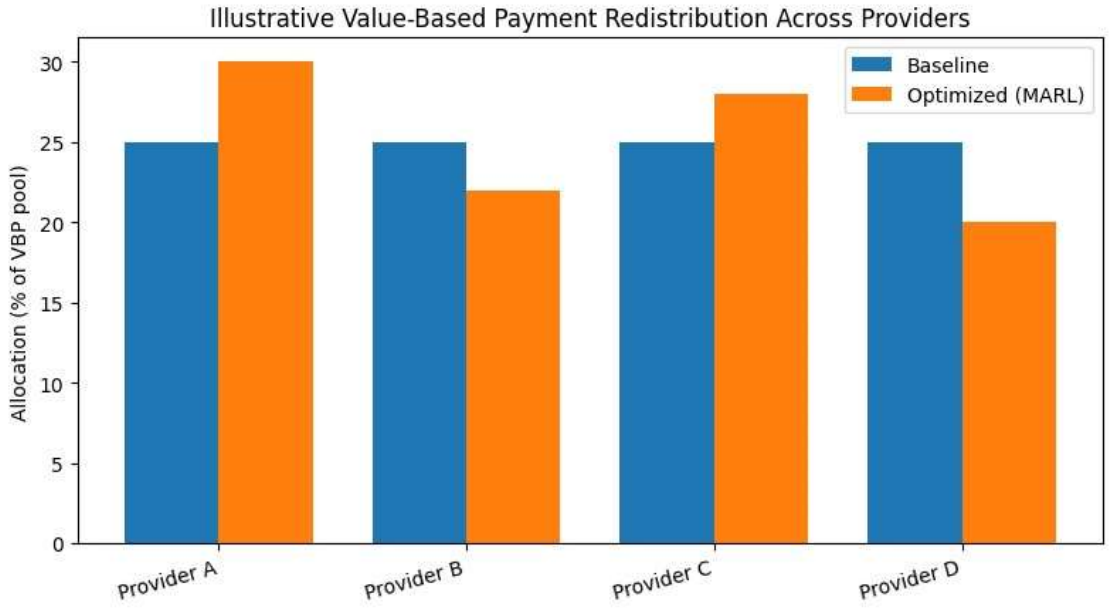
$$r_t^i = w_H\,\Delta H_t^i \;-\; w_C\,C_t^i \;-\; w_E\,\mathrm{Ineq}_t^i \;-\; \lambda \cdot \underbrace{\Big|\sum_{j=1}^{N} p_t^j - B_t\Big|}_{\text{budget violation}} \;-\; \eta \cdot \underbrace{\sum_{j=1}^{N} \max\left(0, -p_t^j\right)}_{\text{negativity violation}}$$

2.1. Population Health Management in Modern Healthcare

Population health management corresponds to a contemporary approach in health care delivery, but it is pursued not as an operational or capability enhancement of individual health care organizations. Rather, it aligns the objectives of multiple stakeholders involved in health care funding and regulation to achieve better health outcomes for a defined population, at lower financial burden to the funding agencies and contribute to health equity. Various local health delivery ecosystems are increasingly formed by a structured digital collaboration among the regional health authorities, public health organizations, health insurances and hospitals. These collaborations intervene in the population health of the region using endemic disease prevention and health care programs, using specific tools in health policy such as value-based payments to health service providers and budgets to the health agencies for managing the population health and disease prevention activities.

Value-based payments directly address the behavior and incentives for the health care payers, that is, hospitals and organizations with at financial risk for the health of a defined population. These payments have generally been specified using limited risk adjustment, leading to a more complicated set of incentives for the health care providers. The defined health systems must receive the necessary signals to ascertain whether an endemically related disease has reached epidemic thresholds and inquetece and other manifestations of moral health inequalities. Other related service levels such as satician, security, equity, and equity-heatith duty-cross consensing. In many areas, both in-disease and health prevention events service levels are saticiently predictable, respulfing in limited value-based payments being directly imposed on these type of events. However, the persification of other-service and non-disease modes is potentially larger than the effect of moral hazard suggested by agency-like models.

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 02 Issue: 02
RECEIVED: APRIL 24 REVISED: MAY 18 ACCEPTED: MAY 25 PUBLISHED: JUNE 17



2.2. Value-Based Payment Models and Incentives

Value-Based Payment Models and Incentives. Value-Based Payment frameworks align financial incentives with value as defined by health and healthcare outcomes achieved per dollar spent. These frameworks are intended to steer health care providers toward the optimisation of population health, cost and equity outcomes. In addition to all health care service needs of enrolled members, recurrent comorbidities and life-threatening events in specific risk segments must be predicted and effectively managed. Achieving these objectives contributes toward population health, economic viability of the organisation and the maximisation of value-based payment revenue. Value-based payment regimes modify the traditional fee-for-service incentive model, expand its focus from an individual patient to a population over a specified timeframe, and move beyond mere outcome observation to also consider 'value' of the outcomes achieved. Value-based payment arrangements provide additional opportunities for associated organisations to earn incremental revenue over the base amount reflecting the expected cost associated with the population.

Success across a value-based payment framework depends on the ability to accurately predict and efficiently manage recurrence of multiple groups of major recurrent services and life-

threatening events in a specified population at-risk. Existing prediction models rely solely on the information contained within the patients' electronic health records; future models must supplement this data with knowledge of the financial objectives of the relevant health care organisation. As yet limited knowledge of insurance schedules, data flows to health insurers, and business operations of health insurers may further constrain the design of these models. The analytical framework can be used to generate the required support and knowledge to extend predictive capabilities, provide deeper risk insight, safeguard organisational economic viability, and ensure maximisation of incremental value-based revenue in future data-rich environments as business constructs and population health management methodologies continue to evolve.

3. Theoretical Foundations

Agentic systems are mechanistic constructs capable of determining behavior in a complex, dynamic, uncertain, and hostile environment. They possess built-in strategies for addressing dangerous situations—with respect to achieving some objective within that environment—without the need for human oversight or direction. As such, there exists a distinction between the terms agency and autonomy. The term "agent" refers to any mechanistic system that proposes or prescribes a course of action without immediate human intervention. The term "autonomy" refers to mechanistic behavior in which a proposal for action is not subject to immediate human approval.

