Bounding Counterfactual Outcomes of Health Insurance Delay-and-Deny Practices

Martin B. Haugh
Imperial College Business School, Imperial College, m.haugh@imperial.ac.uk
Raghav Singal
Tuck School of Business, Dartmouth College, singal@dartmouth.edu

Problem definition. Health insurance delay-and-deny practices, such as requiring prior authorization, frequently restrict access to timely and necessary medical care. These policies can have severe consequences, including delayed diagnosis and treatment, leading to poor patient outcomes or even death. Counterfactual analysis offers a way to assess the impact of these practices by bounding the likelihood that a patient would have survived under alternative policies given the severe consequence under the original policy (*probability of necessity*). However, quantifying counterfactual probabilities is challenging in dynamic systems, such as those arising in healthcare, where disease progression is often latent and evolves over time.

Methodology and results. We develop a principled framework to bound counterfactual probabilities, specifically the probability of necessity (PN), in generalized hidden Markov models (GHMMs). Our approach leverages the structure of GHMMs to construct a feasible space of structural causal models and employs polynomial optimization to compute tight bounds on PN. We integrate domain-specific knowledge to further refine these bounds. A data-driven case study on breast cancer screening and treatment demonstrates the power of our framework. We show that incorporating domain-specific knowledge can reduce the width of PN bounds significantly; thus, providing actionable insights. Our computational techniques are scalable, enabling high-quality solutions for time horizons as large as 100 periods within a few hours.

Managerial implications. Our framework offers a rigorous tool for evaluating the impact of healthcare policies, such as requiring prior authorization or incorrectly denying a medical screening, on patient outcomes. The ability to estimate counterfactual probabilities has applications in policy and legal analysis, and healthcare management. For instance, the bounds on PN can inform court cases by quantifying the impact of an (incorrect) delay-and-deny insurance policy on a particular patient who experienced a detrimental outcome. Beyond healthcare, our methodology is applicable to dynamic latent-state systems in other domains.

Key words. Causal inference, optimization, hidden Markov models, public policy, healthcare
History: This version: January 2026. Previous versions: January 2025 and September 2025.

1. Introduction

In May 2023, *The New York Times* (Rabin 2023) reported the U.S. Preventive Services Task Force changed the age at which it says women should start getting regular mammograms to detect breast cancer. Since 2009, the task force had recommended periodic breast cancer screenings for women 50 and over. The new recommendation lowered the age to 40. In one of the reader comments, a woman asked an interesting question. She explained that she had felt a lump in her breast at age 38 and decided to have a mammogram. The doctor's office had to call her insurance company to

pre-approve (*prior authorization*) the screening, which it eventually did. A few months later, she was diagnosed with stage-1 breast cancer. "Imagine what would have happened if ... my insurance would have denied covering the mammogram?", she asked. Presumably, she wouldn't have had the mammogram, and her cancer would have been allowed to progress to a more dangerous state. The reader's question is a *counterfactual* question as it imagines what might have happened in a counterfactual world where screening was denied given that in the actual world screening was covered, cancer was discovered, and the reader (presumably) survived.

In fact, prior authorization requests are frequently delayed/denied in the U.S. (Chaffin and Wernau 2024). While the overall denial rate across the health insurance industry is not publicly disclosed, data is available for Medicare Advantage plans which serve over 33 million beneficiaries (Freed et al. 2024). An analysis by KFF, a large healthcare nonprofit in the U.S., reported a prior authorization denial rate of 7.4% for 46.2 million requests submitted through Medicare Advantage plans in 2022, a notable increase from the 5.6%-5.8% range observed in 2019-2021 (Biniek et al. 2024). While such denials aim to limit unnecessary treatments and control costs, they are often "incorrect" for critical services such as advanced imaging (e.g., MRIs) and post-acute care (e.g., inpatient rehabilitation). In an audit conducted by the Office of Inspector General on the 2019 data, 13% of prior authorization denials from Medicare Advantage plans met Medicare coverage rules and therefore should have been authorized (Grimm 2022). These delays and denials, especially when inappropriate, can result in severe patient outcomes. A recent *New York Times* survey revealed that one in three doctors reported such delays or denials leading to serious medical complications or even deaths, with many doctors agreeing that prior authorization has been "weaponized" by insurance companies to boost profits (Stockton 2024). In fact, in 2023, *ProPublica* reported on this issue, referring to prior authorization as "denials for dollars business" (Miller et al. 2023). This issue has garnered widespread attention, reigniting debate following the December 2024 killing of UnitedHealthcare CEO Brian Thompson. A recent *Wall Street Journal* investigation interviewing doctors provides a stark example (Mathews 2024): "I had a patient. She had a history of breast cancer ... The appropriate scan for her was ... CT. I ordered the test. It was denied by her health insurance company ... the patient later died. This is ... not an uncommon example."

