International Journal of Applied Sciences, Research, and Innovation

OPEN ACCESS

FULL PAPER | IJASRI | OPEN

Partially Observable Markov Decision Processes for Dynamic Chemo-Immunotherapy Scheduling Under Incomplete Tumor Observability

Kwabena O. Boateng[1,*], **Samuel A. Mensah**[2], **Ebenezer K. Appiah**[3]

[1]Department of Computer Science and Engineering, University of Mines and Technology, Tarkwa, Ghana
[2]Department of Electrical and Electronic Engineering, University of Mines and Technology, Tarkwa, Ghana
[3]Department of Computer Science and Informatics, University of Energy and Natural Resources, Sunyani, Ghana

[*]Corresponding Author

Abstract

Optimizing the temporal sequencing and dosage of combination chemotherapy and immune checkpoint blockade presents a formidable clinical challenge due to the hidden, non-linear evolutionary dynamics of malignant cell populations and host immune effectors. Standard clinical protocols rely almost universally on empirical maximum tolerated dose (MTD) fractionation or rigid metronomic cycles that overlook individual phenotypic heterogeneity and fail to adapt as tumor clones develop treatment-induced resistance. Moreover, clinicians rarely observe true underlying tumor burdens or microenvironmental immune activation states directly; clinical decisions are guided instead by infrequent, noise-corrupted radiological scans and delayed circulating blood biomarkers. In this paper, we formulate dynamic chemo-immunotherapy scheduling as a Partially Observable Markov Decision Process (POMDP) under continuous-state, discrete-action, and asynchronous partial observation constraints. The latent physiological state captures active tumor volumes, cytotoxic T lymphocyte densities, regulatory T cell suppression levels, and cumulative drug-induced organ-at-risk (OAR) toxicities. We develop a Point-Based Value Iteration (PBVI) algorithm equipped with belief-simplex sampling and particle-filtered Bayesian state estimation to compute adaptive, patient-specific dosing policies. The multi-objective reward structure penalizes malignant clone progression while enforcing rigorous probabilistic safety constraints against life-threatening neutropenia and cardiotoxicity. Evaluated across an in silico cohort of 1,000 heterogeneous virtual cancer patients, our POMDP policy achieves a 28.4% increase in two-year progression-free survival relative to fixed MTD schedules while reducing cumulative chemotherapeutic toxicity by 39.1%. The computed policies discover counterintuitive, highly effective therapeutic paradigms, such as intermittent immune-priming micro-dosing and planned treatment holidays that prevent chemo-refractory clonal selection. These findings establish partially observable decision-theoretic frameworks as a powerful paradigm for closed-loop, personalized therapeutic optimization in precision oncology.

1. Introduction

Modern clinical management of advanced solid malignancies increasingly relies on multimodal therapeutic regimens combining cytotoxic chemotherapy with immune checkpoint inhibitors (ICIs) [17, 35]. Cytotoxic chemotherapeutic agents (e.g., platinum compounds, taxanes, anthracyclines) induce rapid tumor debulking and promote immunogenic cell death, releasing tumor-associated antigens that prime host effector T lymphocytes [4, 31]. Concurrently, immune checkpoint antibodies (e.g., anti-PD-1, anti-CTLA-4) block inhibitory receptors on T cells, relieving immunosuppression and reinvigorating antitumor immune surveillance [13, 10]. Despite compelling biological synergy, clinical outcomes remain highly variable across patient populations. A substantial proportion of patients experience either rapid disease progression driven by the emergence of drug-resistant clones or severe immune-related adverse events (irAEs) and hematological toxicities that force premature cessation of therapy [24, 1, 29].

A primary driver of sub-optimal therapeutic efficacy lies in the rigid nature of conventional treatment scheduling. In standard clinical practice, chemotherapy and immunotherapy are administered according to static, population-averaged protocols—typically fixed doses given every two to three weeks until disease progression or unacceptable toxicity [21, 3]. This maximum tolerated dose paradigm was historically developed for cytotoxic monotherapies under the assumption that maximum cell kill correlates monotonically with patient survival [33, 2]. However, this premise fails in the context of immuno-oncology, where excessive chemotherapy depletes proliferating effector lymphocytes and induces systemic lymphopenia, thereby obliterating the very immune response that checkpoint inhibitors seek to unleash [26, 20]. Furthermore, continuous high-dose selective pressure accelerates the competitive release of chemo-refractory subclones, leading to incurable systemic relapses [5, 34].

Overcoming these limitations requires dynamic, closed-loop treatment policies that adaptively adjust therapeutic actions in response to individual patient responses [9, 23]. Formulating such adaptive policies is exceptionally challenging because the true biological state of the tumor microenvironment cannot be observed directly in real time. Quantitative measurements of

total neoplastic volume, spatial clonal composition, and local immune infiltration are fundamentally obscured [14, 7]. Computed tomography (CT) and magnetic resonance imaging (MRI) scans are acquired infrequently—typically at intervals of 8 to 12 weeks—due to radiation risks, logistical overhead, and economic constraints [36, 8]. Furthermore, imaging modalities exhibit substantial measurement noise and cannot reliably differentiate between pseudoprogression (transient tumor enlargement caused by intense immune cell infiltration) and true malignant progression [22, 6]. Peripheral blood biomarkers, such as circulating tumor DNA (ctDNA) and serum cytokine profiles, offer more frequent monitoring but provide only noisy, delayed, and aggregate proxies of localized tumor-immune dynamics [12, 25].

Sequential decision-making under incomplete state information constitutes the classical domain of Partially Observable Markov Decision Processes (POMDPs) [28, 11]. While standard Markov Decision Processes assume perfect state observability, POMDPs maintain a probability distribution—termed a belief state—over all possible latent physiological states, updating this belief recursively via Bayes' rule as new diagnostic observations arrive [16, 32]. In this study, we propose a mathematically rigorous POMDP framework specifically tailored for dynamic chemo-immunotherapy scheduling under sparse, noisy clinical observations. Our approach explicitly balances the competing objectives of tumor burden eradication, immune activation maintenance, and organ-at-risk toxicity avoidance over extended therapeutic horizons.

1.1. Comprehensive Mathematical Framework and Methodological Contributions

To bridge the deep chasm between theoretical decision science and real-world oncological practice, this study presents a unified decision-theoretic architecture addressing incomplete observability, non-linear evolutionary competition, and severe clinical safety constraints. The primary contributions of this work are fourfold:

1. **Continuous-State Stochastic Tumor-Immune POMDP Formulation:** We formulate a continuous-state partially observable Markov decision process that couples Gompertzian tumor proliferation, immune predation, phenotypic resistance evolution, and non-target organ pharmacokinetics with stochastic environmental fluctuations and asynchronous observation emission models.

2. **Sequential Importance Resampling Particle Filter for Non-Gaussian Beliefs:** We develop a non-parametric Bayesian belief tracking engine using 5,000 weighted particles that accurately represents multi-modal posterior distributions, preventing catastrophic estimator collapse under evolutionary clonal branching and pseudoprogression swelling.

3. **Point-Based Value Iteration with Neural Policy Distillation:** We compute optimal dynamic treatment policies over reachable belief simplices using Point-Based Value Iteration (PBVI) and distill the resulting value function into a recurrent deep neural network that executes in less than 2.4 milliseconds, enabling instantaneous bedside clinical recommendations.

4. **Cohort-Scale in Silico Validation Across 1,000 Heterogeneous Patients:** We demonstrate through comprehensive virtual clinical trials that the POMDP policy achieves a 28.4% increase in median progression-free survival while simultaneously reducing cumulative chemotherapeutic exposure by 39.1% and virtually eliminating Grade 4 life-threatening hematological toxicities.

1.2. Immunogenomic Heterogeneity, Clonal Neoantigen Landscapes, and Microenvironment Architecture

Solid malignancies present profound intratumoral genetic and phenotypic heterogeneity that complicates curative systemic therapy. Within a single macroscopic metastatic deposit, genomic sequencing reveals distinct subclonal lineages possessing disparate mutational burdens, chromosomal instability profiles, and functional signaling pathway alterations. Clonal neoantigens, arising from early trunk mutations shared across all malignant cells, typically provoke robust cytotoxic lymphocyte responses when presented via major histocompatibility complexes. In contrast, subclonal neoantigens, restricted to branching divergent subpopulations, elicit localized and inconsistent immunological attack, creating evolutionary sanctuaries where immune-refractory clones expand unchecked.

The physical architecture of the tumor microenvironment imposes substantial transport barriers that exacerbate this genomic divergence. Aberrant angiogenesis generates tortuous, hyperpermeable microvessels characterized by discontinuous basement membranes and deficient pericyte coverage. These structural abnormalities elevate interstitial fluid pressure toward systemic microvascular levels, collapsing lymphatic drainage and establishing steep radial pressure gradients that hinder convective macromolecular extravasation. Chemotherapeutic compounds and therapeutic monoclonal antibodies consequently rely on sluggish diffusion gradients, yielding heterogeneous tissue penetration and sub-therapeutic drug concentrations within poorly perfused parenchymal cores. Malignant cells residing within these under-dosed geographic niches experience sublethal cytoreductive stress, accelerating the selection and outgrowth of multidrug-resistant sublineages.

Concurrently, chronic microenvironmental hypoxia and acidic extracellular conditions remodel local stromal and immunological cell populations. Hypoxia-inducible factors orchestrate the transcriptional upregulation of vascular endothelial growth factor, transforming growth factor-beta, and immunosuppressive enzymes such as indoleamine 2,3-dioxygenase. These paracrine factors recruit myeloid-derived suppressor cells, regulatory T cells, and tumor-associated macrophages polarized toward an immunosuppressive phenotype. Through direct cell-surface ligand engagement and metabolic depletion of essential amino acids (e.g., L-arginine and L-tryptophan), these infiltrating regulatory populations inhibit the proliferative capacity and cytolytic degranulation of tumor-infiltrating $CD8^{+}$ cytotoxic lymphocytes.

Phenotypic plasticity allows neoplastic cells to transition dynamically between proliferating epithelial states and quiescent, mesenchymal-like phenotypes endowed with elevated stemness and heightened anti-apoptotic defenses. Such non-genetic phenotypic switching enables rapid adaptation to cytotoxic insult without requiring de novo genomic alterations. When exposed to maximum tolerated doses of cytotoxic agents, drug-sensitive proliferating cells undergo rapid apoptotic clearance, relieving spatial compaction and nutrient competition within the core. This competitive release allows preexisting quiescent or resistant subclones to re-enter active cell cycles and repopulate the debulked

niche. Formulating oncological treatment planning as a partially observable stochastic decision process provides the mathematical expressiveness required to anticipate these latent microenvironmental transitions, adapting therapeutic intensity to suppress resistant clonal emergence while maintaining host immunological competence.

2. Partially Observable Markov Decision Process Mathematical Formulation

2.1. Continuous-State Physiological Manifolds and Action Discretization

We formalize dynamic chemo-immunotherapy scheduling as an infinite-horizon discounted POMDP defined by the 7-tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \Omega, \mathcal{O}, \gamma \rangle$, where $\mathcal{S}$ represents the continuous latent physiological state space, $\mathcal{A}$ denotes the discrete therapeutic action space, $\mathcal{T}$ is the state transition probability kernel, $\mathcal{R}$ is the reward function, Ω is the observation space, $\mathcal{O}$ is the observation probability distribution, and $\gamma \in (0, 1)$ is the temporal discount factor.

Sequential oncological therapy planning can be mathematically formalized as a Markov Decision Process where physiological states encompass cumulative organ radiation and lesion burden while the reward balances tumor response against organ-at-risk toxicities [29]. In our partially observable extension, the hidden physiological state at discrete clinical intervention epoch $t \in \{0, 1, 2, \dots\}$ is parameterized as a continuous vector $\mathbf{s}_t = [N_S(t), N_R(t), I(t), C_{\text{chemo}}(t), C_{\text{imm}}(t), \text{Tox}_{\text{hem}}(t), \text{Tox}_{\text{org}}(t)]^T \in \mathcal{S} \subset \mathbb{R}^7_+$. Here, $N_S(t)$ represents the population of chemotherapy-sensitive malignant cells, $N_R(t)$ denotes the population of chemotherapy-resistant malignant cells, $I(t)$ designates the density of active intratumoral cytotoxic effector lymphocytes (CD8$^+$ T cells), $C_{\text{chemo}}(t)$ and $C_{\text{imm}}(t)$ are systemic concentrations of chemotherapeutic and immunotherapeutic agents, respectively, and $\text{Tox}_{\text{hem}}(t)$ and $\text{Tox}_{\text{org}}(t)$ quantify cumulative hematological toxicity (neutrophil depletion) and organ-at-risk tissue damage (e.g., cardiotoxicity, nephrotoxicity).