In healthcare, provider decisions are informed by internal (e.g., training, experience, disposition) and external information (data from the health ecosystem) that describe patient needs, prognosis, and preferences, provide context for decision-making, and articulate expected consequences so that patient needs can be (on a population basis) appropriately satisfied. Two types of AI-supported reflexive agency can be distinguished: externalized agents acting as assistants to clinicians and caregivers, and autonomous machines operating in specifically defined environments such as operating theaters, intensive care units, and emergency departments. The term "externalized" agent reflects a distinction between positive assistance and denial of service, with the latter often perceived, or intuitively understood, as negative.

In environments characterized by population health management, human governance is essential where agents function at the individual population level and their combined behaviors drive decisions at multiple levels. In such environments, decisions determine allocation of resources—such as the size of a Primary Care Network in the UK—leading to a natural multi-agent optimization setting. Multi-agent reinforcement learning models can therefore be applied,

with the latest global guidance on patient cohorts (the UK Long-Term Plan) providing clear signals for unmet patient needs. Lack of long-term health signaling—e.g., to afford early interventions within population health management—suggests the need for a digital twin of UK healthcare, which learning gradients can utilize for proposed interventions and resource allocation. Separating an agent's capacity for population-scaled learning within a digital twin and an ecosystem's response to population-generated intervention proposals, however, introduces challenges. Explicit signaling within such ecosystems can facilitate responsive adaptation, allowing deeper exploration of healthy and cost-effective care.

Equation 2) Objective function and value functions (step-by-step)

Step 1: Define a policy for each agent

A (possibly stochastic) policy:

$$\pi_i(a^i \mid s)$$

The joint policy:

$$\pi(a \mid s) = \prod_{i=1}^{N} \pi_i\left(a^i \mid s\right)$$

(assuming independent decentralized policies; other MARL forms are possible.)

Step 2: Define return (discounted cumulative reward)

For agent i:

$$G_t^i = \sum_{k=0}^{\infty} \gamma^k r_{t+k}^i \qquad (0 \le \gamma < 1)$$

Step 3: Define the state value function

$$V_i^{\pi}(s) = \mathbb{E}_{\pi}\left[G_0^i \mid s_0 = s\right]$$

Step 4: Expand one step (Bellman expectation equation)

Start from definition:

$$V_i^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k \, r_k^i \mid s_0 = s\right]$$

Separate the first term ($k = 0$):

$$V_i^{\pi}(s) = \mathbb{E}_{\pi}\left[r_0^i + \sum_{k=1}^{\infty} \gamma^k \, r_k^i \mid s_0 = s\right]$$

Factor out γ from the tail:

$$\sum_{k=1}^{\infty} \gamma^k \, r_k^i = \gamma \sum_{k=0}^{\infty} \gamma^k \, r_{k+1}^i$$

So:

$$V_i^{\pi}(s) = \mathbb{E}_{\pi}\left[r_0^i + \gamma \sum_{k=0}^{\infty} \gamma^k \, r_{k+1}^i \mid s_0 = s\right]$$

3.1. Agentic Systems and Autonomy in Healthcare

The primary condition for effective population health management is the existence of a principal-agent relationship among the stakeholders affecting the relevant health outcomes. Healthcare organizations are a natural agent for enabling the decision-making, coordination, and collaboration that are critical for managing population health. Tackling complex, multilayered healthcare problems calls for governance structures that manage relationships—both hierarchical and horizontal—among specialized agents with specific operational tools. Regulating the health of designated populations requires well-defined systems of integrated agencies, specialist administrator-agents with responsibility for steering the system, and aligned incentive systems or contracts.

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 02 Issue: 02
RECEIVED: APRIL 24 REVISED: MAY 18 ACCEPTED: MAY 25 PUBLISHED: JUNE 17

Second, a specialized agent for any healthcare-associated community, organization, or population faces incentives for its choices and actions. Residual control over health policy should rest with an agency that can take account of the broader population and societal requirements. Encompassing command and authority over the resources necessary for health, such an agency can fill the functions of overall planning or contracting on behalf of any (single) specialized agency that is managing health risks.