Such delays and denials can not only lead to severe patient outcomes but also to lawsuits. For instance, in 2022, multiple physicians and thirty patients filed a class action lawsuit against the Hawaii Medical Service Association (HMSA) (Nitta vs HMSA 2024). The plaintiffs allege that HMSA's prior authorization requirements led to the denial of essential medical services, such as CT and MRI scans, adversely affecting patient care. One of the plaintiffs, Scott Norton, had his physician recommend an MRI in 2022 to diagnose his condition. HMSA initially refused to cover the scan, approving only physical therapy. By the time the MRI was authorized – almost three

months later – Norton was diagnosed with advanced prostate cancer that had metastasized to his back, spine, hip, and ribs. He passed away in November 2023. In an interview, Norton's widow, Charlotte Husen, said, "I do think that if he had gotten what his doctor wanted him to get early on it would have been a different outcome altogether" (Huff 2024). In February 2024, Circuit Judge Robert Kim ruled in favor of the plaintiffs. HMSA has appealed this decision, and in October 2024, the Hawaii Supreme Court agreed to hear the case, underscoring its potential to impact healthcare delivery across the state. Norton is just one of the thirty patients involved in this lawsuit. Other notable cases include the landmark Fox vs. Health Net trial, which resulted in an $89 million award (Eckholm 1993), and the ongoing Lokken vs. UnitedHealth lawsuit (Mole 2023).

These legal cases highlight systemic issues where health insurance delay-and-deny practices hinder access to timely and essential medical care. The outcomes of such lawsuits could have far-reaching implications for healthcare policies and practices. Decisions in these cases often rely on multiple factors, including the *probability of necessity* (PN). In this context, PN quantifies the likelihood that a patient would not have experienced a severe outcome (e.g., disease progression or death) had the medical screening not been denied, given that the severe outcome was observed after the denial. This counterfactual measure underpins earlier statements like "imagine what would have happened if" and "if he had gotten what his doctor wanted him to get early on". By quantifying the extent to which an adverse outcome (e.g., a patient's death) can be attributed to a specific event (e.g., a denied screening), understanding PN offers valuable evidence for lawsuits and policy debates. It underscores the critical importance of evaluating the impact of delay-and-deny practices on patient outcomes – a concern that extends beyond the U.S., as demonstrated by the recent implementation of the so-called Martha's Rule in the U.K. (NHS 2024).

In this work, we propose a principled framework to bound PN in the context of health insurance delay-and-deny practices. Our approach consists of two components: (1) modeling the progression of disease at the patient level, and (2) employing this model to algorithmically bound PN.

For the first component, we leverage generalized hidden Markov models (GHMMs), which are well-established in both academia and practice for modeling such processes (see detailed review in §2). As elaborated in §3, GHMMs are particularly suited for capturing the dynamics of disease progression because they allow for the patient state – such as the stage of cancer – to be both (a) hidden/latent and (b) dynamic. This reflects real-world scenarios where the true state of a disease is often unknown and inferred only through noisy diagnostic signals. For example, in a cancer setting, the actual stage of cancer (or the absence thereof) is typically hidden prediagnosis, with screenings providing incomplete or noisy observations. Moreover, the disease progresses dynamically over time, influenced by factors such as whether the patient is receiving treatment.

While a probabilistic model such as a GHMM is useful for disease modeling, it cannot uniquely identify counterfactual quantities like PN. As demonstrated in Example 1 below, the key challenge lies in the fact that infinitely many structural causal models (SCMs) can be consistent with a given probabilistic model, and each SCM can result in a different estimate of PN. To address this, we draw from the partial identification literature (discussed further in §2) and construct a feasible space of SCMs consistent with the underlying GHMM. Within this space, we solve two optimization problems: one to upper bound the PN and another to lower bound it. While inspired by the partial identification literature, our methodology is novel in its application to the hidden and dynamic nature of GHMMs. To establish the foundation for our methodological contributions, we begin with a simplified example where neither hidden states nor dynamic elements are present.