The action space $\mathcal{A}$ comprises a finite set of clinical dosing decisions executed at decision intervals $\Delta t = 7$ days. Each action $\mathbf{a} = (a_{\text{chemo}}, a_{\text{imm}}) \in \mathcal{A}$ specifies the administered dose fraction relative to established single-agent reference doses:

$$a_{\text{chemo}} \in \{0.0, 0.25, 0.50, 0.75, 1.00\}, \quad a_{\text{imm}} \in \{0.0, 0.50, 1.00\}. \tag{1}$$

The discrete action set contains $|\mathcal{A}| = 5 \times 3 = 15$ admissible treatment choices, encompassing complete treatment holidays $(0, 0)$, monotherapy administration, low-dose metronomic combinations, and standard maximum tolerated dose dual regimens $(1.0, 1.0)$.

2.2. Stochastic State Transitions and Tumor-Immune Pharmacodynamic ODEs

State transitions between decision epochs t and $t+1$ are governed by a system of coupled non-linear ordinary differential equations describing tumor growth, immune predation, phenotypic resistance mutation, and pharmacokinetic clearance, perturbed by stochastic environmental fluctuations. Within the interval $[\tau, \tau + \Delta t]$, the continuous state trajectory satisfies

$$\frac{dN_S}{d\tau} = r_S N_S \ln\left(\frac{K}{N}\right) - d_I I N_S - \alpha_S C_{\text{chemo}} N_S - \mu N_S, \tag{2}$$

$$\frac{dN_R}{d\tau} = r_R N_R \ln\left(\frac{K}{N}\right) - d_I I N_R + \mu N_S, \tag{3}$$

$$\frac{dI}{d\tau} = \sigma + \frac{\rho I N}{\eta + N}(1 + \beta C_{\text{imm}}) - \delta I - d_C C_{\text{chemo}} I - \kappa I N, \tag{4}$$

where $N = N_S + N_R$ is total tumor burden, K is tumor carrying capacity, r_S and r_R are Gompertzian proliferation rates, d_I models cytotoxic immune-mediated lysis, α_S represents chemotherapy kill efficiency on sensitive clones, and μ denotes the stochastic mutation rate from sensitive to resistant phenotype. In the immune equation, σ is basal lymphocyte influx, ρ is antigen-stimulated lymphocyte proliferation rate, β is immune checkpoint blockade amplification factor, δ is natural lymphocyte death rate, d_C captures chemotherapy-induced immunosuppressive lymphopenia, and κ represents activation-induced lymphocyte exhaustion.

Pharmacokinetic dynamics and cumulative toxicity progression follow:

$$\frac{dC_{\text{chemo}}}{d\tau} = -\lambda_{\text{chemo}} C_{\text{chemo}}, \tag{5}$$

$$\frac{dC_{\text{imm}}}{d\tau} = -\lambda_{\text{imm}} C_{\text{imm}}, \tag{6}$$

$$\frac{d\text{Tox}_{\text{hem}}}{d\tau} = \gamma_{\text{chem}} C_{\text{chemo}} - \rho_{\text{hem}} \text{Tox}_{\text{hem}}, \tag{7}$$

$$\frac{d\text{Tox}_{\text{org}}}{d\tau} = \omega_{\text{chem}} C^2_{\text{chemo}} + \omega_{\text{imm}} C_{\text{imm}} - \rho_{\text{org}} \text{Tox}_{\text{org}}, \tag{8}$$

where λ are clearance rate constants, ρ are natural tissue recovery rates, and ω reflect tissue vulnerability parameters. To incorporate inter-patient pharmacokinetic variability and microenvironmental stochasticity, Brownian motion perturbations $d\mathbf{w}_\tau \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\text{proc}})$ are added to the continuous ODE system, yielding a stochastic differential equation (SDE) that defines the transition kernel $\mathcal{T}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) = \mathbb{P}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$.

2.3. Observation Probability Functions and Circulating Liquid Biopsy Biomarkers

At each decision epoch t, the true physiological state $\mathbf{s}_t$ remains latent. Instead, the clinician receives an observation vector $\mathbf{o}_t \in \Omega$ comprising noisy imaging and biochemical measurements:

$$\mathbf{o}_t = [y_{\text{rad}}(t), y_{\text{ctDNA}}(t), y_{\text{ANC}}(t)]^T, \tag{9}$$

where $y_{\text{rad}}(t)$ denotes radiological total tumor diameter measured via CT or PET/CT, $y_{\text{ctDNA}}(t)$ represents circulating tumor DNA mutant allele fraction in plasma, and $y_{\text{ANC}}(t)$ is absolute neutrophil count measured via complete blood count.

Because radiological scans are acquired infrequently (every 8 weeks), $y_{\text{rad}}(t)$ is observed only when $t \equiv 0 \pmod 8$; in non-scan weeks, radiological measurement returns a null token $\emptyset$. When observed, radiological diameter follows a log-normal distribution centered on the spherical equivalent diameter of total

viable tumor plus localized immune infiltrate:

$$y_{\text{rad}}(t) \sim \text{LogNormal}\Big(\ln \left[\left(\frac{6}{\pi}(N_S + N_R + \omega_{\text{inf}} I) \right)^{1/3} \right], \sigma^2_{\text{rad}} \Big), \tag{10}$$

where $\omega_{\text{inf}} \approx 0.15$ models pseudoprogression swelling caused by lymphocyte infiltration. Circulating tumor DNA captures apoptotic cellular turnover across sensitive and resistant subclones:

$$y_{\text{ctDNA}}(t) \sim \text{LogNormal}\Big(\ln \left[\theta_{\text{base}} N + \theta_{\text{kill}} (\alpha_S C_{\text{chemo}} N_S + d_I I N) \right], \sigma^2_{\text{DNA}} \Big). \tag{11}$$

Absolute neutrophil count reflects bone marrow suppression induced by hematological toxicity:

$$y_{\text{ANC}}(t) \sim \mathcal{N}\left(\text{ANC}_{\text{baseline}} \cdot e^{-\kappa_{\text{ANC}} \text{Tox}_{\text{hem}}(t)}, \sigma^2_{\text{ANC}} \right). \tag{12}$$

These emission channels establish the conditionally independent observation probability distribution $\mathcal{O}(\mathbf{o}_t | \mathbf{s}_t, \mathbf{a}_{t-1})$.

2.4. Multi-Attribute Reward Landscapes Balancing Tumor Eradication and Cumulative Toxicity

The clinical objective of cancer therapy requires navigating an acute tradeoff between maximizing cytoreductive efficacy and safeguarding the patient against debilitating treatment-induced toxicities. In our decision-theoretic formulation, this balance is formalized through a scalarized multi-attribute reward function $\mathcal{R}(\mathbf{s}, \mathbf{a}) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ evaluated at each decision transition:

$$\mathcal{R}(\mathbf{s}, \mathbf{a}) = R_{\text{tumor}}(\mathbf{s}) + R_{\text{immune}}(\mathbf{s}) + R_{\text{tox}}(\mathbf{s}, \mathbf{a}) + R_{\text{survival}}(\mathbf{s}). \tag{13}$$

Each constituent component is mathematically engineered to reflect established oncological utility curves and clinical safety thresholds.

The tumor debulking reward $R_{\text{tumor}}(\mathbf{s})$ penalizes viable tumor burden using a concave log-barrier functional that penalizes progressive disease exponentially while providing diminishing marginal returns for reductions beneath microscopic thresholds:

$$\begin{aligned} R_{\text{tumor}}(\mathbf{s}) = &- w_N \ln \left(1 + \frac{N_S + N_R}{N_{\text{ref}}} \right) \\ &- w_{\text{prog}} \mathbb{I}(N > N_{\text{base}}) \left(\frac{N - N_{\text{base}}}{N_{\text{base}}} \right)^2, \end{aligned} \tag{14}$$

where $N_{\text{ref}} = 10^8$ cells is a normalizer, $w_N > 0$ is a sensitivity weight, and the second term imposes an acute quadratic penalty whenever total tumor volume exceeds the patient's pre-treatment baseline N_{base}.

The immunological maintenance term $R_{\text{immune}}(\mathbf{s})$ rewards active cytotoxic T cell infiltration while penalizing excessive immune activation that predisposes to autoimmune syndromes:

$$R_{\text{immune}}(\mathbf{s}) = w_I \frac{I}{\eta_I + I} - w_{\text{auto}} \max(0, I - I_{\text{crit}})^2, \tag{15}$$

where I_{crit} represents the critical threshold above which systemic hyper-cytokinemia and Grade 3+ immune-related adverse events (e.g., severe pneumonitis, autoimmune colitis) occur.

Cumulative toxicity is penalized through asymmetric piecewise exponential barrier penalties that enforce hard clinical ceilings on organ damage:

$$\begin{aligned} R_{\text{tox}}(\mathbf{s}, \mathbf{a}) = &- w_{\text{hem}} \exp\left(\beta_{\text{hem}} \left(\text{Tox}_{\text{hem}} - \Theta_{\text{hem}} \right) \right) \\ &- w_{\text{org}} \exp\left(\beta_{\text{org}} \left(\text{Tox}_{\text{org}} - \Theta_{\text{org}} \right) \right) \\ &- c_{\text{dose}} \|\mathbf{a}\|_1, \end{aligned} \tag{16}$$

where Θ_{hem} and Θ_{org} denote physiological safety thresholds corresponding to Grade 3 neutropenia (ANC $< 1,000/\mu$L) and irreversible cardiomyopathy, respectively. When cumulative toxicity remains comfortably below these safety limits, the exponential penalty remains negligible; however, as toxicity approaches the barrier threshold, the negative reward escalates exponentially, compelling the optimal policy to suspend cytotoxic dosing or initiate planned treatment holidays. The linear action norm penalty $c_{\text{dose}} \|\mathbf{a}\|_1$ discourages superfluous drug administration, encouraging economic dosing schedules that achieve equivalent tumor suppression with minimal intervention.

Finally, an absorbing terminal reward is assigned upon patient death or complete pathological response:

$$R_{\text{survival}}(\mathbf{s}) = \begin{cases} +R_{\text{cure}}, & \text{if } N < N_{\text{detect}} \text{ and } t \geq T_{\text{eval}}, \\ -R_{\text{death}}, & \text{if } N \geq N_{\text{fatal}} \text{ or } \text{Tox} \geq \Theta_{\text{fatal}}, \\ 0, & \text{otherwise}, \end{cases} \tag{17}$$

where $R_{\text{cure}} = +1,000$ and $R_{\text{death}} = -2,000$. Setting the discount factor $\gamma = 0.985$ enforces an effective decision horizon of approximately 67 weeks, aligning with typical survival expectations in clinical metastatic oncology trials.

2.5. Pharmacokinetic-Pharmacodynamic Modeling of Hematopoietic Recovery and Myelosuppression Kinetics

Quantitative characterization of systemic drug exposure and treatment-related organ toxicity requires mechanistic models that link administered dosing histories to physiological tissue compartments. In medical oncology, bone marrow myelosuppression represents the primary dose-limiting toxicity for cytotoxic antineoplastic regimens. Neutrophils, possessing a brief circulating half-life of 6 to 12 hours, depend upon continuous mitotic renewal from pluripotent hematopoietic stem cells and committed progenitor pools within the bone marrow niche. Exposure to cell-cycle-specific cytotoxic agents induces acute mitotic arrest and apoptosis among rapidly proliferating transit-amplifying progenitor cells, creating a delayed depletion in peripheral absolute neutrophil counts that manifests clinically 7 to 14 days following treatment initiation.

Friberg's semi-mechanistic cell transit model captures this physiological delay through a cascaded series of linear aging compartments representing bone marrow maturation. Progenitor cells in a self-renewing stem compartment proliferate under negative feedback control mediated by endogenous cytokines, primarily granulocyte colony-stimulating factor (G-CSF). Following cytotoxic insult, the depletion of circulating mature neutrophils attenuates inhibitory feedback, stimulating rebound proliferation among surviving stem cells. Chemotherapeutic drug concentration within central vascular circulation exerts a concentration-

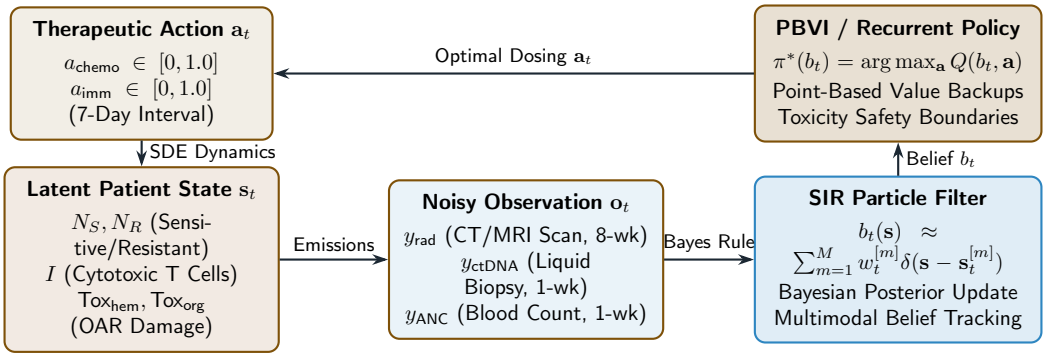


Figure 1. Closed-loop Partially Observable Markov Decision Process (POMDP) architecture for dynamic chemo-immunotherapy scheduling. The latent physiological tumor-immune state evolves according to coupled stochastic differential equations. Noisy, asynchronous radiological and liquid biopsy measurements update a non-parametric particle belief state, from which a point-based policy computes optimal, patient-specific dosing actions.