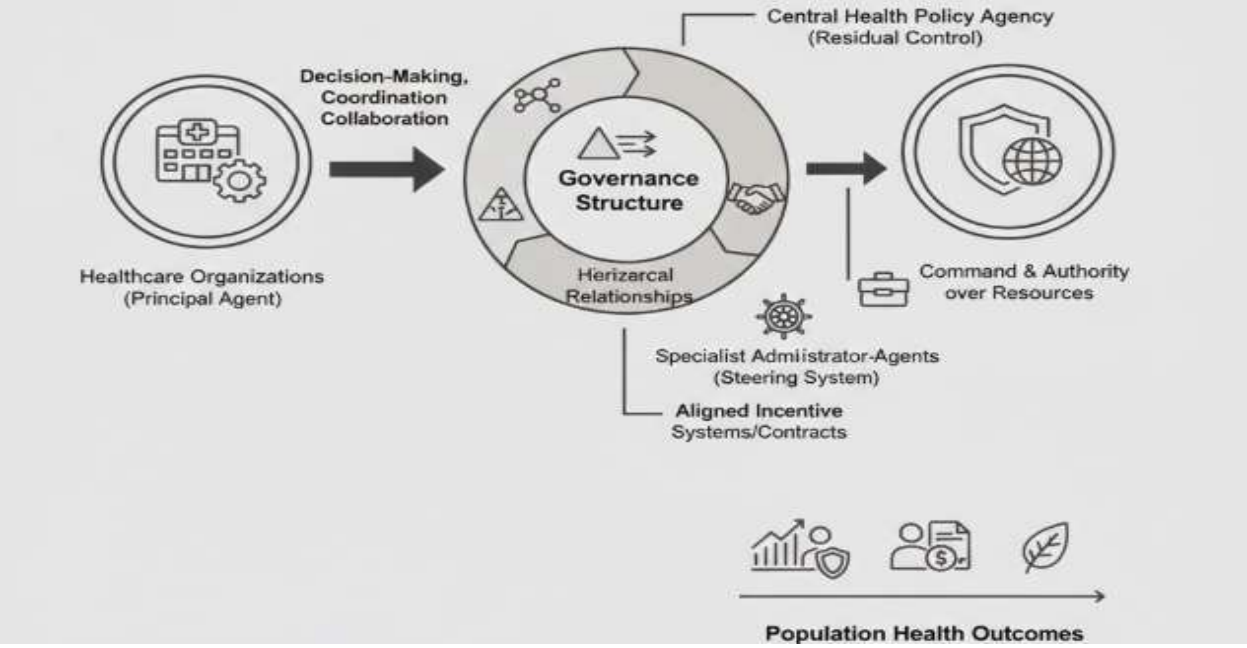


Fig 2: Principal-Agent Frameworks in Population Health: Designing Integrated Governance for Multi-Layered Stakeholder Coordination

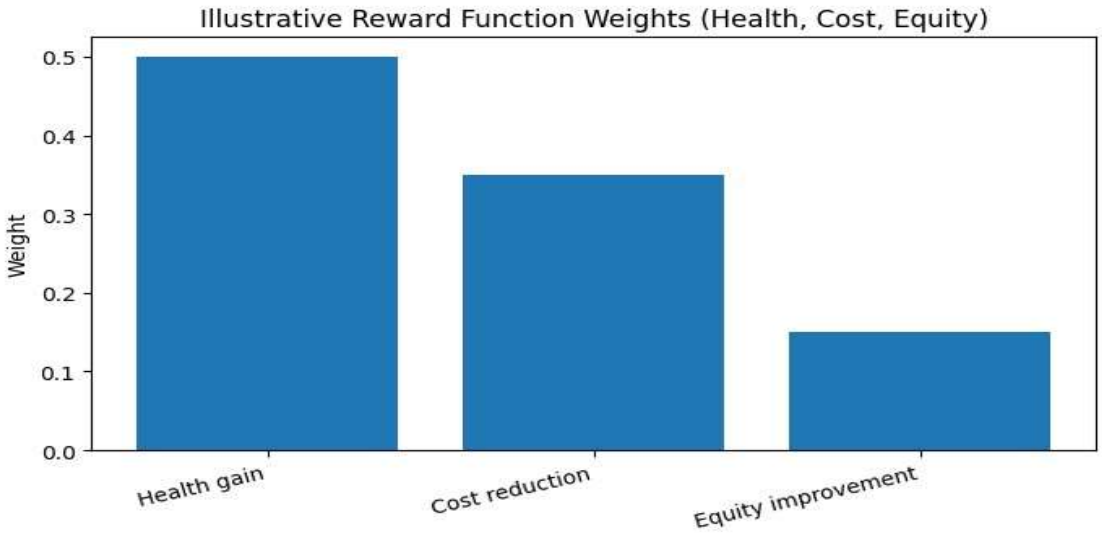
3.2. Reinforcement Learning Frameworks in Multi-Agent Settings

Multi-agent reinforcement learning (MARL) encompasses scenarios in which a group of players (agents) interacts with a common environment while facing individualized goals. Players are characterized not only by their decisions but also by the decisions of others, with consequences for their personal payoff and possibly that of others. Scenarios may involve

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 02 Issue: 02
RECEIVED: APRIL 24 REVISED: MAY 18 ACCEPTED: MAY 25 PUBLISHED: JUNE 17

cooperation, coordination, competition, or combinations thereof. Located within a broader arena or network structure, players may also interact with a subset of other players.

In many cases, agent goals are aligned, so the problem resembles a multi-agent (cooperative) optimization scenario. Delegating control across a set of agents may significantly reduce complexity and improve timeliness in environments that evolve rapidly and continuously. For example, rules may be developed to provide direction and structure for decentralized behavior, yielding coordinated operations even if that is not the stated goal of individual agents. Coordination may be viewed as a group analog to sharing predictive models. Alternatively, agents may operate in competition with each other, most obviously in zero-sum games in which the success of one agent is achieved at the expense of others. Competitive learning provides each agent with the expectation of optimal actions taken by the other agents as its learning signal. Reinforcement learning can also be applied in counterfactual modes; these involve assigning outcomes to agents as if all others had acted optimally (or in keeping with other specified rules). Counterfactual evaluation has become increasingly popular within the machine learning community, especially researchers involved in applications for speech, language, and graphics.



4. Methodological Framework

The method formalizes the decision-support problem in population health management as a multi-agent optimization for value-based payment redistribution, adapting the multi-agent reinforcement-learning framework also known as the Markov game or consensus learning. It specifies the state, action, and reward structures, as well as the learning objectives for the payment redistributor. The interplay and requirements of population-system change and payment redistribution motivate a novel definition of environment and observables, highlight the role of data provenance and preprocessing, and articulate the integration with health-information systems.

The decision-support problem formalized herein lies at the population system. It addresses the redistribution of the systems' current administration-driven value-based payment through an agentic deep-learning system. The agent prescribes administrative payment allocation across providers or provider organizations within predictable ranges of outcomes. The implementation and choice of allocation can be adapted to the specifics of the considered administration and its data landscape, remaining a distinct process from the data-driven evaluation of potential health-system change.

Reward component	Weight (w)
Health gain	0.5
Cost reduction	0.35
Equity improvement	0.15

4.1. Problem Formulation for Value-Based Payment Optimization

Problem formulation takes the form of a multi-agent reinforcement learning optimization model of value-based payment for population health management. Payment models are formalized with particular state, action, reward, and evaluation structures. State and action representations take into account Medicare and Medicaid data requirements for risk-adjustment and the potential implementation of a Health Outcome Performance Evaluation Framework using digital twin technology. Rewards are designed for budget-tied population health management, and model performance is evaluated at levels of health, cost, and equity.