Example 1. Consider a probabilistic model of two random variables X and Y, where $X \in \{0,1\}$ represents whether a medical treatment was substandard (1) or standard (0), and $Y \in \{0,1\}$ indicates whether the patient died (1) or survived (0). Denote the outcome as $Y_x := Y \mid (X = x)$ and assume the marginal probabilities $\mathbb{P}(Y_0)$ and $\mathbb{P}(Y_1)$ are known. Consider a patient who died under substandard treatment ($Y_1 = 1$). We are interested in determining the counterfactual outcome had the treatment been standard. This involves studying the random variable $Y_0 \mid (Y_1 = 1)$, with the probability of necessity defined as $\text{PN} := \mathbb{P}(Y_0 = 0 \mid Y_1 = 1)$. However, estimating PN is not straightforward because it requires knowledge of the *joint* distribution of (Y_0, Y_1) (or equivalently, the SCM underlying this probabilistic model). Without additional assumptions, the joint distribution cannot be uniquely identified from the marginals alone. Moreover, there are infinitely many joint distributions consistent with any given marginals, and each one yields a different estimate of PN.

Example 1 demonstrates that even in a static and fully observable setting, a probabilistic model cannot uniquely identify PN. As detailed in §2, this limitation is well-recognized in the partial identification literature, where linear programming formulations exist to bound PN in such static scenarios. However, in our framework we focus on more complex probabilistic and dynamic latent-state models that present additional challenges.

To handle the latent states, we must infer the posterior distribution that describes how the patient's state evolves over time, conditional on observed data. By "posterior distribution", we mean the conditional distribution obtained by conditioning on observed data, consistent with standard GHMM and graphical modeling usage, and not a Bayesian posterior over the GHMM parameters, which are assumed to be fixed and known. While this task is generally intractable, we leverage the structure of the GHMM and apply filtering and simulation algorithms to efficiently sample from the posterior. These posterior samples then serve as inputs to the optimization problems used to bound PN. The dynamic nature of the GHMM further complicates the problem, transforming the optimization formulations from linear programs into polynomial programs. Additionally, because

we rely on posterior samples of the hidden states, these polynomial optimizations are solved via sample average approximation (SAA). Consequently, while the resulting bounds are approximate rather than exact, their statistical consistency is guaranteed through standard results on SAA. In summary, our primary contribution lies in synthesizing and extending methodologies from several distinct fields to address a novel and impactful problem. Specifically, we integrate techniques from causal inference, state-space modeling, simulation, and optimization.

Before summarizing our other contributions, we emphasize that estimating a counterfactual quantity such as PN differs from *off-policy evaluation* (OPE). This can be see via Example 1, where $\text{PN} = \mathbb{P}(Y_0 = 0 \mid Y_1 = 1)$ but the OPE problem is to estimate $\mathbb{P}(Y_0 = 0)$, which does not depend on the joint distribution of (Y_0, Y_1). Naturally, in a legal context, the quantity of interest would be PN, as it is retrospective and conditions on what actually transpired. This distinction, while subtle, makes our problem considerably more challenging. Throughout this paper, we use the term "counterfactual" to signify that our analysis explicitly conditions on the observed data while simultaneously accounting for a change in the underlying policy.

We make several additional contributions, which we now briefly summarize. First, we demonstrate how domain-specific knowledge can be incorporated into our framework as (linear) constraints in the optimization problems used to bound PN. For example, in healthcare, knowledge such as a "better" treatment leading to a "better" patient outcome can be embedded. Incorporating such constraints leads to tighter bounds on PN, which is practically desirable. We illustrate this in our data-driven breast cancer case study (another contribution), where the impact of these constraints on the bounds is shown to be non-trivial. Finally, we address the computational challenges of our methodology by (1) leveraging the sparsity inherent in the application and (2) reformulating the optimizations and introducing a relaxation of the feasibility space. These techniques enable us to compute quantifiably high-quality bounds for large-scale instances (e.g., spanning up to 100 periods) in a matter of hours, demonstrating the scalability of our approach.