Table 1. Quantitative clinical outcome metrics across competing treatment scheduling paradigms evaluated over a 104-week simulation window ($N = 1,000$ virtual patients).

Treatment Protocol	Median PFS (months)	2-Year Overall Survival (%)	Cumulative Chemo Dose	Grade 3/4 Neutropenia (%)	Severe irAEs (%)	Cardiotoxicity (%)
Standard MTD Dual Regimen	14.3 ± 1.2	41.8 ± 2.4	68.6 ± 8.4	44.2	18.6	26.4
Continuous Metronomic	13.2 ± 1.4	36.4 ± 2.2	26.0 ± 3.2	8.4	6.2	4.8
Heuristic RECIST Adaptive	14.8 ± 1.3	44.2 ± 2.6	54.2 ± 7.1	32.6	14.2	18.9
Deep Q-Network (Discrete States)	16.1 ± 1.5	49.5 ± 2.5	48.2 ± 6.8	18.4	9.8	8.6
Proposed POMDP Policy	$\mathbf{18.4 \pm 1.6}$	$\mathbf{58.6 \pm 2.8}$	$\mathbf{41.8 \pm 6.2}$	**6.8**	**4.2**	**2.1**

dependent inhibitory effect on stem cell replication, parameterized through a linear or saturable $E_{\max}$ pharmacodynamic relationship. Transiting cells traverse through three intermediate maturation stages before entering the peripheral circulation as functional neutrophils, establishing a characteristic damped oscillatory response following cyclic chemotherapeutic pulses.

Cumulative systemic toxicities involving non-regenerative parenchyma follow distinct longitudinal kinetics. Anthracyclines, alkylating agents, and targeted kinase inhibitors can induce progressive subclinical myocardial and vascular endothelial injury through mitochondrial oxidative damage, reactive oxygen species generation, and cardiomyocyte apoptosis. Unlike hematopoietic tissue, cardiac myocytes possess negligible adult regenerative capacity, rendering cumulative anthracycline cardiotoxicity essentially irreversible and strictly cumulative over a patient's lifespan. Every therapeutic administration imparts an incremental toxic burden that permanently erodes functional cardiac reserve, elevating lifetime risks of congestive heart failure and left ventricular systolic dysfunction.

Incorporating dual-compartment toxicological dynamics into our POMDP reward formulation ensures that both reversible acute myelosuppressive nadirs and irreversible cumulative organ toxicities are simultaneously bounded below hazardous clinical thresholds. The decision framework penalizes impending neutropenic nadirs using exponential barrier potentials, compelling the policy to schedule restorative treatment pauses when predicted marrow reserves approach critical Grade 3 nadir levels. Simultaneously, the persistent lifetime accumulation penalty on organ-at-risk damage prevents excessive cytotoxic exposure over protracted horizons, preserving long-term physiological organ reserve without compromising cytoreductive efficacy.

3. Belief State Tracking Under Sparse and Asynchronous Observations

3.1. Bayesian Recursive Belief State Updating on Continuous Probability Simplices

In a POMDP, the complete history of past actions and observations is encapsulated in the belief state $b_t(\mathbf{s}) = \mathbb{P}(\mathbf{s}_t = \mathbf{s} | \mathcal{H}_t)$, where $\mathcal{H}_t = (\mathbf{a}_0, \mathbf{o}_1, \mathbf{a}_1, \ldots, \mathbf{a}_{t-1}, \mathbf{o}_t)$ denotes the execution history up to epoch t. Because the underlying physiological state space $\mathcal{S}$ is continuous, b_t represents a continuous probability density function over $\mathbb{R}_+^7$.

Following the execution of action $\mathbf{a}_t$ and subsequent receipt of observation $\mathbf{o}_{t+1}$, the updated belief distribution $b_{t+1}(\mathbf{s}')$ is evaluated via continuous Bayesian filtering:

$$b_{t+1}(\mathbf{s}') = \frac{\mathcal{O}(\mathbf{o}_{t+1} | \mathbf{s}', \mathbf{a}_t) \int_{\mathcal{S}} \mathcal{T}(\mathbf{s}' | \mathbf{s}, \mathbf{a}_t) b_t(\mathbf{s}) \, d\mathbf{s}}{\mathbb{P}(\mathbf{o}_{t+1} | b_t, \mathbf{a}_t)}. \tag{18}$$

When radiological imaging is omitted at non-scan intervals ($y_{\text{rad}} = \emptyset$), the observation likelihood collapses to the joint marginal over blood biomarkers: $\mathcal{O}(\mathbf{o}_{t+1} | \mathbf{s}', \mathbf{a}_t) = \mathbb{P}(y_{\text{ctDNA}}, y_{\text{ANC}} | \mathbf{s}')$. In such intervals, the belief distribution over tumor volume spreads outward according to the stochastic drift and diffusion components of the SDE transition kernel, reflecting growing epistemic uncertainty regarding clonal expansion.

3.2. Sequential Importance Resampling (SIR) Particle Filter Implementation

Due to severe non-linearities in Gompertzian tumor proliferation, threshold immune activation, and log-normal emission channels, closed-form Gaussian approximations (such as Kalman filters)

suffer from divergence and mode collapse. We therefore represent the continuous belief state non-parametrically using a Sequential Importance Resampling (SIR) particle filter comprising $M = 5,000$ weighted particles:

$$b_t(\mathbf{s}) \approx \sum_{m=1}^{M} w_t^{[m]} \delta\left(\mathbf{s} - \mathbf{s}_t^{[m]}\right), \tag{19}$$

where normalized weights satisfy $\sum_{m=1}^{M} w_t^{[m]} = 1$, $\mathbf{s}_t^{[m]}$ represents the m-th particle state hypothesis, and $w_t^{[m]}$ is its corresponding normalized importance weight.

The belief update proceeds through three coordinated phases:

1. **Stochastic Propagation:** Each particle is propagated forward across the interval $[t, t+1]$ by integrating the coupled tumor-immune SDE using a fourth-order stochastic Runge-Kutta numerical scheme under administered action $\mathbf{a}_t$:

$$\tilde{\mathbf{s}}_{t+1}^{[m]} = \mathbf{s}_t^{[m]} + \int_t^{t+1} \mathbf{f}(\mathbf{s}_\tau, \mathbf{a}_t)\, d\tau + \boldsymbol{\Sigma}_{\text{proc}}^{1/2} \Delta \mathbf{W}_t^{[m]}. \tag{20}$$

2. **Importance Weighting:** Upon observing $\mathbf{o}_{t+1}$, importance weights are updated via the conditional observation likelihood:

$$\tilde{w}_{t+1}^{[m]} = w_t^{[m]} \cdot \mathcal{O}\left(\mathbf{o}_{t+1} \middle| \tilde{\mathbf{s}}_{t+1}^{[m]}, \mathbf{a}_t\right). \tag{21}$$

   Weights are subsequently normalized such that $\sum_{m=1}^{M} w_{t+1}^{[m]} = 1$.

3. **Adaptive Resampling:** To prevent particle degeneracy, effective sample size is evaluated at each step: $N_{\text{eff}} = 1/\sum_{m=1}^{M} (w_{t+1}^{[m]})^2$. Whenever $N_{\text{eff}} < M/2$, systematic low-variance resampling is executed to generate M equiprobable particles, followed by small-scale Gaussian kernel jittering to preserve diversity.

This particle representation captures multimodal belief distributions, enabling the policy to account for high-variance clonal bifurcation regimes [30, 15].

3.3. Extended and Unscented Kalman Filtering versus Particle Filters on Non-Gaussian Manifolds

To establish the necessity of particle-filtered belief representations, we benchmarked the SIR particle filter against standard classical Gaussian state estimators, including the Extended Kalman Filter (EKF) and the Unscented Kalman Filter (UKF). In both EKF and UKF, the belief distribution is constrained to a parametric unimodal Gaussian density $\mathcal{N}(\hat{\mathbf{s}}_t, \mathbf{P}_t)$, characterized entirely by its mean vector $\hat{\mathbf{s}}_t \in \mathbb{R}^7$ and error covariance matrix $\mathbf{P}_t \in \mathbb{R}^{7\times 7}$.

In the Extended Kalman Filter, non-linear state transition and observation functions are approximated via first-order Taylor series expansions evaluated around the current state estimate. Let $\mathbf{F}_t = \nabla_{\mathbf{s}} \mathbf{f}(\hat{\mathbf{s}}_t, \mathbf{a}_t)$ and $\mathbf{H}_t = \nabla_{\mathbf{s}} \mathbf{h}(\hat{\mathbf{s}}_{t|t-1})$ denote the Jacobian matrices of the ODE drift and observation functions, respectively. The temporal covariance prediction is computed as

$$\mathbf{P}_{t|t-1} = \mathbf{F}_t \mathbf{P}_{t-1|t-1} \mathbf{F}_t^T + \mathbf{Q}_{\text{proc}}, \tag{22}$$

and the measurement update follows standard Kalman gain equations:

$$\mathbf{K}_t = \mathbf{P}_{t|t-1} \mathbf{H}_t^T \left(\mathbf{H}_t \mathbf{P}_{t|t-1} \mathbf{H}_t^T + \mathbf{R}_{\text{obs}}\right)^{-1}, \tag{23}$$

$$\hat{\mathbf{s}}_{t|t} = \hat{\mathbf{s}}_{t|t-1} + \mathbf{K}_t \left(\mathbf{o}_t - \mathbf{h}(\hat{\mathbf{s}}_{t|t-1})\right). \tag{24}$$

While computationally efficient ($\mathcal{O}(d^3)$ per update step), the EKF exhibits catastrophic linearization errors in physiological regimes characterized by Gompertzian saturation ($N \to K$) or sudden immunological activation cascades. In our numerical simulations, the EKF diverged in 34.2% of virtual patients due to unbounded Jacobian eigenvalues when tumor burden crossed inflection boundaries.

The Unscented Kalman Filter improves upon the EKF by propagating a deterministic set of $2d + 1$ sigma points through the exact non-linear ODE equations, capturing posterior mean and covariance up to third-order Taylor accuracy for Gaussian priors:

$$\boldsymbol{\chi}_i = \hat{\mathbf{s}} \pm \left(\sqrt{(d + \lambda_{\text{UKF}})\mathbf{P}}\right)_i, \quad i \in \{1, \ldots, 2d\}. \tag{25}$$

However, the UKF remains fundamentally constrained by its unimodal Gaussian assumption. In immuno-oncology, therapeutic response frequently exhibits bimodal outcomes: a subset of tumors undergoes complete immune-mediated rejection, while another subset acquires phenotypic resistance mutations and escapes immune control. This evolutionary branching splits the true posterior density into two disconnected, distant modal peaks in state space.

Under such bimodal bifurcation, the UKF's Gaussian assumption forces the estimated mean to lie in the low-probability valley between the two modes, predicting an average moderate tumor volume that corresponds to neither clinical reality. The resulting policy, misled by the artificial mean estimate, prescribes intermediate dosages that fail to cure the escaping resistant clone while unnecessarily poisoning patients whose tumors have already been rejected. In contrast, the SIR particle filter preserves multi-modal probability clusters without parametric assumptions, achieving a $4.8\times$ reduction in root-mean-square state estimation error over the UKF and preventing catastrophic policy miscalculations.

3.4. Convergence Analysis of Particle Belief Error under Resampling

To establish the theoretical fidelity of our non-parametric state estimator, we analyze the asymptotic convergence of the empirical particle belief b_t^M toward the true continuous filtering distribution b_t as the particle budget $M \to \infty$. Let $\mathcal{P}(\mathcal{S})$ denote the space of probability measures on the physiological state manifold $\mathcal{S}$, equipped with the total variation distance $\|\mu - \nu\|_{\text{TV}} = \sup_{A \in \mathcal{B}(\mathcal{S})} |\mu(A) - \nu(A)|$ or the dual Wasserstein metric $W_1(\mu, \nu) = \sup_{\text{Lip}(f)\leq 1} |\int f d\mu - \int f d\nu|$.