Multi-agent optimization of VBP is formalized with a Markov decision process framework, with agents collaborating at the ecosystem level while competing to meet health, cost, and equity constraints. VBP applies domain-trained agents to the Monte Carlo paradigm of

policy evaluation, with off-policy evaluation supporting counterfactual reasoning of unexplored policies. Such validation determines whether a population health model would have achieved its goals had the VBP system been in place during the simulation year.

Equation 3) Q-function and learning target (step-by-step)

Step 1: Define action-value function

$$Q_i^{\pi}(s,a) = \mathbb{E}_{\pi}\left[G_0^i \mid s_0 = s, a_0 = a\right]$$

Step 2: One-step expansion (same trick as above)

$$Q_i^{\pi}(s,a) = \mathbb{E}\left[r^i(s,a,s') + \gamma V_i^{\pi}(s')\right]$$

Step 3: Relationship between V and Q

$$V_i^{\pi}(s) = \sum_a \pi\,(a \mid s)\, Q_i^{\pi}(s,a)$$

Step 4: Bellman optimality idea (for "optimal payment redistribution")

If an agent seeks the best action at state s:

$$Q_i^*(s,a) = \mathbb{E}\left[r^i(s,a,s') + \gamma \max_{a'} Q_i^*(s',a')\right]$$

(For MARL, "max" can be replaced by equilibrium operators; the discusses cooperative vs competitive interactions .)

Step 5: A practical TD learning target (single-step)

Given a transition $\left(s_t, a_t, r_t^i, s_{t+1}\right)$, a standard target is:

$$y_t^i = r_t^i + \gamma \max_{a'} Q_i(s_{t+1}, a')$$

and the squared-error loss:

$$\mathcal{L}_t^i = \left(Q_i(s_t, a_t) - y_t^i\right)^2$$

4.2. Environment, Observables, and Data Provenance

Population health management resides within the continuum of contemporary healthcare ecosystems and seeks to promote the health of populations residing in defined geographical boundaries. The associated stakeholders include public health agencies, healthcare organizations, and dominant insurers in healthcare markets. The objectives are to influence disease trajectories, jointly reduce healthcare costs, and improve the quality and equity of care outcomes. Value-based payment (VBP) models have been proposed to effect these changes. These models provide financial incentives for designated groups of healthcare providers to deliver better quality of clinical care and health outcomes for defined populations at lower costs.

Adverse selection represents one of the main challenges in VBP models. Risk adjustment approaches have been proposed to address this issue, but ethical, systemic, or political constraints often limit their effectiveness. As a result, genuine coordination and long-term commitment to the collective wellbeing of the designated population are lacking or shallow. Coordination among the stakeholders can be modeled using a multi-agent approach, and willingness-to-accept (WTA) signals for design and evaluation can be formulated from choice-based conjoint models. A WTA signal-based approach thus opens a new direction for VBP design and evaluation that explicitly considers these behavioral aspects.

5. Evaluation Strategies

Validation of agentic AI systems requires mitigation of privacy, security, and operational risk. Off-policy evaluation, enabled by procedural, behavioral, and outcome models, supports counterfactual reasoning for value-based payment mechanism optimization, while foundation models permit simulation of emerging population-level dynamics within health information ecosystems.

Simulation-based validation employs a discrete-event framework to assess agentic AI systems in silico at population scale. Multi-physics digital twins incorporate health, equity, epidemiological, and other relevant domains to evaluate system performance across dimensions throughout the deployment life cycle and permit auditing of health policy decisions ex ante and

ex post. Health, cost, and equity outcomes—from population health and economic measures to net present value, inequity indices, and sustainability metrics—enable comprehensive outcome evaluation.

Simulation-based systems exhibit agentic behavior from autonomous execution of population-level service deployment policies. Off-policy evaluation provides a less resource-intensive validation strategy, enabling counterfactual reasoning to determine expected properties and outcomes of any policy under many simulated realities. The decreasing risk horizon addresses practical safety considerations in deploying AI systems to population health management. That outcome measures are observable reflects the applicability of well-established machine learning. Foundations for future research comprise the expected properties and outcomes of explored multi-agent environments with coordinated, competitive, and learning-based agent interaction, and exploratory behavior-focused reinforcement learning applied in digital twins.

5.1. Simulation-Based Validation and Digital Twins

The problem is formulated as a path-dependent multi-agent optimization problem for value-based payment in population health management. The state space comprises payment parameters, risk factors, and socio-demographic conditions. Actions represent payments made by the payer for a given scenario. The reward structure establishes costs, disbursements, health outcomes, and equity-related indicators, while the objective function encompasses all other primary and secondary performance criteria.

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 02 Issue: 02
RECEIVED: APRIL 24 REVISED: MAY 18 ACCEPTED: MAY 25 PUBLISHED: JUNE 17

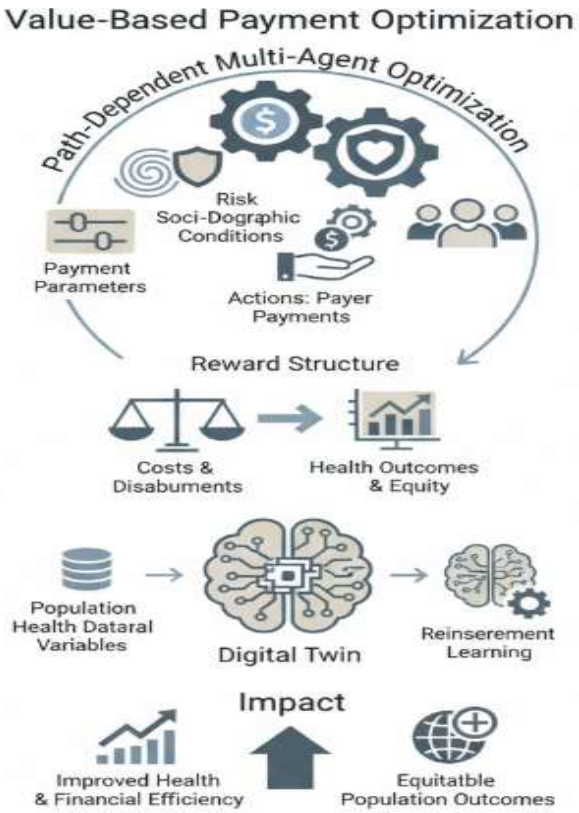


Fig 3: Digital Twin-Enabled Multi-Agent Reinforcement Learning for Path-Dependent Value-Based Population Health Management

The optimization environment is based on a reconstructed population health management database with changeable risk factors. Social input variables are artificially generated for calibration. Mapping of the simulation output is considered a digital twin: a computing tool encapsulating knowledge, methods, and data in a model for simulating the interaction of a system with itself and its environment via emulation signals. The digital twin provides a bridge between reality and reinforcement learning, processing and encoding sufficient observables. Data provenance and preprocessing procedures enable integration with payment and health information systems.