Ultimately, the extent to which our framework proves useful in practice depends on the adoption of GHMM-like models for representing disease progression under various treatment and screening regimes. As we discuss in §2.1, such models are in widespread use across a range of healthcare applications. If this trend continues, then our framework stands to have increasing practical relevance, both in legal contexts and across domains such as public health, epidemiology, and personalized medicine. In these areas, the so-called probabilities of causation (PCs), including PN, are of inherent interest to researchers, practitioners, and policy experts alike. In legal settings, in particular, the use of causal quantities like PN is already well established. For example, Lagakos and Mosteller (1986), Greenland (1999), Tian and Pearl (2000), Cai and Kuroki (2005), Chapter 9 of Pearl (2009),

and Mueller et al. (2022) (among others) all highlight the legal relevance of PN. Notably, in a medical malpractice case, Lagakos and Mosteller (1986) suggested that PN should correspond to the proportion of total damages awarded to the plaintiff. Of course, there will always be debate around the correctness/robustness of any model used to estimate/bound PN. However, this is no different from current practice: models are already used in legal settings to guide decision-making – including judgments and out-of-court settlements – despite their inherent assumptions and limitations.

Paper Outline. We review relevant literature in §2, formulate the problem in §3, and present our solution methodology in §4. In §5, we show how domain-specific knowledge can be embedded in our framework, followed by a breast cancer case study in §6. We elaborate on our techniques to enhance computational scalability in §7 and conclude in §8.

2. Literature Review

Our research builds on three literature streams: (1) (hidden) Markov models in healthcare, (2) partial identification in causal inference, and (3) counterfactual analysis in dynamic models. We note in passing a growing body of causality-related research in the OR/MS community (e.g., Eberhardt et al. 2025, Bertsimas et al. 2022). For instance, Eberhardt et al. (2025) propose an integer-programming method for causal discovery, focusing on learning underlying graphical models while Bertsimas et al. (2022) propose a distributionally robust optimization framework for causal inference under unobserved confounding. We also note that Kaplan-Meier (Kaplan and Meier 1958) and Cox proportional hazards models (Cox 1972) are widely used in biostatistics to estimate population-level survival probabilities or hazard ratios. They allow for competing risks without assuming time-constancy, but do not directly address PN. Our approach targets this legally/medically relevant estimand, using partial (rather than point) identification via a time-constancy assumption. G-estimation (Robins 1994) methods, while powerful in point-identified settings, do not yet extend to partial identification, making our framework a complementary tool.

2.1. Hidden Markov Models in Healthcare

HMMs and their variants are widely employed in healthcare for modeling disease progression, optimizing screening strategies, and evaluating treatment outcomes. In the OR/OM literature, Maillart et al. (2008), Ayer et al. (2012), and Zhang et al. (2012) utilized partially observable Markov models to capture cancer progression and optimize decision-making for mammography and colorectal cancer screening. Similarly, Markov models are extensively used in the medical literature (Sonnenberg and Beck 1993, Alagoz et al. 2010) and in practical applications, such as the University of Wisconsin Breast Cancer Simulator (UWBCS) (UWBCS 2013). These models' ability to account for latent states and dynamic transitions makes them particularly well-suited for representing disease progression. Notably, the UWBCS is employed by the National Cancer

Institute to address critical policy questions related to breast cancer screening and treatment. For further discussion on the practical significance of such models, see §3.1 of Ayer et al. (2012).

The novel contribution of our work lies in integrating GHMMs into a counterfactual inference framework to analyze delay-and-deny practices in health insurance. Specifically, we demonstrate how tight bounds on the probability of necessity (PN) can be obtained, extending the utility of GHMMs to address counterfactual queries.

2.2. Partial Identification in Causal Inference

The partial identification literature seeks to bound counterfactual quantities when causal effects cannot be precisely identified due to unobserved variables or latent confounding. Early work by Manski (1990) provided analytic, albeit often loose bounds. Balke and Pearl (1994) introduced linear programming formulations for sharp bounds, extended by Tian and Pearl (2000). More recently, Duarte et al. (2024) showed that general partial identification problems could be framed as polynomial optimization problems, yielding tight bounds. Zhang et al. (2022) demonstrated that counterfactual bounds could be obtained via polynomial programs, though their approach approximated solutions using MCMC by placing priors over the response function distributions. In contrast, we optimize over the response function distributions rather than assuming any prior.