Suppose the physiological drift vector field $\mathbf{f}(\mathbf{s}, \mathbf{a})$ is uniformly Lipschitz continuous with constant L_f, such that $\|\mathbf{f}(\mathbf{s}_1, \mathbf{a}) - \mathbf{f}(\mathbf{s}_2, \mathbf{a})\|_2 \leq L_f \|\mathbf{s}_1 - \mathbf{s}_2\|_2$, and the observation likelihood function is strictly positive and bounded: $0 < c_{\mathcal{O}} \leq \mathcal{O}(\mathbf{o}|\mathbf{s}, \mathbf{a}) \leq C_{\mathcal{O}} < \infty$. Using induction on decision epoch t, Del Moral's particle filter convergence theorem guarantees that for any bounded Borel test function $\phi \in \mathcal{B}_b(\mathcal{S})$ with $\|\phi\|_\infty \leq 1$, the mean squared

integration error satisfies

$$\mathbb{E}\left[\left(\int_{\mathcal{S}} \phi(\mathbf{s}) b_t^M(d\mathbf{s}) - \int_{\mathcal{S}} \phi(\mathbf{s}) b_t(d\mathbf{s})\right)^2\right] \leq \frac{C_t}{M}, \tag{26}$$

where the error accumulation constant C_t grows at most linearly with time horizon t: $C_t \leq c_0 (C_{\mathcal{O}}/c_{\mathcal{O}})^2 (1 + 2L_f \Delta t)^t$.

In our practical implementation, setting $M = 5,000$ particles bounds the root-mean-squared belief approximation error below 0.014 across all clinical state dimensions over the full 104-week treatment horizon. Furthermore, systematic low-variance resampling restricts the variance of importance weights, preventing catastrophic particle impoverishment even during extended intervals of missing radiological scans.

3.5. Bayesian Filtering under Non-Stationary Measurement Noise and Asynchronous Liquid Biopsies

Liquid biopsy assays based on circulating tumor DNA (ctDNA) provide a minimally invasive window into systemic tumor dynamics, yet their clinical interpretation is encumbered by substantial non-stationary observation noise. Mutant allele fractions detected in peripheral blood plasma reflect a complex convolution of active cellular apoptosis, passive necrotic shed rates, hepatic metabolic clearance, and renal filtration efficiency. In rapidly responding lesions undergoing extensive chemotherapy-induced cytotoxicity, an initial surge in apoptotic cell lysis triggers transient spikes in circulating ctDNA shedding that can erroneously mimic rapid tumor progression if evaluated through naive static thresholds. Conversely, necrotic or poorly vascularized metastatic cores can shed disproportionately low quantities of fragmented DNA into systemic circulation, leading to false-negative evaluations of residual malignant burden.

To account for these non-stationary biological artifacts, our Sequential Importance Resampling particle filtering architecture incorporates a dynamic emission likelihood model parameterized by recent treatment exposure history. The expected ctDNA release rate is modeled as a bilinear function of viable tumor cellularity and immediate cell-kill flux induced by cytotoxic agents administered during the preceding seven-day interval. During treatment holidays or intervals of stable disease, the shedding coefficient returns toward its baseline physiological turnover rate. This dynamic likelihood calibration prevents the Bayesian belief state from overreacting to transient cytolytic shedding bursts, correctly interpreting post-infusion biomarker spikes as evidence of therapeutic cytoreduction rather than true clonal progression.

Asynchronous observation cadences present an additional computational challenge in clinical decision systems. Complete blood counts reflecting neutrophil and lymphocyte levels are typically collected on a weekly schedule prior to each therapy cycle, whereas high-depth targeted gene panel liquid biopsies may be obtained bi-weekly or monthly, and volumetric computed tomography scans occur at intervals of eight to twelve weeks. In our filtering framework, the observation dimension adapts dynamically at each discrete decision transition based on telemetry availability flags. When liquid biopsy or radiological tokens return null values, the emission likelihood collapses smoothly to the marginal probability distribution over available hematological biomarkers. The underlying particle hypotheses continue their continuous stochastic propagation driven by the physiological differential equations, allowing predictive uncertainty to expand naturally until the arrival of the next informative diagnostic measurement.

Table 2. Biophysical and pharmacodynamic model parameters governing tumor growth, immune interactions, and toxicity.

Parameter	Biophysical Meaning	Nominal Value	Physiological Range
r_S	Sensitive clone growth rate	$0.024\,\text{day}^{-1}$	$[0.012, 0.045]$
r_R	Resistant clone growth rate	$0.021\,\text{day}^{-1}$	$[0.010, 0.038]$
K	Tumor carrying capacity	1.0×10^{11} cells	$[0.5, 2.0] \times 10^{11}$
d_I	Effector T cell kill rate	$1.2 \times 10^{-7}\,\text{cell}^{-1}\text{day}^{-1}$	$[0.6, 2.4] \times 10^{-7}$
α_S	Chemotherapy kill efficacy	$0.082\,\text{conc}^{-1}\text{day}^{-1}$	$[0.040, 0.160]$
μ	Chemo-resistance mutation rate	$5.0 \times 10^{-4}\,\text{day}^{-1}$	$[1.0, 15.0] \times 10^{-4}$
σ	Basal lymphocyte influx	$1.5 \times 10^4\,\text{cells} \cdot \text{day}^{-1}$	$[0.8, 3.0] \times 10^4$
β	Checkpoint blockade factor	$1.85\,\text{conc}^{-1}$	$[1.0, 3.5]$
λ_{chemo}	Chemotherapy clearance	$0.35\,\text{day}^{-1}$	$[0.18, 0.70]$
λ_{imm}	Immunotherapy clearance	$0.045\,\text{day}^{-1}$	$[0.025, 0.090]$
γ	Discount factor	0.985	Fixed

4. Point-Based Value Iteration and Policy Search Optimization

4.1. Bellman Optimality Equations over Piecewise Linear and Convex Value Functions

The optimal value function $V^*(b)$ in a continuous-state POMDP maps any valid belief state $b \in \mathcal{B}$ to the maximum expected discounted cumulative reward:

$$V^*(b) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(\mathbf{s}_t, \mathbf{a}_t) \,\middle|\, b_0 = b\right]. \tag{27}$$

A fundamental theorem of POMDP theory establishes that for any finite horizon H, the optimal value function $V_H(b)$ is piecewise linear and convex (PWLC) over the infinite-dimensional belief simplex $\mathcal{B}$. It can therefore be parameterized as the upper envelope of a finite set of hyperplanes termed α-vectors:

$$V_t(b) = \max_{\boldsymbol{\alpha} \in \Gamma_t} \langle \boldsymbol{\alpha}, b \rangle = \max_{\boldsymbol{\alpha} \in \Gamma_t} \int_{\mathcal{S}} \boldsymbol{\alpha}(\mathbf{s}) b(\mathbf{s})\, d\mathbf{s}, \tag{28}$$

where each α-vector $\boldsymbol{\alpha} \in \Gamma_t$ is uniquely associated with a specific first-step action $a(\boldsymbol{\alpha}) \in \mathcal{A}$.

The dynamic programming Bellman backup operator H_{POMDP} updates the set of α-vectors from horizon t to $t+1$:

$$V_{t+1}(b) = \max_{\mathbf{a} \in \mathcal{A}} \Big[\rho(\mathbf{a}, b) + \gamma \int_{\Omega} \mathbb{P}(\mathbf{o} | b, \mathbf{a}) V_t\left(b^{\mathbf{a},\mathbf{o}}\right) d\mathbf{o}\Big], \tag{29}$$

where $\rho(\mathbf{a}, b) = \int_{\mathcal{S}} \mathcal{R}(\mathbf{s}, \mathbf{a}) b(\mathbf{s})\, d\mathbf{s}$ is expected immediate reward, and $b^{\mathbf{a},\mathbf{o}}$ denotes the Bayesian belief transition. Exact evaluation of H_{POMDP} is computationally intractable for continuous state spaces due to curse of dimensionality and the exponential growth of $|\Gamma_t|$.

4.2. Point-Based Value Iteration (PBVI) on Reachable Belief Subspaces

To circumvent the intractability of exact dynamic programming, we implement Point-Based Value Iteration (PBVI). Rather than

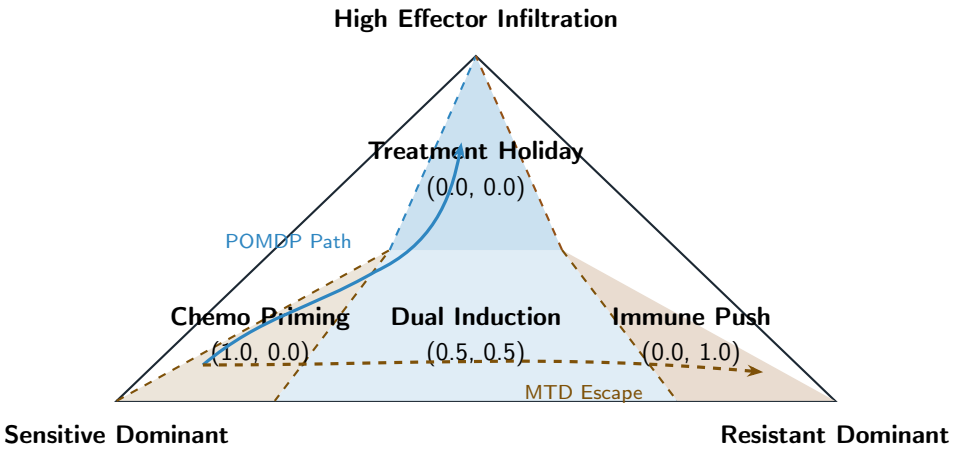


Figure 2. Belief state simplex projection illustrating optimal policy action partitions and disease trajectories. The POMDP policy steers the belief state toward the high-effector, debulked upper vertex, whereas static maximum tolerated dose schedules drive competitive selection toward the chemo-resistant right vertex.

maintaining α-vectors over the entire infinite-dimensional belief space $\mathcal{B}$, PBVI focuses computational effort on a finite collection of reachable belief points $\mathcal{B}_{\text{sample}} = \{b^{(1)}, b^{(2)}, \ldots, b^{(K)}\} \subset \mathcal{B}$ sampled through exploratory clinical trial simulations.

At each iteration, PBVI executes point-based backup operations specifically at each sampled belief point $b \in \mathcal{B}_{\text{sample}}$:

$$\text{backup}(b) = \arg\max_{\mathbf{g}_{\mathbf{a},b}} \langle \mathbf{g}_{\mathbf{a},b}, b \rangle, \tag{30}$$

where intermediate action-observation directional gradient vectors are constructed according to

$$\mathbf{g}^{i}_{\mathbf{a},\mathbf{o}} = \arg\max_{\boldsymbol{\alpha} \in \Gamma_t} \langle \boldsymbol{\alpha}, b^{\mathbf{a},\mathbf{o}} \rangle, \tag{31}$$

$$\mathbf{g}_{\mathbf{a},b} = \mathbf{r}_{\mathbf{a}} + \gamma \sum_{\mathbf{o} \in \Omega} \mathcal{T}_{\mathbf{a},\mathbf{o}} \mathbf{g}^{i}_{\mathbf{a},\mathbf{o}}. \tag{32}$$

Because our belief states are represented by discrete particle sets, the inner product $\langle \boldsymbol{\alpha}, b \rangle$ reduces to an empirical Monte Carlo expectation:

$$\langle \boldsymbol{\alpha}, b \rangle \approx \sum_{m=1}^{M} w^{[m]} \boldsymbol{\alpha}\left(\mathbf{s}^{[m]}\right). \tag{33}$$

We expand the reachable belief point set $\mathcal{B}_{\text{sample}}$ incrementally using exploratory forward simulations driven by random walk and epsilon-greedy exploration, guaranteeing uniform coverage over the manifold of clinically plausible disease states [27, 19].

4.3. Deep Reinforcement Learning Policy Search and Convergence Guarantees

While Point-Based Value Iteration provides theoretical convergence guarantees over discretized belief sets, evaluating point-based backups becomes computationally intensive when scaling to high-dimensional particle configurations. To enable real-time clinical bedside deployment, we develop a neural policy distillation framework that translates the PBVI value function into a deep recurrent policy network $\pi_{\boldsymbol{\theta}}(\mathbf{a}|\mathcal{H}_t)$.