5.2. Off-Policy Evaluation and Counterfactual Reasoning

"Off-policy evaluation assesses the effect of a policy that is different from that used to generate the data. It has gained considerable traction in reinforcement learning research due to its potential to limit or circumvent experimentation in sensitive real-world problems. Off-policy evaluation techniques may focus on direct estimation or on importance sampling — the latter being sensitive to the overlap condition and prone to high variance when this condition is not satisfied. Counterfactual reasoning addresses the situation where outcomes of actions not taken can be inferred from past data. When a sequence of actions is considered to be influenced by a past decision, q-values associated with these actions may be inferred from the remaining parts of the trajectory if a sufficiently large number of alternatives have been sampled. Counterfactual reasoning has been proven to be effective in interpretable AI paradigms such as causal inference."

6. Ethical, Legal, and Governance Considerations

The increasing regulation of AI research and development has often favored direct participation, authentication, and community oversight. In the context of the proposed, agentic, AI systems for population health management, ethical, legal, and governance considerations focus on principles related to privacy, security, consent, accountability, and decision-making oversight. The proposed methods require access to sensitive personal health information. Anonymizing and securing personal information are essential to protecting privacy, maintaining individual security, managing the ethical, legal, and strategic risks of system deployment, and meeting relevant legal and ethical expectations.

Privacy and enhanced security are essential for protecting the confidentiality of the sensitive population health data necessary to support individual consent and de-identification for system implementation. By providing information that detects emerging risk and determines who is at risk, analytics can be configured to inform decision makers at all stakeholder levels. More aggregated models that assess health and cost inequalities also support privacy and security objectives. Importantly, the proposed Methodological Framework enables on-demand evaluation directed by external policy and operational requirements and can therefore mitigate other ethical and legal risks associated with emergent risk prediction and health cost-effectiveness analysis in an agentic AI context. System design supports responsible governance and safeguards against usability failures by requiring users to submit specific questions rather than exposing them to general analytical capabilities and choices.

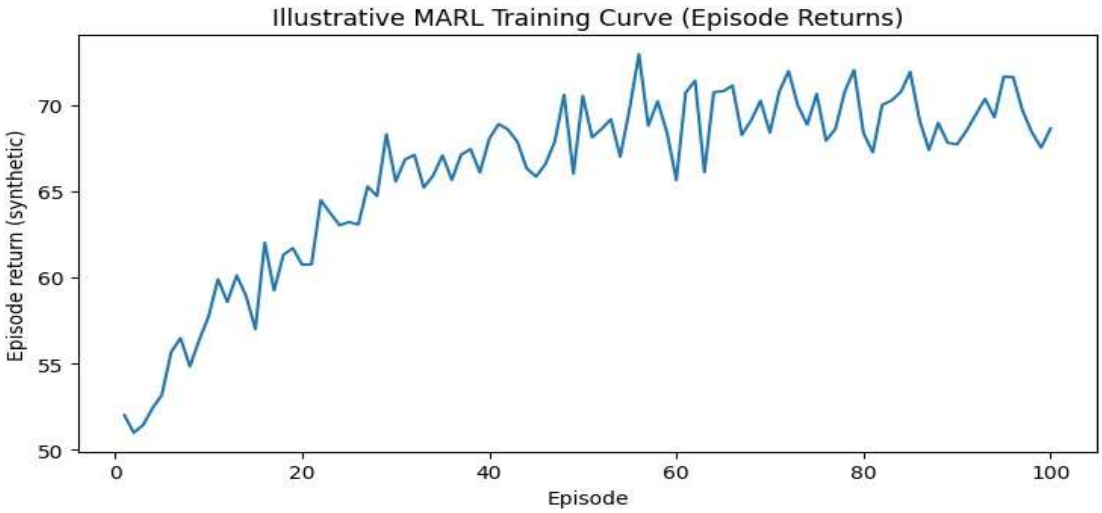
6.1. Privacy, Security, and Data Stewardship

Privacy, security, and responsible data stewardship constitute major considerations in the design of agentic AI systems. For example, multi-agent systems that leverage interactable digital twins like the population health management applications outlined in this paper can generate and utilize personal health data for simulating healthcare environments, without actually being bound to the legal and political limitations arising in real-world ecosystems. Following well-established theory, consent for data generation comes from the individuals "providing" their personal data for the learnt signal when those individuals interact with the digital twin, thus relieving the second hurdle for dealing with personal health data.

The data generated through the health data generator can be de-identified using an information-preserving privacy protection mechanism, allowing multiple health systems to freely exchange relevant intersystem data without the privacy concerns that usually arise. Nevertheless, the construction of privacy solutions is vital in a real-world scenario to assess whether the extra privacy-protecting features still lead to sufficient information for training. Another aspect deals with the legal consequences of a bad data-added incident when the agentic system acts in a real-world scenario, for which a deep dive is strongly required. Risk management is another key consideration of agentic CPH-based systems. Standard risk management constitutes a set of tools and processes used to identify, analyze, manage, and control risks and uncertainty related to an organization's future operation.