Our approach builds on this literature by specializing to GHMMs, which introduces unique advantages and challenges. First, the GHMM structure allows us to assume the probabilistic model is known, which allows us to focus on a natural class of SCMs and enables efficient sampling of hidden states via filtering algorithms. Second, the sparsity inherent in healthcare applications, such as the structure of emission and transition probabilities, makes the computationally intensive polynomial optimization tractable. This is in contrast to generic SCMs, where solving the optimization is often approximated or intractable.

2.3. Counterfactual Analysis in Dynamic Models

Recent work has explored counterfactual analysis in dynamic models, but these approaches often assume strong restrictions on the underlying SCM. Buesing et al. (2019) explicitly fixed the SCM, while Oberst and Sontag (2019) and Tsirtsis et al. (2021) used the Gumbel-max SCM to model counterfactual stability. Although these methods provide unique counterfactual estimates, they do not consider the range of feasible SCMs. Our framework relaxes this assumption by constructing bounds over the space of feasible SCMs, providing insight into the variability of counterfactual quantities like PN. Additionally, we address the conjecture by Oberst and Sontag (2019) that counterfactual stability is uniquely modeled by the Gumbel-max SCM, showing instead that it can be enforced via linear constraints in the underlying optimizations.

Unlike prior works, Haugh and Singal (2026) construct counterfactual bounds for expected casino winnings attributable to cheating (EWAC) in the well-known dishonest casino HMM. Their counterfactual of interest – the EWAC – enables the use of linear programming to construct the bounds. In contrast, our focus on healthcare applications introduces greater challenges such as hidden state inference and polynomial optimization, making our problem much more complex. In summary, our work is among the first to bound counterfactual queries in dynamic latent-state models, bridging the gap between causality and real-world applications such as healthcare policy evaluation. Finally, we note that a preliminary version of our work appeared in *ICML* (Haugh and Singal 2023).

3. Problem Formulation

We present the hidden Markov model in §3.1 and define the probability of necessity in §3.2.

3.1. The Generalized Hidden Markov Model (GHMM)

Consistent with the academic literature and models employed in practice, we adopt a dynamic latent-state model to capture disease progression at a patient level. This model, illustrated in Figure 1, operates over T discrete time periods. In each period t, the patient's disease state $H_t \in \mathbb{H}$ (a finite set) is unobserved. Based on H_t and the insurance coverage policy $X_t \in \mathbb{X}$ (finite), an observable outcome or *emission* $O_t \in \mathbb{O}$ (finite) is generated, such as a screening result. The *emission probability* is represented as $e_{hxi} := \mathbb{P}(O_t = i \mid H_t = h, X_t = x)$ for all (h, x, i). Following the emission, the hidden state H_t transitions to H_{t+1} with a *transition probability* $q_{hih'} := \mathbb{P}(H_{t+1} = h' \mid H_t = h, O_t = i)$. The model is defined by three components: $\mathbf{M} \equiv (\mathbf{p}, \mathbf{E}, \mathbf{Q})$, where $\mathbf{p} := [p_h]_h$ denotes the *initial state distribution* with $p_h := \mathbb{P}(H_1 = h)$ for all h, $\mathbf{E} := [e_{hxi}]_{h,x,i}$, and $\mathbf{Q} := [q_{hih'}]_{h,i,h'}$. Similar to the "$Y_x$" notation in Example 1, we define $O_{hx} := O_t \mid (H_t = h, X_t = x)$ for all (h, x) and $H_{hi} := H_{t+1} \mid (H_t = h, O_t = i)$ for all (h, i), leveraging the time-homogeneity of emissions and state transitions. Thus, $e_{hxi} = \mathbb{P}(O_{hx} = i)$ and $q_{hih'} = \mathbb{P}(H_{hi} = h')$ are analogous to the marginals in Example 1.