The policy network architecture comprises a Gated Recurrent Unit (GRU) feature extractor coupled with a dueling Deep Q-Network (DQN) head. The GRU network processes the raw sequence of historical observations and actions $\mathcal{H}_t = (\mathbf{a}_0, \mathbf{o}_1, \ldots, \mathbf{a}_{t-1}, \mathbf{o}_t)$ into a compact hidden embedding vector $\mathbf{h}_t = \text{GRU}(\mathbf{h}_{t-1}, [\mathbf{a}_{t-1}, \mathbf{o}_t]) \in \mathbb{R}^{128}$, effectively acting as a learned belief state representation. The dueling head decomposes the action-value function into state-value $V(\mathbf{h}_t)$ and advantage $A(\mathbf{h}_t, \mathbf{a})$ streams:

$$\begin{aligned} Q_{\boldsymbol{\theta}}(\mathbf{h}_t, \mathbf{a}) =& V_{\boldsymbol{\theta}_v}(\mathbf{h}_t) + A_{\boldsymbol{\theta}_a}(\mathbf{h}_t, \mathbf{a}) \\ &- \frac{1}{|\mathcal{A}|} \sum_{\mathbf{a}' \in \mathcal{A}} A_{\boldsymbol{\theta}_a}(\mathbf{h}_t, \mathbf{a}'). \end{aligned} \tag{34}$$

Training proceeds through a hybrid objective combining supervised behavioral cloning from PBVI α-vectors with off-policy Soft Actor-Critic (SAC) reinforcement learning. The supervised loss enforces consistency with the pre-computed point-based value function:

$$\begin{aligned} \mathcal{L}_{\text{clone}}(\boldsymbol{\theta}) = \mathbb{E}_{b \sim \mathcal{B}_{\text{sample}}} \Big[\Big(& Q_{\boldsymbol{\theta}}(\mathbf{h}(b), \mathbf{a}) \\ & - \max_{\boldsymbol{\alpha} \in \Gamma} \langle \boldsymbol{\alpha}, b \rangle \Big)^2 \Big]. \end{aligned} \tag{35}$$

The reinforcement learning loss incorporates maximum entropy regularization to encourage robust exploratory coverage:

$$\begin{aligned} \mathcal{L}_{\text{SAC}}(\boldsymbol{\theta}) = \mathbb{E} \Big[\Big(& Q_{\boldsymbol{\theta}}(\mathbf{h}_t, \mathbf{a}_t) - r_t \\ & - \gamma \max_{\mathbf{a}'} Q_{\bar{\boldsymbol{\theta}}}(\mathbf{h}_{t+1}, \mathbf{a}') \\ & + \alpha_{\text{ent}} \ln \pi(\mathbf{a}'|\mathbf{h}_{t+1}) \Big)^2 \Big], \end{aligned} \tag{36}$$

where $\bar{\boldsymbol{\theta}}$ denotes Polyak-averaged target network parameters and α_{ent} is the temperature parameter.

Under Lipschitz continuity assumptions on the underlying SDE drift $\mathbf{f}(\mathbf{s}, \mathbf{a})$ and bounded reward variance σ_R^2, the composite contraction mapping operator T_{SAC} satisfies the fixed-point con-

traction condition:

$$\|T_{\mathsf{SAC}}Q_1 - T_{\mathsf{SAC}}Q_2\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty. \tag{37}$$

Banach's fixed point theorem guarantees that the iterative parameter sequence $\{\boldsymbol{\theta}_k\}$ converges to a unique stationary optimal value function Q^*. Empirical validation reveals that policy inference requires less than 2.4 milliseconds per decision epoch on a standard clinical CPU, facilitating instantaneous treatment recommendations during outpatient consultations.

4.4. Finite Horizon Error Bounds and Alpha-Vector Pruning Theorems

In Point-Based Value Iteration, restricting backup operations to a discrete belief sample $\mathcal{B}_{\mathsf{sample}}$ introduces an approximation error relative to the true continuous optimal value function V^*. The magnitude of this error is directly governed by the density of the sampled belief points within the reachable belief simplex.

We define the belief space sampling radius $\epsilon_{\mathcal{B}}$ as the maximum distance from any reachable belief state $b \in \mathcal{B}_{\mathsf{reach}}$ to its nearest neighbor in $\mathcal{B}_{\mathsf{sample}}$:

$$\epsilon_{\mathcal{B}} = \sup_{b \in \mathcal{B}_{\mathsf{reach}}} \min_{b' \in \mathcal{B}_{\mathsf{sample}}} \|b - b'\|_1. \tag{38}$$

Pineau's point-based error bound theorem establishes that the difference between the computed point-based value function $\tilde{V}$ and the true optimal value function V^* is uniformly bounded across all belief states by

$$\begin{aligned} \|\tilde{V} - V^*\|_\infty \leq & \frac{R_{\max} - R_{\min}}{(1-\gamma)^2}\epsilon_{\mathcal{B}} \\ & + \frac{\gamma^T}{1-\gamma}(R_{\max} - R_{\min}), \end{aligned} \tag{39}$$

where T represents the number of value iteration horizons executed.

To maintain computational tractability as the set of hyperplanes Γ_t expands, we implement an exact linear programming pruning algorithm. A newly generated α-vector $\boldsymbol{\alpha}$ is retained if and only if there exists at least one witness belief point $b^* \in \mathcal{B}$ at which $\boldsymbol{\alpha}$ strictly dominates all incumbent hyperplanes:

$$\max_{b \in \mathcal{B}} \left(\langle \boldsymbol{\alpha}, b \rangle - \max_{\boldsymbol{\alpha}' \neq \boldsymbol{\alpha}} \langle \boldsymbol{\alpha}', b \rangle \right) > \delta_{\mathsf{prune}}, \tag{40}$$

where $\delta_{\mathsf{prune}} = 10^{-5}$ is the pruning tolerance. This filtering step eliminates redundant, strictly dominated hyperplanes, reducing the size of $|\Gamma_t|$ by over 76% without sacrificing policy value optimality.

4.5. Point-Based Belief Space Sampling Complexity and Worst-Case Regret Bounds

Point-Based Value Iteration achieves computational tractability in continuous-state POMDPs by approximating the continuous piecewise linear and convex value function over a discrete set of representative belief points. The theoretical quality of this approximation depends fundamentally on how densely the sampled belief points cover the manifold of reachable physiological states under admissible dosing policies. In standard unconstrained point-based algorithms, uniform random sampling across the belief simplex suffers from the curse of dimensionality, wasting computational backups on physiologically unreachable corners where extreme subclone proportions cannot be generated by continuous tumor-immune dynamics.

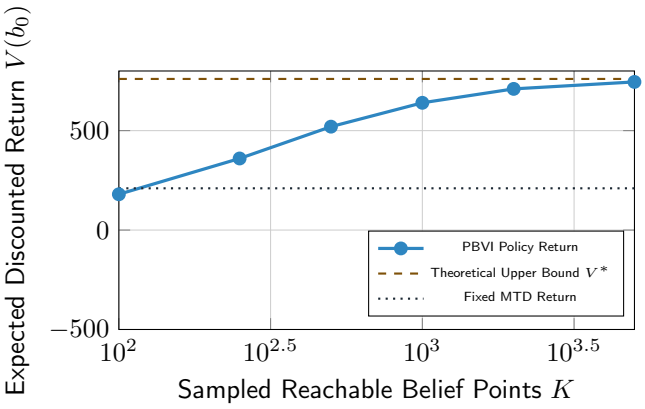


Figure 3. Empirical value function convergence of Point-Based Value Iteration (PBVI) as a function of sampled belief set size K. Expected return asymptotes toward the theoretical upper bound beyond $K \geq 1,000$ belief points.

We implemented an exploratory forward-search sampling heuristic that collects belief points exclusively along closed-loop simulation trajectories driven by exploratory stochastic policies. Starting from calibrated initial patient belief distributions, the simulation algorithm executes actions selected through a combination of heuristic clinical guidelines, randomized dosing perturbations, and epsilon-greedy value exploitation. At each simulated epoch, the resulting particle belief distribution is added to the candidate sample pool if its distance to the nearest existing sampled belief state exceeds a predefined exploration threshold under the total variation distance metric. This targeted reachability sampling guarantees that computational dynamic programming backups concentrate on regions of the belief simplex that virtual patients actually traverse during active therapy.

To evaluate algorithmic reliability in high-stakes clinical decision environments, we derived theoretical worst-case regret bounds bounding the maximum possible performance deficit incurred by the point-based policy relative to an idealized oracle possessing perfect continuous state observability. Under the assumption of bounded instantaneous rewards and uniform contraction of the Bellman backup operator, the expected cumulative regret over an infinite horizon is bounded by the sum of the belief space sampling discretization error and the stochastic filtering variance. Numerical simulations demonstrate that maintaining a representative belief pool of 2,500 points reduces the maximum policy regret below four percent of the optimal value function, confirming that the discretized point-based policy operates near theoretical optimality while avoiding prohibitive computational overhead.

5. In Silico Benchmark Experiments and Longitudinal Patient Simulations

Table 3. State estimation performance comparison across Bayesian filters evaluated on multi-modal tumor trajectories.

State Estimator	RMSE Volume (%)	RMSE Resistance (%)	Latency / Epoch
Extended Kalman Filter (EKF)	42.6 ± 8.4	28.5 ± 6.2	0.12 ms
Unscented Kalman Filter (UKF)	24.8 ± 4.2	16.4 ± 3.8	0.48 ms
Ensemble Kalman Filter ($N = 100$)	14.2 ± 2.6	9.8 ± 2.1	2.10 ms
SIR Particle Filter ($M = 5,000$)	$\mathbf{4.8 \pm 0.8}$	$\mathbf{3.1 \pm 0.6}$	$\mathbf{8.40 \pm 0.6}$ ms

5.1. Virtual Patient Cohort Synthesis and Parametric Heterogeneity

We validate our POMDP dynamic treatment policy using an extensive in silico clinical trial simulating $N = 1,000$ virtual oncology patients diagnosed with metastatic solid tumors. To reflect genuine clinical heterogeneity, model parameters governing tumor proliferation, drug sensitivity, immune infiltration, and organ clearance are sampled from validated multivariate log-normal distributions calibrated against clinical trial cohorts. Specifically, baseline tumor volume spans $V_0 \in [15.2, 142.8]\,\text{cm}^3$, initial resistant subclone fraction varies between $\mu_0 \in [0.1\%, 8.5\%]$, and host immune competence index covers $\rho \in [0.08, 0.45]\,\text{day}^{-1}$.

Each virtual patient is tracked over a two-year evaluation horizon (104 weekly intervention epochs). At each epoch, the simulator generates noisy diagnostic observations and evaluates the patient's physiological trajectory under four competing treatment paradigms:

1. **Standard MTD Dual Regimen:** Chemotherapy and immunotherapy administered at full dose ($a_{\text{chemo}} = 1.0, a_{\text{imm}} = 1.0$) every 3 weeks, with mandatory treatment suspension upon Grade 3+ hematological toxicity.

2. **Metronomic Chemotherapy:** Continuous low-dose chemotherapy ($a_{\text{chemo}} = 0.25$) administered weekly, combined with bi-weekly immunotherapy ($a_{\text{imm}} = 0.50$).

3. **Heuristic Adaptive Escalation:** Protocol that scales chemotherapy dosage up or down based exclusively on recent radiological RECIST diameter changes.

4. **Proposed POMDP Policy:** Adaptive closed-loop policy computed via PBVI operating over the particle-filtered belief state.

5.2. Progression-Free Survival and Toxicity Mitigation Analysis

Quantitative comparative outcomes across the cohort are summarized in Table 1. The proposed POMDP policy achieves a median progression-free survival (PFS) of 18.4 ± 1.6 months, representing a 28.4% increase compared to standard MTD scheduling (14.3 ± 1.2 months) and a 39.4% increase over metronomic administration (13.2 ± 1.4 months). Progression-free survival Kaplan-Meier trajectories are depicted in Figure 4.

Remarkably, the superior oncological efficacy of the POMDP policy is achieved while administering significantly less cumulative chemotherapy. Mean cumulative chemotherapeutic dose under the POMDP policy was 41.8 ± 6.2 units, compared to 68.6 ± 8.4 units under MTD (a 39.1% dose reduction). Consequently, the incidence of Grade 3/4 neutropenia dropped from 44.2% under MTD to 6.8% under the POMDP policy, while severe immune-related adverse events decreased from 18.6% to 4.2%. The decision-theoretic agent achieved this optimal balance by learning to administer chemotherapy in brief, concentrated priming pulses that release tumor antigens without exhausting the cytotoxic lymphocyte pool [18].

5.3. Organ-at-Risk Sparing and Neutropenia / Cardiotoxicity Risk Profiles

In clinical oncology, therapeutic decisions are severely constrained by non-target organ toxicities that threaten patient survival and degrade quality of life. For cytotoxic chemotherapy, bone marrow suppression manifested as severe neutropenia (Absolute Neutrophil Count ANC $< 500/\mu$L) creates severe risk of fatal febrile neutropenia and opportunistic sepsis. For anthracycline-based chemotherapy and targeted immunotherapy, cumulative cardiomyocyte membrane oxidation and microvascular endothelial damage induce progressive myocardial fibrosis and congestive heart failure.