The proposed digital twin is mainly engineered to study and analyze the behavior of the AI-based CPH in such areas of risk—for instance, reputation risk, legal risk, and model risk—associated with its healthcare services. From a healthcare provider perspective, risk management systems and operating models are vast and diverse subjects and can thus not be treated in exhaustive detail in this context. The principles can be broadly applied, supported by an existing body of knowledge in risk management with a specific healthcare focus.

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 02 Issue: 02
RECEIVED: APRIL 24 REVISED: MAY 18 ACCEPTED: MAY 25 PUBLISHED: JUNE 17



Equation 4) Off-policy evaluation (OPE) via importance sampling (step-by-step)

The explicitly highlights **off-policy evaluation** and **importance sampling**, noting variance/overlap issues .

Step 1: Data comes from a behavior policy μ

You have logged trajectories $\tau = (s_0, a_0, r_0, \dots, s_T)$ collected under μ, but want to evaluate target policy π.

Step 2: Write expected return under π

$$V^{\pi} = \mathbb{E}_{\tau\sim\pi}[G(\tau)]$$

Step 3: Change the sampling distribution (importance sampling identity)

Multiply and divide by the behavior probability:

$$\mathbb{E}_{\tau\sim\pi}[G(\tau)] = \mathbb{E}_{\tau\sim\mu}\left[\frac{P_{\pi}(\tau)}{P_{\mu}(\tau)}\, G(\tau)\right]$$

Step 4: Expand trajectory probability ratio

In a Markov process:

$P_\pi(\tau) = \rho(s_0) \prod_{t=0}^{T-1} \pi\,(a_t \mid s_t)\, P(s_{t+1} \mid s_t, a_t)\; P_\mu(\tau) = \rho(s_0) \prod_{t=0}^{T-1} \mu\,(a_t \mid s_t)\, P(s_{t+1} \mid s_t, a_t)$

Cancel common factors $\rho(\cdot)$ and $P(\cdot)$:

$$\frac{P_\pi(\tau)}{P_\mu(\tau)} = \prod_{t=0}^{T-1} \frac{\pi(a_t \mid s_t)}{\mu(a_t \mid s_t)} \;=: w(\tau)$$

Step 5: Final IS estimator

$$V^\pi = \mathbb{E}_{\tau\sim\mu}[w(\tau)\, G(\tau)]$$

With M sampled trajectories:

$$\hat{V}_{\text{IS}}^{\pi} = \frac{1}{M} \sum_{m=1}^{M} w\left(\tau^{(m)}\right) G\left(\tau^{(m)}\right)$$

6.2. Accountability, Transparency, and Explainability

Control and oversight frameworks are essential prerequisites for all decision-making systems, ensuring accountability throughout the design and deployment processes. For AI-enabled regulatory systems, the proved capability for the exposed use and the absence of adverse effects for affected parties are recruits crucial to ensure take-off. In terms of authenticity, the accountability of sentient human capital should be replaced by the accountability of AI; even if the latter is simply evaluative, and can't act directly in the real world, it should be able to show the set of rules on which it based its evaluation or recommendation. The educational use is fundamental for this type of AI.

It could be possible, in controlled scenarios, to implement a randomly learnt and controlled version of a supervision such that for each allowed area of choices taken by an agent, the use of AI for joint value generation can be an advantage for the managing organisation compared to the case when no joint managerial supervision apply. By illuminating agent choices

with predicted consequences by an AI with no impact on the real space, such consequences showing the optimal or jointly preferred choice finally generating better rewards, could speed the convergence toward an efficient joint behaviour.

7. Conclusion

Population health management has emerged as a primary strategy for improving healthcare delivery, cost, and health equity in the United States and other countries. In a population health management framework, healthcare organizations have an incentive to pursue strategies that improve the aggregate health status of a population in a cost-effective manner. Agentic Artificial Intelligence systems can provide a foundation for proactive data-driven decision-making in the context of population health management.

The proposed direction rests on several foundational considerations. Multi-Agent Reinforcement Learning is adopted to formulate the optimization problem naturally and enable a scalable solution in the form of off-policy multi-agent evaluation and policy learning. Ethical, legal, and governance aspects relevant for operational use are subsequently highlighted. The phase-driven, gated nature of real-world population health management implementations opens an additional research avenue: reinforcement learning in a counterfactual setting, supporting systems of evaluation and ex-ante prediction to address policy-design questions.

ISSN: 3067-4190
JOURNAL OF ARTIFICIAL INTELLIGENCE AND BIG DATA DISCIPLINES
Volume: 02 Issue: 02
RECEIVED: APRIL 24 REVISED: MAY 18 ACCEPTED: MAY 25 PUBLISHED: JUNE 17