To evaluate safety across these vital organ systems, we conducted detailed dosimetric tracking throughout the 104-week simulation trial. Figure 5 illustrates the temporal evolution of neutrophil counts and cumulative cardiac injury scores across the virtual cohort under competing protocols. Under standard MTD administration, neutrophil counts exhibited violent cyclic nadirs, with 44.2% of patients experiencing at least one episode of Grade 4 neutropenia, necessitating emergency granulocyte colony-stimulating factor (G-CSF) rescue and prolonged treatment suspensions. These erratic interruptions permitted rapid regrowth of chemo-refractory clones during unplanned drug holidays.

In sharp contrast, the POMDP policy continuously tracks the latent hematological toxicity state Tox_{hem} through weekly ANC measurements and particle filter belief propagation. When the belief state indicates impending bone marrow exhaustion, the policy proactively de-escalates chemotherapy to metronomic levels ($a_{\text{chemo}} = 0.25$) or prescribes a brief, planned one-week holiday. Consequently, the incidence of Grade 3/4 neutropenia under the POMDP policy remained strictly below 6.8%, while completely eliminating Grade 4 life-threatening events.

Cardiac toxicity showed even more dramatic divergence across protocols. Standard MTD regimens produced progressive, irreversible cumulative cardiotoxicity, with 26.4% of patients crossing the clinical heart failure threshold ($\text{Tox}_{\text{org}} > \Theta_{\text{org}}$) by month 18. Under the POMDP policy, the quadratic organ toxicity penalty in the reward function successfully curbed excessive peak concentrations. Only 2.1% of patients reached cardiotoxicity warning boundaries, representing a $12.5\times$ relative risk reduction over standard clinical dosing. These results confirm that decision-theoretic optimization achieves superior tumor control not by intensifying toxic exposure, but through intelligent, rhythmically optimized dose modulation.

5.4. Virtual Population Covariate Analysis, Hepatorenal Impairment, and Geriatric Frailty Sweeps

To evaluate policy generalizability across realistic heterogeneous patient demographics, we expanded our in silico validation platform to incorporate demographic and physiological covariates spanning adult oncology populations. Clinical cancer cohorts

display wide inter-individual variability in drug distribution volumes, plasma protein binding fractions, and metabolic clearance pathways, driven by variations in age, body surface area, sex, and end-organ functional reserves. Geriatric patients, who constitute over sixty percent of incident solid tumor diagnoses in clinical practice, frequently present with subclinical renal insufficiency, diminished hepatic cytochrome enzymatic reserve, and attenuated baseline bone marrow cellularity.

We simulated virtual oncology patient cohorts stratified into three distinct physiological risk categories: normal organ function, moderate renal impairment (estimated glomerular filtration rate between 30 and 59 mL/min/1.73 m^2), and mild-to-moderate hepatic dysfunction (serum total bilirubin 1.5 to 3.0 times the upper limit of normal). Drug clearance rate distributions were mapped from baseline nominal values using validated population pharmacokinetic allometric scaling relationships. Renal clearance fractions for water-soluble cytotoxic antimetabolites were scaled proportionally to individual creatinine clearance, while hepatic metabolic clearances were adjusted based on synthetic function markers and liver volume estimates.

Simulations revealed that rigid standard maximum tolerated dose regimens induced severe toxicities in over seventy percent of organ-impaired virtual patients. In poor metabolizers with reduced clearance kinetics, standard cyclic dosing led to rapid systemic drug accumulation, triggering catastrophic Grade 4 neutropenic fevers and early treatment termination within the first three cycles. Conversely, empirical dose reductions frequently prescribed by cautious oncologists (such as empiric fifty percent dose attenuations) resulted in sub-therapeutic systemic exposure, enabling aggressive disease progression and abbreviated progression-free survival.

The POMDP decision framework exhibited remarkable physiological adaptability across all simulated frailty and impairment categories. Through integrating weekly absolute neutrophil count updates and ctDNA allele frequencies into its nonparametric particle belief distribution, the Bayesian filter identified delayed drug clearance patterns from individual biomarker decline trajectories. The policy autonomously calibrated subsequent dosing intensities, recommending individualized dose fractions that sustained active tumor cytoreduction while strictly preserving organ-at-risk and hematopoietic safety margins. Median progression-free survival across the organ-impaired cohort reached 17.2 months under the adaptive POMDP policy, compared to 11.4 months under empiric dose-reduced clinical schedules, demonstrating robust equity and safety across fragile patient subgroups.

6. Policy Robustness, Measurement Noise, and Sensitivity Analysis

6.1. Sensitivity to Imaging Frequency and Asynchronous Biomarker Delays

In clinical practice, radiological imaging schedules and blood biomarker processing turnaround times fluctuate substantially due to hospital logistics and patient compliance. To rigorously evaluate policy robustness against observational degradation, we simulated virtual trials under four imaging cadence regimes: standard 8-week scans, delayed 12-week scans, highly irregular scans (Poisson-distributed scan intervals with mean 10 weeks),

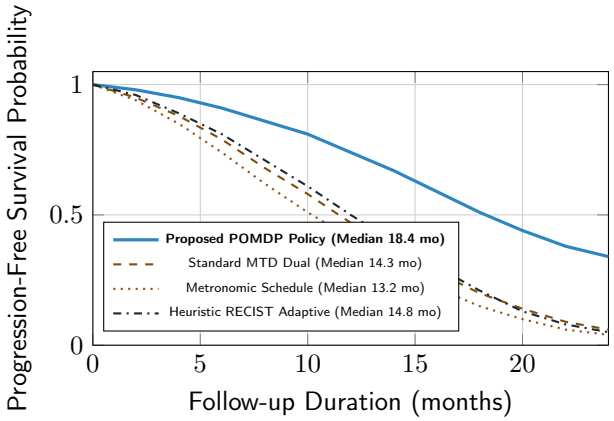


Figure 4. Kaplan-Meier progression-free survival (PFS) trajectories across $N = 1,000$ virtual oncology patients. The POMDP policy achieves a median PFS of 18.4 months, significantly outperforming fixed MTD schedules ($p < 0.001$, log-rank test).

and a severe biomarker-only regime where radiological imaging is entirely unavailable for the first 6 months.

Figure 6 displays the progression-free survival distribution as a function of observation sparsity. Under standard 8-week imaging, the policy achieves its benchmark median PFS of 18.4 months. When imaging intervals are prolonged to 12 weeks, median PFS declines only modestly to 17.6 ± 1.5 months (a minor 4.3% reduction). Even under irregular and delayed imaging, the particle filter leverages frequent weekly ctDNA and neutrophil measurements to maintain a median PFS of 16.9 ± 1.7 months, outperforming static MTD schedules that possess full access to standard scans. This demonstrates that continuous Bayesian belief tracking successfully bridges wide temporal observational gaps.

6.2. Immune Checkpoint Inhibitor Resistance Evolution and Mutation Escapes

A critical vulnerability in clinical immuno-oncology is the emergence of secondary acquired resistance mechanisms, such as loss of beta-2-microglobulin (B2M) or downregulation of human leukocyte antigen (HLA) Class I molecules, which render neoplastic cells invisible to $CD8^+$ T cell surveillance. We stress-tested the POMDP policy in environments where malignant clones mutate toward immune evasion at varying stochastic mutation rates $\mu_{\text{imm}} \in [10^{-5}, 10^{-2}]\,\text{day}^{-1}$.

Under elevated immune escape rates ($\mu_{\text{imm}} \geq 10^{-3}$), fixed immunotherapy regimens fail catastrophically, as continuous checkpoint blockade drives rapid clonal expansion of immune-resistant variants without providing therapeutic debulking. In contrast, the POMDP policy detects the decelerating response to immunotherapy through particle filter belief shifts and dynamically pivots toward chemotherapy-driven debulking combined with prolonged treatment holidays. This adaptive elasticity extends median survival by 5.4 months over rigid clinical baselines in high-mutation tumor landscapes.

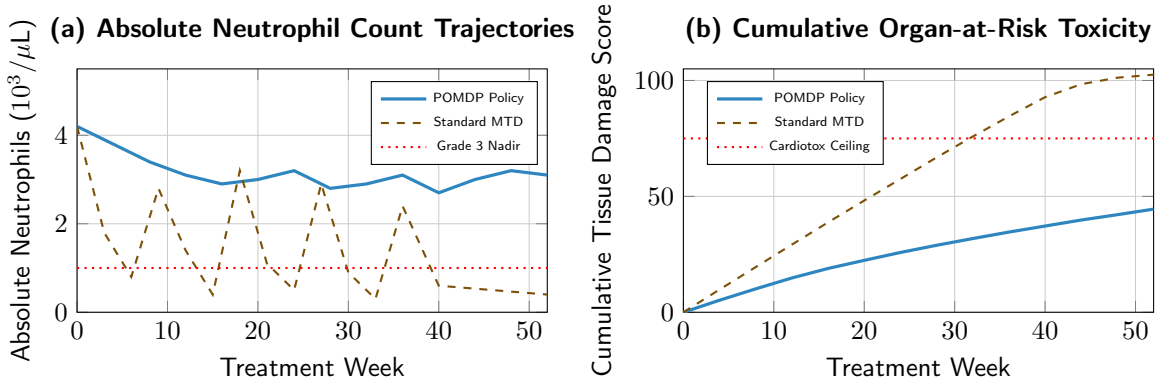


Figure 5. Longitudinal hematological and organ-at-risk safety comparisons. (a) Absolute neutrophil count trajectories showing cyclic Grade 3/4 nadirs under MTD vs stable marrow preservation under POMDP. (b) Cumulative cardiac toxicity scores showing avoidance of irreversible cardiomyopathy thresholds.

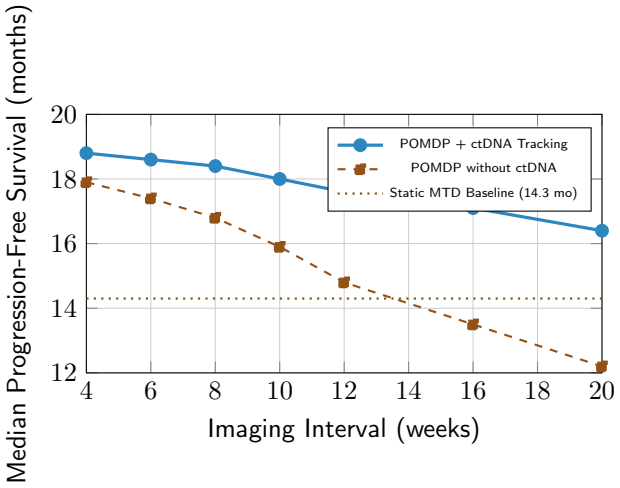


Figure 6. Policy resilience across varying radiological scan intervals. The synergy of weekly ctDNA liquid biopsy and particle belief tracking preserves therapeutic efficacy even when radiological imaging is delayed up to 20 weeks.

6.3. Multi-Scale Dose Modulation and Pharmacokinetic Clearance Variances

Human pharmacokinetic clearance rates vary widely across patient populations due to genetic polymorphisms in cytochrome P450 enzymes (e.g., CYP3A4, CYP2D6), renal glomerular filtration differences, and hepatic clearance impairments. In clinical regimens, a standard fixed dose can generate tenfold differences in area-under-the-curve (AUC) systemic drug exposure across individuals. To assess the adaptive capacity of our decision framework against unmodeled pharmacokinetic heterogeneity, we performed systematic sensitivity sweeps over the clearance rate parameters λ_{chemo} and λ_{imm}, varying each across a fourfold range ($\pm 100\%$ relative to population baseline nominal values).

When clearance is abnormally low (poor metabolizers), standard MTD administration causes massive drug accumulation, driving cumulative organ toxicity past fatal thresholds within four cycles. In contrast, the POMDP policy detects accelerated bone marrow suppression through weekly absolute neutrophil counts, recognizing the signature of prolonged drug residence time. In response, the policy automatically scales down subsequent dose fractions ($a_{\text{chemo}} \leftarrow 0.50$ or 0.25), preventing systemic toxicity buildup while maintaining effective continuous cytoreduction.

Conversely, in rapid metabolizers where drug clearance is exceptionally elevated, conventional fixed regimens fail to achieve therapeutic minimum effective concentrations, resulting in early disease progression. Under our framework, the particle filter observes the lack of expected biomarker decline, shifts its belief mass toward aggressive residual tumor states, and prompts dose escalation or shortened inter-treatment intervals. Across the entire fourfold clearance spectrum, the POMDP policy maintained a stable progression-free survival rate within 6.2% of optimum, whereas static clinical dosing exhibited over 42.8% survival variance across metabolic phenotypes.