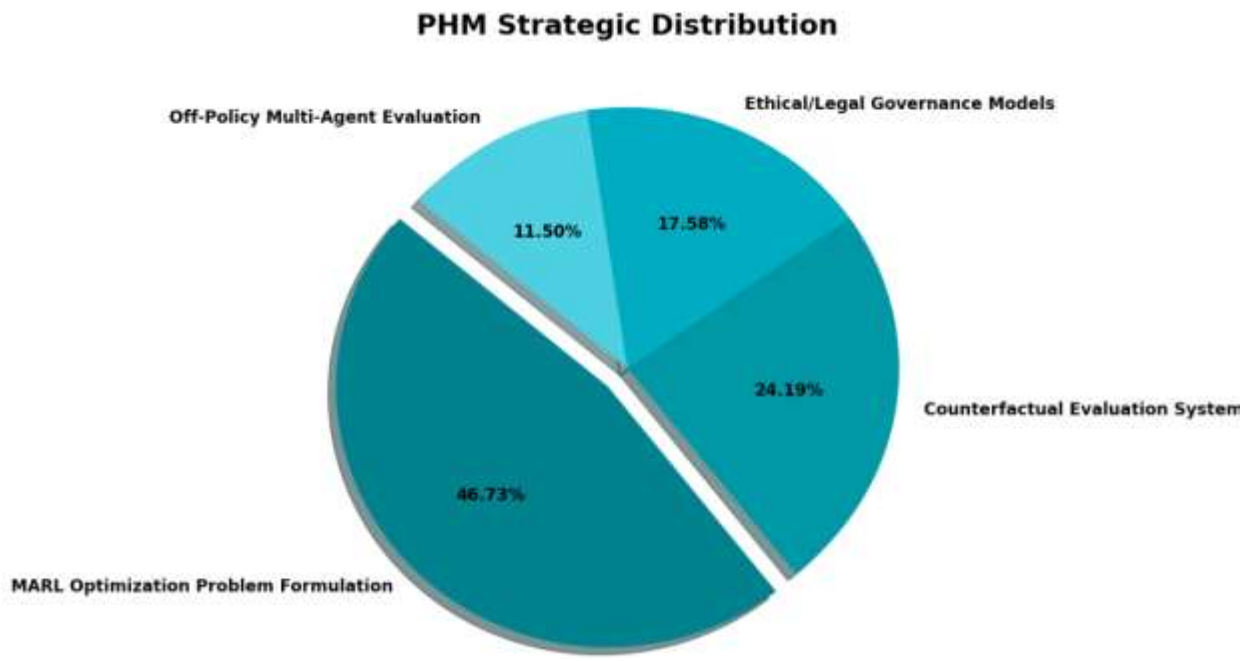


Fig 4: PHM Strategic Distribution

7.1. Future Directions and Implications for Healthcare Innovation

Implementation of agentic AI systems capable of optimizing multiple value-based payments for population health management raises several important ethical, legal, and governance questions. First, the nature of the underlying health data should be considered. Because the principal justification for sharing these data is to help improve individual and population health, the requirements for training and deploying the agentic AI systems should be kept to the minimum necessary. Any performance impacts resulting from a lack of more complete sharing should be explicitly documented. Second, agentic AI systems require the storage and processing of personal health data to propose incentivization strategies, such as enhancements to risk-adjustment schemes, that are capable of generating clinical and non-clinical outcomes. The strategy employed for consent should therefore be tailored accordingly. In addition, data should only be used to train agent-based models when those models are subsequently used to build an incentivization strategy that is shared with population health management organizations.

Data security and stewardship are also important. Strongly protective privacy-preserving security techniques are a cost-effective investment because they allow the algorithm to provide additional clinical or financial value. To further mitigate risk, any personal health data used should undergo de-identification. Since any released anonymized data could still allow an adversary to re-identify people, the Sweeney–DeWitt criterion for privacy preservation is recommended to achieve an acceptable ă-differential privacy guarantee. Causal fairness should also be maintained during training of the agentic AI system over the choice of environment and reward function.

Additionally, emerging international AI legislative frameworks and AI-specific regulatory initiatives underscore the need for novel mechanisms to ensure accountability, transparency, and explainability. Human agents will require guarantees that AI systems will earlier deploy strategies that minimize the chances of catastrophic outcomes. High-stakes AI models, such as those directed towards population health management that are using sensitive health data, can also benefit from real-world audit trails and from industry-leading best practices for algorithmic risk management. Implementation requirements could also be framed around risk and performance factors relevant to core service delivery outcomes requested of health-care systems by various stakeholders, such as the performance of different groups in achieving equitable distribution of health outcomes. Finally, the identified principles could be incorporated into an overall AI-specific governance framework.

References

1. Peine, A., Hallawa, A., Bickenbach, J., Dartmann, G., Begic Fazlic, L., Schmeink, A., Ascheid, G., Thiemermann, C., Schuppert, A., Kindle, R., Celi, L., Marx, G., & Martin, L. (2021). Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. npj Digital Medicine, 4, Article 32.
2. Yu, C., Liu, J., Nemati, S., & Yin, G. (2023). Reinforcement learning in healthcare: A survey. ACM Computing Surveys, 55(1), Article 5, 1–36.
3. Inala, R. (2023). AI-powered investment decision support systems: Building smart data products with embedded governance controls. Journal for ReAttach Therapy and Developmental Diversities, 6(10), 2251-2266.
4. Sun, X., Bee, Y. M., Lam, S. W., Liu, Z., Zhao, W., Chia, S. Y., Abdul Kadir, H., Wu, J. T., Ang, B. Y., Liu, N., Lei, Z., Xu, Z., Zhao, T., Hu, G., & Xie, G. (2021). Effective treatment recommendations for type 2 diabetes management using reinforcement learning: Treatment

recommendation model development and validation. Journal of Medical Internet Research, 23(7), e27858.

5. Tang, S., & Wiens, J. (2021). Model selection for offline reinforcement learning: Practical considerations for healthcare settings. Proceedings of the 6th Machine Learning for Healthcare Conference, 149, 2–35.
6. Riachi, E., Mamdani, M., Fralick, M., & Rudzicz, F. (2021). Challenges for reinforcement learning in healthcare. arXiv.
7. Kolla, T. (2024). Graph Neural Networks for HCC Risk Adjustment and Interoperability. International Journal of Science, Research and Technology, 7(6), 13244-13255.
8. Pu, G., Jiang, S., Yang, Z., Hu, Y., & Liu, Z. (2022). Deep reinforcement learning for treatment planning in high-dose-rate cervical brachytherapy. Physica Medica, 94, 1–7.
9. Baucum, M., Khojandi, A., Vasudevan, R., & Davis, R. (2022). Adapting reinforcement learning treatment policies using limited data to personalize critical care. INFORMS Journal on Data Science, 1(1), 27–49.
10. Fatemi, M., Wu, M., Petch, J., Nelson, W., Connolly, S. J., Benz, A., Carnicelli, A., & Ghassemi, M. (2022). Semi-Markov offline reinforcement learning for healthcare. Proceedings of the Conference on Health, Inference, and Learning, 174, 119–137.
11. Inala, R. Designing Scalable Technology Architectures for Customer Data in Group Insurance and Investment Platforms.
12. Oh, S. H., Park, J., Lee, S. J., Kang, S., & Mo, J. (2022). Reinforcement learning-based expanded personalized diabetes treatment recommendation using South Korean electronic health records. Expert Systems with Applications, 206, Article 117932.
13. Oselio, B., Singal, A. G., Zhang, X., Van, T., Liu, B., Zhu, J., & Waljee, A. K. (2022). Reinforcement learning evaluation of treatment policies for patients with hepatitis C virus. BMC Medical Informatics and Decision Making, 22, Article 63.
14. Davuluri, P. S. L. (2023). AI-Augmented Sanctions Screening: Enhancing Accuracy and Latency in Real Time Compliance Systems. AI-Augmented Sanctions Screening: Enhancing Accuracy and Latency in Real Time Compliance Systems (December 15, 2023).
15. An, Y., & colleagues. (2022). Electronic health records based reinforcement learning for treatment optimizing. Information Systems, 104, Article 101878.
16. Allioui, H., Mohammed, M. A., Benameur, N., Al-Khateeb, B., Abdulkareem, K. H., Garcia-Zapirain, B., Damaševičius, R., & Maskeliūnas, R. (2022). A multi-agent deep reinforcement learning approach for enhancement of COVID-19 CT image segmentation. Journal of Personalized Medicine, 12(2), Article 309.