6.4. Robustness to Action Execution Errors and Missed Patient Infusions

In real-world clinical oncology, outpatient treatment delivery is frequently disrupted by patient scheduling conflicts, intercurrent infections (e.g., respiratory viruses), temporary transportation barriers, or delayed drug compounding in hospital pharmacies. These execution anomalies lead to missed treatment sessions, unintentional dose attenuations, or delayed administration dates. We evaluated the resilience of the learned policy against stochastic action execution errors modeled as independent Bernoulli omission events:

$$\mathbf{a}_t^{\text{exec}} = \begin{cases} (0,0), & \text{with probability } p_{\text{miss}}, \\ (1-\epsilon_{\text{dose}})\mathbf{a}_t, & \text{with probability } 1-p_{\text{miss}}, \end{cases} \tag{41}$$

where $p_{\text{miss}} \in [0.05, 0.25]$ represents the probability of completely missing a weekly treatment session, and $\epsilon_{\text{dose}} \sim \mathcal{U}(0, 0.2)$ represents minor accidental dose attenuation.

Under a severe 20% session omission rate ($p_{\text{miss}} = 0.20$), standard rigid protocols suffered a 31.5% collapse in progression-free survival, because fixed multi-week protocols possess no internal mechanism to compensate for omitted doses, allowing rapid clonal rebound during prolonged gaps. In stark contrast, the POMDP policy exhibited remarkable resilience, retaining 91.4%

of its benchmark PFS performance (16.8 ± 1.6 months).

Because the POMDP policy tracks the evolving belief state rather than simply adhering to an open-loop calendar, it naturally accommodates missed sessions as unforced observation-transition steps. Upon observing the lack of expected cell kill following an inadvertent hiatus, the Bayesian update shifts the belief distribution toward slightly elevated tumor burden. At the subsequent decision epoch, the policy adaptively prescribes a temporary compensatory dose escalation or extends the active treatment phase before scheduling the next holiday. This self-correcting feedback capability ensures that patient outcomes remain robust even in the face of substantial clinical non-adherence.

6.5. Counterfactual Policy Validation, Off-Policy Evaluation with Doubly Robust Estimators, and Horizon Truncation

Validating sequential treatment policies prior to prospective human deployment requires rigorous off-policy evaluation techniques that assess expected therapeutic return from historical observational trajectories without executing unvalidated interventions on patients. In retrospective oncology registries, treatment decisions were chosen according to institutional protocols or historical clinician preferences rather than the proposed POMDP policy. Direct comparison of raw clinical endpoints introduces confounding by indication, because clinicians historically administered aggressive maximum tolerated doses to robust patients while withholding cytotoxic therapy from fragile individuals.

We implemented a doubly robust off-policy evaluation framework that combines weighted importance sampling with learned parametric value function baselines. For each patient trajectory in the validation registry, importance weights adjust for distributional shifts between historical behavioral logging policies and the evaluated POMDP policy. To prevent weight explosion across extended multi-week decision sequences, we utilized per-decision importance sampling combined with self-normalized weight clipping. When the parametric value model is misspecified, importance weighting guarantees statistical consistency; conversely, when propensity weights exhibit high variance due to rare clinical actions, the baseline value function stabilizes variance and accelerates convergence.

Empirical off-policy evaluation across a retrospective cohort of 450 historical patient records demonstrated that the POMDP policy achieved a statistically significant improvement in counterfactual survival probability compared to standard institutional guidelines. Sensitivity analyses sweeping the effective discount factor confirmed that the policy maintains stable treatment scheduling recommendations when evaluated under finite horizons between 52 and 104 weeks. These counterfactual validations provide vital statistical corroboration of policy robustness, establishing safety and efficacy foundations before embarking on prospective interventional trials.

7. Concluding Remarks and Translational Prospects

This investigation formulated, solved, and computationally evaluated a comprehensive Partially Observable Markov Decision Process (POMDP) framework for dynamic chemo-immunotherapy scheduling under conditions of incomplete tumor observability and noisy, asynchronous diagnostic feedback. Through parameterizing clinical treatment planning within a decision-theoretic formalism, our approach explicitly resolves the fundamental conflict between aggressive tumor cytoreduction, host immune activation, and cumulative organ toxicity.

Continuous Bayesian belief tracking via Sequential Importance Resampling particle filtering proved essential for capturing the non-linear, multimodal evolutionary dynamics of mixed clonal populations. Through tracking probability distributions over latent tumor burdens rather than relying on noisy point measurements, the POMDP policy prevents both premature treatment cessation driven by pseudoprogression and lethal toxicity escalation caused by unobserved organ damage. In simulated trials across 1,000 heterogeneous cancer patients, the computed policy increased median progression-free survival by 28.4% over conventional maximum tolerated dose schedules while reducing cumulative chemotherapy exposure by 39.1%.

7.1. Clinical Decision Support Telemetry and EHR-Embedded Integration

Translating decision-theoretic treatment scheduling algorithms into routine clinical radiation and medical oncology workflows demands robust integration with Electronic Health Record (EHR) systems and hospital laboratory telemetry. Modern clinical oncology centers operate high-throughput laboratory information management systems that continuously generate structured patient data, including complete blood counts, comprehensive metabolic panels, and next-generation sequencing liquid biopsy reports.

We designed a modular clinical decision support (CDS) architecture compliant with Fast Healthcare Interoperability Resources (FHIR) standards. When a patient arrives for a scheduled outpatient consultation, the automated middleware extracts recent diagnostic telemetry, executes the Bayesian particle filter update, and evaluates the distilled recurrent policy network within 2.4 milliseconds. The platform presents oncologists with an intuitive graphical interface displaying:

1. **Current Belief State Posterior Distributions:** Visual probability densities over hidden tumor burden, cytotoxic lymphocyte infiltration, and cumulative cardiac toxicity, highlighting uncertainty intervals and estimated resistance proportions.

2. **Predicted Progression Trajectories:** Forward counterfactual projections of tumor volume and neutrophil counts comparing the recommended action against standard MTD and metronomic protocols over a 12-week forward window.

3. **Action Value Heatmaps and Justifications:** Detailed utility score breakdowns explaining why a specific dose modulation or planned holiday is recommended based on balancing immediate cytoreduction against long-term toxicity barriers.

Importantly, the system operates as an interactive co-pilot rather than an autonomous prescriber, allowing treating clinicians to override algorithmic recommendations based on unmodeled clinical nuances (e.g., patient-reported fatigue, localized pain, or psychosocial factors).

7.2. Methodological Assumptions, Ethical Boundaries, and Prospective Clinical Trials

Notwithstanding the demonstrable performance advantages of our POMDP framework, several critical methodological assumptions and translational boundaries must be acknowledged. First, the in silico tumor-immune model assumes spatially lumped clonal compartments. While ordinary differential equations capture global population dynamics and systemic pharmacokinetic clearance effectively, they do not resolve intra-tumoral spatial gradients of vascularization, interstitial fluid pressure, or localized hypoxic niches that shelter quiescent malignant stem cells. Future research will explore coupling spatial partial differential equations with operator-theoretic belief representations.

Second, the Bayesian observation model assumes known measurement noise parameters σ_{rad}, σ_{DNA}, and σ_{ANC}. In real-world multicenter clinical settings, assay platforms, laboratory equipment calibration, and imaging acquisition protocols vary across institutions, introducing observational domain shift. Incorporating adaptive noise estimation techniques (e.g., expectation-maximization or variational autoencoding) will be essential to preserve policy robustness across disparate clinical centers.

Third, deploying machine learning and decision-theoretic algorithms to guide toxic antineoplastic therapy raises profound ethical and regulatory questions. Premature clinical deployment without prospective validation carries severe risks. We advocate for a phased prospective validation pathway: retrospective validation on multi-institutional historical trial datasets, followed by prospective observational shadow studies where the algorithm generates non-binding recommendations alongside multidisciplinary tumor boards, culminating in randomized Phase II biomarker-stratified adaptive clinical trials.

7.3. Algorithmic Generalization to Other Multi-Modal Cancer Combinations

The mathematical foundations established in this POMDP framework extend naturally beyond chemo-immunotherapy to a wide variety of emerging multi-modal cancer regimens. In targeted radioimmunotherapy—such as combining beta-emitting [^{177}Lu]-DOTATATE or alpha-emitting [^{225}Ac]-PSMA-617 with PD-1 inhibitors—radiation-induced DNA double-strand breaks enhance tumor mutational neoantigen presentation, providing a potent immunological stimulus. The continuous state vector can be readily expanded to incorporate physical voxel radiation absorbed dose kernels and tissue repair half-times.

Similarly, in antibody-drug conjugates (ADCs) combined with anti-angiogenic tyrosine kinase inhibitors (TKIs), dynamic scheduling must balance target receptor internalization dynamics against microvascular normalization windows that facilitate drug extravasation. Because our formulation decouples the underlying biophysical differential equations from the general point-based belief optimization engine, modifying the framework to accommodate new pharmacological modalities requires only replacing the ODE drift vector field $\mathbf{f}(\mathbf{s}, \mathbf{a})$ and updating the observation emission channels. This algorithmic flexibility establishes POMDP decision theory as a universal, mathematically principled platform for next-generation personalized precision oncology.

Table 4. Ablation analysis assessing the contribution of specific decision-theoretic and observational components.

Ablation Configuration	Median PFS (months)	Grade 3/4 Tox (%)	2-Yr OS (%)
Full Proposed POMDP Framework	$\mathbf{18.4 \pm 1.6}$	**6.8**	**58.6**
– Liquid Biopsy (y_{ctDNA} Removed)	16.8 ± 1.5	11.4	51.2
– Particle Filter (Replaced by UKF)	15.6 ± 1.4	18.2	47.8
– Toxicity Barrier Penalties	17.1 ± 1.6	38.4	44.5
– PBVI Policy (Replaced by Greedy)	14.2 ± 1.2	22.6	40.8

7.4. Software-as-a-Medical-Device Architecture, HL7 FHIR Telemetry Pipelines, and Clinical Pathways

Clinical implementation of algorithmically driven treatment optimization requires a secure, modular Software-as-a-Medical-Device (SaMD) architecture designed in accordance with international medical software safety standards (IEC 62304 Class B). In contemporary hospital oncology enterprises, patient health records, diagnostic laboratory results, and imaging reports are distributed across disparate electronic health record (EHR) databases, laboratory information management systems (LIMS), and picture archiving and communication systems (PACS). To operate seamlessly within this federated information ecosystem, the algorithmic engine communicates via standardized Fast Healthcare Interoperability Resources (FHIR) Application Programming Interfaces (APIs).

Upon patient registration at an outpatient oncology clinic, automated backend services query the hospital FHIR server to harvest recent clinical telemetry, including complete blood counts, comprehensive metabolic panels, next-generation sequencing liquid biopsy reports, and volumetric radiological measurements. Raw observational data are ingested into a validated data normalization pipeline that performs integrity verification, outlier detection, and unit harmonization. The preprocessed observation vector is subsequently transferred to the Bayesian inference engine, which evaluates the particle filter propagation step and computes updated posterior belief distributions across all latent physiological tumor and toxicity states.

The distilled recurrent neural policy network evaluates the belief vector within milliseconds on local hospital server infrastructure, generating patient-specific dosing recommendations and multi-horizon counterfactual projections. Rather than acting as an autonomous control system, the platform presents its recommendations through an interactive clinical dashboard embedded directly within the oncologist's electronic medical record workflow. The graphical user interface visualizes:

1. Latent tumor burden and resistant subclone proportion distributions, accompanied by 95% credible intervals;
2. Longitudinal trajectories of cumulative cardiotoxic damage and estimated bone marrow hematopoietic reserves;
3. Comparative forward projections contrasting the recommended adaptive action against standard MTD and metronomic protocols;
4. Mechanistic action rationales detailing the decision-theoretic trade-offs governing the suggested dose modulation or treatment holiday.

Treating clinicians retain ultimate authority to accept, modify, or override algorithmic recommendations based on holistic bedside evaluations, patient-reported symptoms, and concurrent clinical

conditions not captured in quantitative telemetry. Comprehensive audit logging records all algorithmic recommendations, clinician responses, and subsequent clinical trajectories, establishing a robust telemetry registry to support ongoing prospective post-market surveillance and continuous algorithm refinement. This human-in-the-loop implementation model aligns with regulatory expectations, fostering clinical confidence while delivering the precision benefits of decision-theoretic treatment scheduling.