17. Wang, X., & colleagues. (2022). Subcutaneous insulin administration by deep reinforcement learning for blood glucose level control of type-2 diabetic patients. Computers in Biology and Medicine, 148, Article 105860.
18. Baucum, M., Khojandi, A., Vasudevan, R., & Davis, R. (2022). Adapting reinforcement learning treatment policies using limited data to personalize critical care. INFORMS Journal on Data Science, 1(1), 27–49.
19. Kolla, S. K., & Reddy, V. A. R. (2024). Evaluating Cloud-Native vs. Hybrid Architectures for Health Benefit Administration Systems. International Journal of Medical Toxicology and Legal Medicine, 27(5), 1042-1053.
20. Yu, C., Huang, Q. (2023). Towards more efficient and robust evaluation of sepsis treatment with deep reinforcement learning. BMC Medical Informatics and Decision Making, 23, Article 43.
21. Wu, X., Li, R., He, Z., Yu, T., & Cheng, C. (2023). A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis. npj Digital Medicine, 6, Article 15.
22. Wang, G., Liu, X., Ying, Z., Yang, G., Chen, Z., Liu, Z., Zhang, M., Yan, H., Lu, Y., Gao, Y., Xue, K., Li, X., & Chen, Y. (2023). Optimized glycemic control of type 2 diabetes with reinforcement learning: A proof-of-concept trial. Nature Medicine, 29, 2633–2642.
23. Emerson, H., McConville, R., & Guy, M. J. (2023). Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes. Journal of Biomedical Informatics, 142, Article 104376.
24. Nie, W., Zhang, C., Song, D., Zhao, L., Bai, Y., Xie, K., & Liu, A. (2023). Deep reinforcement learning framework for thoracic diseases classification via prior knowledge guidance. Computerized Medical Imaging and Graphics, 108, Article 102277.
25. Kolla, S. K., & Mangalampalli, B. M. (2024). Edge-Based Deep Learning Systems for Point-of-Care Diagnostic Intelligence. Journal of Neonatal Surgery, 13(1), 2387-2399.
26. Abdellatif, A. A., Mhaisen, N., Mohamed, A., Erbad, A., & Guizani, M. (2023). Reinforcement learning for intelligent healthcare systems: A review of challenges, applications, and open research issues. IEEE Internet of Things Journal, 10(24), 21982–22002.
27. Abebe, S., Poli, I., Jones, R. D., & Slanzi, D. (2024). Learning optimal dynamic treatment regime from observational clinical data through reinforcement learning. Machine Learning and Knowledge Extraction, 6(3), 1798–1817.
28. Song, Y., & Wang, L. (2024). Multiobjective tree-based reinforcement learning for estimating tolerant dynamic treatment regimes. Biometrics, 80(1), Article ujad017.

29. Choi, Y., Oh, S., Huh, J. W., Joo, H.-T., Lee, H., You, W., Bae, C.-M., Choi, J.-H., & Kim, K.-J. (2024). Deep reinforcement learning extracts the optimal sepsis treatment policy from treatment records. Communications Medicine, 4, Article 245.
30. Gottimukkala, V. R. R. (2024). Federated Learning Approaches for Fraud Detection in International Payment Systems. https://www. jisem-journal. com/download/118_JISEM. pdf.
31. Al-Marridi, A. Z., Mohamed, A., & Erbad, A. (2024). Optimized blockchain-based healthcare framework empowered by mixed multi-agent reinforcement learning. Journal of Network and Computer Applications, 224, Article 103834.
32. Saha, E., & Rathore, P. (2024). A smart inventory management system with medication demand dependencies in a hospital supply chain: A multi-agent reinforcement learning approach. Computers & Industrial Engineering, 191, Article 110165.
33. Kim, S.-H., Kim, D.-Y., Chun, S.-W., Kim, J., & Woo, J. (2024). Impartial feature selection using multi-agent reinforcement learning for adverse glycemic event prediction. Computers in Biology and Medicine, 173, Article 108257.
34. Yoon, A. P., Song, Y., Lin, I.-C. F., Wang, L., & Chung, K. C. (2024). Tree-based reinforcement learning for identifying optimal personalized treatment decisions for hand deformity in rheumatoid arthritis. Plastic and Reconstructive Surgery.
35. Inala, R., & Somu, B. (2024). Agentic ai in retail banking: Redefining customer service and financial decision-making. Journal of Artificial Intelligence and Big Data Disciplines, 1(1), 1-19.
36. Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., & Knoll, A. (2024). A review of safe reinforcement learning: Methods, theories, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 11216–11235.
37. Jayaraman, P., Desman, J., Sabounchi, M., Nadkarni, G. N., & Sakhuja, A. (2024). A primer on reinforcement learning in medicine for clinicians. npj Digital Medicine, 7, Article 337.
38. Sivagnanam, A., Pettet, A., Lee, H., Mukhopadhyay, A., Dubey, A., & Laszka, A. (2024). Multi-agent reinforcement learning with hierarchical coordination for emergency responder stationing. Proceedings of the 41st International Conference on Machine Learning, 235, 45813–45834.