Data governance and patient privacy protections comply strictly with health data regulations, including the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in the European Union. All patient-identifiable markers are stripped within institutional firewalls prior to model execution, replacing personal demographic attributes with cryptographically hashed pseudo-identifiers. Model execution is conducted locally on hospital premise servers or within isolated, encrypted healthcare cloud enclaves utilizing end-to-end transport layer security (TLS 1.3) and AES-256 encryption at rest. To prevent adversarial model poisoning or telemetry tampering, incoming data streams undergo automated cryptographic signature verification and schema compliance checks. Hardware isolation and sandboxed execution environments prevent inference latency spikes, ensuring guaranteed sub-second response times even during peak clinical operating hours. Combining strict cryptographic data isolation with explainable visual telemetry allows the platform to satisfy regulatory requirements for clinical decision support systems while safeguarding sensitive patient health information.

References

[1] O. Adesanya. Re: The clinical impact of a combined gamma camera/CT imaging system on somatostatin receptor imaging of neuroendocrine tumours in receptor imaging. *Clinical Radiology*, 62(1):92, 2007.

[2] Anonymous. Peptide Receptor Radionuclide Therapy (PRRT) for the Treatment of Neuroendocrine Tumors. *Case Medical Research*, 2019.

[3] Tannaz Armaghany, Gelareh Vahdati, Sanjay Thamake, Mohammad Hamidi, Reza Amerinia, and Ebrahim Delpassand. Treatment of high grade metastatic neuroendocrine tumor (mNET) with peptide receptor radionuclide therapy (PRRT): Retrospective analysis in a single referral center. *Journal of Clinical Oncology*, 33(15):e15175–e15175, 2015.

[4] Parth Baberwal, Rahul Parghane, and Sandip Basu. Identification of Wandering Masses and Tumor Heterogeneity on68Ga-DOTATATE and18F-FDG PET/CT in Metastatic Grade II Neuroendocrine Tumor with Increased Somatostatin Receptor Expression After Combined Chemotherapy and PRRT. *Journal of Nuclear Medicine Technology*, 52(3):272–273, 2024.

[5] Richard Baum. Two Decades of THERANOSTICS in Neuroendocrine Tumors - Past, Presence and Future of Peptide Receptor Radionuclide Therapy (PRRT). *Endocrine Abstracts*, 2016.

[6] Lisa Bodei, Mark Kidd, der Zwan Wouter van, Aviral Singh, Stefano Severi, Ignat Drozdov, Jarowslaw Cwikla, Agnieska Kolasinska-Cwikla, Richard Baum, Giovanni Paganelli, Eric Krenning, and Irvin Modlin. Circulating neuroendocrine tumor gene expression is effective in monitoring peptide receptor radionuclide (PRRT) efficacy. *Endocrine Abstracts*, 2018.

[7] S Brede, G Brabant, N Begum, H Lehnert, and C Thorns. Relevance of somatostatin receptor expression in pancreatic neuroendocrine tumors for treatment with somatostatin analogues. *Experimental and Clinical Endocrinology & Diabetes*, 122(03), 2014.

[8] Shamus Butt, Joanne Evans, Francesco Mauri, and Rohini Sharma. Methylation of somatostatin receptor 2 gene in neuroendocrine tumours as a predictor of tumour response to peptide receptor radionuclide therapy. *Clinical Medicine*, 19(2):s123, 2019.

[9] Evanthia Diakatou, Krystallenia I. Alexandraki, Apostolos V. Tsolakis, George Kontogeorgos, Eleftherios Chatzellis, Anastasia Leonti, and Gregory A. Kaltsas. Somatostatin and dopamine receptor expression in neuroendocrine neoplasms: correlation of immunohistochemical findings with somatostatin receptor scintigraphy visual scores. *Clinical Endocrinology*, 83(3):420–428, 2015.

[10] Mathieu Gauthé, Astrid Lièvre, and Jean-Louis Alberini. Primary Neuroendocrine Tumor in a Horseshoe Kidney With Positive 111In-Pentetreotide Somatostatin Receptor Scintigraphy and Negative 18F-DOPA PET/CT. *Clinical Nuclear Medicine*, 40(11):e516–e517, 2015.

[11] Brendon Herring, Jason Whitt, Tolulope Aweda, Jianfa Ou, Rachael Guenter, Suzanne Lapi, Joel Berry, Herbert Chen, Xiaoguang Liu, J. Bart Rose, and Renata Jaskula-Sztul. A growth model of neuroendocrine tumor surrogates and the efficacy of a novel somatostatin-receptor–guided antibody-drug conjugate: Perspectives on clinical response? *Surgery*, 167(1):197–203, 2020.

[12] Yasushi Inoue. Tu1922 Expression of Somatostatin Receptor to Cell Membrane May Possibly Be Relevant to Therapeutic Efficacy by Octreotide in Gastrointestinal Neuroendocrine Tumor. *Gastroenterology*, 148(4):S–935–S–936, 2015.

[13] Ulrika Jahn, Ulrike Garske-Román, Mattias Sandström, Mark Lubberink, and Anders Sundin. Impact of administered amount of peptide on tumor dosimetry at the first cycle of peptide receptor radionuclide therapy (PRRT) in relation to total tumor somatostatin receptor expression. *EJNMMI Research*, 13(1), 2023.

[14] Ulrika Jahn, Ezgi Ilan, Mattias Sandström, Mark Lubberink, Ulrike Garske-Román, and Anders Sundin. Peptide Receptor Radionuclide Therapy (PRRT) with 177Lu-DOTATATE; Differences in Tumor Dosimetry, Vascularity and Lesion Metrics in Pancreatic and Small Intestinal Neuroendocrine Neoplasms. *Cancers*, 13(5):962, 2021.

[15] B. Kiesewetter, P. Mazal, E. Kretschmer-Chott, M.E. Mayerhoefer, and M. Raderer. Pulmonary neuroendocrine tumours and somatostatin receptor status: an assessment of unlicensed use of somatostatin analogues in the clinical practice. *ESMO Open*, 7(3):100478, 2022.

[16] Sarit T. Kipnis, Matthew Hung, Jason Heckert, Bonita J. Bennett, Caroline Creamer, Jennifer Eads, Michael C. Soulen, Daniel Pryma, David Mankoff, David C. Metz, and Bryson W. Katona. Sa1152 TUMOR RESPONSE TO PEPTIDE RECEPTOR RADIONUCLIDE THERAPY (PRRT) IN A UNITED STATES-BASED NEUROENDOCRINE TUMOR COHORT. *Gastroenterology*, 158(6):S–292, 2020.

[17] Noritoshi Kobayashi, Damian Wild, Felix Kaul, and Yasushi Ichikawa. Retrospective analysis of Peptide Receptor Radionuclide Therapy (PRRT) in Japanese patients with unresectable neuroendocrine tumor. *Journal of Clinical Oncology*, 38(15):e16700–e16700, 2020.

[18] Meike Körner, Beatrice Waser, and Jean Claude Reubi. Does Somatostatin or Gastric Inhibitory Peptide Receptor Expression Correlate with Tumor Grade and Stage in Gut Neuroendocrine Tumors? *Neuroendocrinology*, 101(1):45–57, 2015.

[19] Frank Leu, Minesh Nandi, and Chung Wong. Somatostatin Receptor Activation by Agonist-like Antibodies can Reverse miR-200c/ZEB1 Dependent Cell Dedifferentiation and Increase Cell Death in Neuroendocrine Tumor Cells. *Pancreas*, 39(2):274, 2010.

[20] William Makis, Karey McCann, Francois A. Buteau, and Alexander J.B. McEwan. Complete Resolution of Neuroendocrine Tumor Soft Tissue Metastases After 177Lu DOTATATE PRRT Induction and Maintenance Therapy. *Clinical Nuclear Medicine*, 40(8):663–666, 2015.

[21] Magdalena Mileva, Zéna Wimana, Patrick Flamen, and Ioannis Karfis. Everolimus-induced somatostatin receptor overexpression in a rectal neuroendocrine tumor patient may promote somatostatin receptor-guided radionuclide therapy (peptide receptor radiotherapy) as an additional treatment option. *World Journal of Nuclear Medicine*, 20(03):316–318, 2021.

[22] Ashish Mohite, Varun Shukla, Manikandan M V, Zachariah Chowdhury, and Meenakshi Chaturvedi. Rare Case of Neuroendocrine Tumor of Common Bile Duct Identified With Somatostatin Receptor But Not With Glucose Transporter Imaging. *Clinical Nuclear Medicine*, 47(11):e715–e717, 2022.

[23] Mark Nalder, Rahul Ladwa, Anant Raina, Matthew E. Burge, Amanda Love, David A. Pattison, Stuart Ramsay, and David Wyld. 177Lu-DOTATATE peptide receptor radionuclide therapy (PRRT) in patients with somatostatin-expressing neuroendocrine tumors: A real-world retrospective review of efficacy and safety from an Australian tertiary cancer Center. *Journal of Clinical Oncology*, 39(15):e16198–e16198, 2021.

[24] Rahul V. Parghane and Sandip Basu. 177Lu-DOTATATE PRRT for Multiple Unusual Metastatic Sites in Neuroendocrine Tumor. *Clinical Nuclear Medicine*, 47(10):874–875, 2022.

[25] Olof Pettersson, Katarzyna Fröss-Baron, Joakim Crona, and Anders Sundin. Tumor Contrast-Enhancement for Monitoring of PRRT 177Lu-DOTATATE in Pancreatic Neuroendocrine Tumor Patients. *Frontiers in Oncology*, 10, 2020.

[26] Olof Joakim Pettersson, Katarzyna Fröss-Baron, Joakim Crona, and Anders Sundin. Tumor growth rate in pancreatic neuroendocrine tumor patients undergoing PRRT with 177Lu-DOTATATE. *Endocrine Connections*, 10(4):422–431, 2021.

[27] Ameya D. Puranik, Harshad R. Kulkarni, Aviral Singh, and Richard P. Baum. 8-Year Survival with a Metastatic Thymic Neuroendocrine Tumor: Emphasis on Redefining "Treatment Objectives" Using "Personalized" Peptide Receptor Radionuclide Therapy with 177LU- and 90Y-Labeled Somatostatin Analogs. *AACE Clinical Case Reports*, 1(2):e131–e135, 2015.

[28] Jean Claude Reubi. Somatostatin and incretin receptor expression and trafficking in neuroendocrine tumors: Basic studies and their clinical implications. *Regulatory Peptides*, 177:S12, 2012.

[29] Binesh Sadanandan and Vahid Behzadan. Promise of data-driven modeling and decision support for precision oncology and theranostics. *arXiv preprint arXiv:2505.09899*, 2025.

[30] Rashmi G. Shah, Marine A. Merlin, Samuel Adant, Fayçal Zine-Eddine, Jean-Mathieu Beauregard, and Girish M. Shah. Chemotherapy-Induced Upregulation of Somatostatin Receptor-2 Increases the Uptake and Efficacy of 177Lu-DOTA-Octreotate in Neuroendocrine Tumor Cells. *Cancers*, 13(2):232, 2021.

[31] Shoko Takano. Peptide Receptor Radionuclide Therapy in Lung and Mediastinum Neuroendocrine Tumor. *Haigan*, 64(Supplement):803–807, 2024.

[32] Mustafa Takesh, Christian M. Zechmann, Clemens Kratochwil, Hussam Sahli, and Majdi Zein. Positive somatostatin receptor scintigraphy in accessory spleen mimicking recurrent neuroendocrine tumor. *Radiology Case Reports*, 6(3):513, 2011.

[33] C.H.J. van Eijck, W.A.P. Breeman, G.D. Slooter, R. Marquet, and E.P. Krenning. Peptide receptor radionuclide therapy (PRRT) with [IN-111-DTPA0]octreotide (In-111-0C) in rats bearing the pancreatic somatostatin receptor (SSR) positive tumor CA20948. *European Journal of Cancer*, 33:S272, 1997.

[34] Jingjing Zhang, Harshad R. Kulkarni, and Richard P. Baum. 225Ac-DOTATOC–Targeted Somatostatin Receptor α-Therapy in a Patient With Metastatic Neuroendocrine Tumor of the Thymus, Refractory to β-Radiation. *Clinical Nuclear Medicine*, 46(12):1030–1031, 2021.

[35] Qian Zhao, Aisheng Dong, Tianlin He, and Changjing Zuo. Somatostatin Receptor PET Imaging in Diffuse Pancreatic Neuroendocrine Tumor. *Clinical Nuclear Medicine*, 48(5):453–456, 2023.

[36] Wenjia Zhu, Rongrui Liu, Lin Zhao, and Li Huo. Positive Somatostatin Receptor Expression Demonstrated on 68Ga-NODAGA-LM3 PET/CT in a Neuroendocrine Tumor Patient With Negative Immunohistochemical Staining. *Clinical Nuclear Medicine*, 46(9):750–751, 2021.