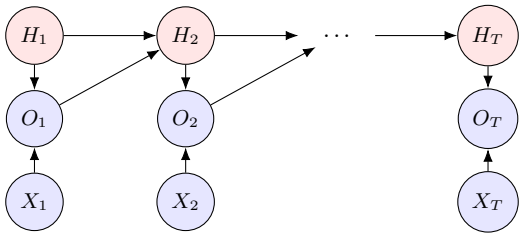


Figure 1 The GHMM. States $H_{1:T} := (H_t)_{t=1}^T$ capture disease progression and are hidden (red). Emissions $O_{1:T}$ (e.g., results of screenings) are observed (blue). $X_{1:T}$ represents the insurance policy (observed). Since $x_{1:T}$ is exogenous, the GHMM could more accurately be described as a GHMM with exogenous inputs.

To make the model more concrete, we map its components to the breast cancer case study detailed in §6. In this context, the time periods correspond to the frequency of mammograms,

which we take to be every six months. The hidden state $H_t \in \{1, \ldots, 7\}$ represents the patient's condition at time t. State 1 denotes that the patient is healthy, while states 2 and 3 correspond to *undiagnosed* in-situ and invasive breast cancer, respectively. States 4 and 5 represent *diagnosed* in-situ and invasive breast cancer, with the assumption that treatment has commenced following diagnosis. Finally, states 6 and 7 are absorbing states that represent recovery from cancer due to treatment and death from cancer, respectively. The observable variable $O_t \in \{1, \ldots, 7\}$ captures the mammogram results or the patient's health status. A value of 1 indicates that no screening was conducted, while a value of 2 corresponds to a negative mammogram result, which could potentially be a false negative. A value of 3 represents a positive mammogram result followed by a negative biopsy, indicating that the patient is healthy but the mammogram produced a false positive. Observations 4 and 5 map to correctly diagnosed in-situ and invasive breast cancer, respectively, where a positive mammogram is confirmed by a positive biopsy. Observations 6 and 7 are used to denote that the patient has either recovered or passed away from breast cancer. The variable $X_t \in \{0, 1\}$ models the insurance company's coverage policy for mammograms. A value of 0 indicates that the insurance company covers the screening, whereas a value of 1 indicates that the company (possibly incorrectly) delays or denies coverage. If the coverage is delayed or denied, no mammogram is performed, and the observation O_t cannot take values of 2, 3, 4, or 5.

As noted in §2, similar models have been extensively studied in the literature and employed in practice. For instance, Ayer et al. (2012) utilize a nearly identical partially observable Markov model to capture cancer progression at the patient level (see Figure 1 in Ayer et al. (2012)). Additionally, the University of Wisconsin Breast Cancer Simulator (UWBCS 2013), widely used to evaluate the effectiveness of breast cancer screening and treatment strategies, employs a comparable discrete-time dynamic model with a 6-month cycle time to represent the progression of breast cancer. These examples underscore the strong academic and practical foundation of our model. Moreover, consistent with existing works (Ayer et al. (2012), UWBCS (2013), and others), we assume the GHMM primitives $(\mathbf{p}, \mathbf{E}, \mathbf{Q})$ are known and calibrate them to real-world data in our §6 case study. This assumption is analogous to assuming the marginals are known in Example 1. It should be clear from our Example 1 discussion that the problem of "estimating" PN remains challenging despite the specified primitives. Before proceeding, we note that our framework could also accommodate a covariate vector $\mathbf{z}$ so that $(\mathbf{p}, \mathbf{E}, \mathbf{Q})$ are all functions of $\mathbf{z}$. This would not complicate the analysis, however, since our focus in bounding PN will focus on a single individual with a particular value of $\mathbf{z}$. As such, there is no need to explicitly model the dependence of $(\mathbf{p}, \mathbf{E}, \mathbf{Q})$ on $\mathbf{z}$.

We are interested in "estimating" PN, the probability a patient would have not died under a screening policy given that she died under a different (presumably delay-and-deny) policy. To this end, we focus on a special state of interest (denoted by h^*) that represents death. Since death is always observable, we assume $H_t = h^*$ iff $O_t = h^*$ for all t. In our breast cancer example, $h^* = 7$.

3.2. Probability of Necessity (PN)

Suppose we observe a sequence of emissions $o_{1:T}$ under the policy $x_{1:T}$. While the true hidden path $h_{1:T}$ remains unobserved, it is known that $h_T = h^*$ (indicating death) at the final period. In the context of breast cancer, the observed emissions for a specific patient might appear as follows:

$$\underbrace{o_1, \ldots, o_{\tau_s-1}}_{\in\{2,3\}}, \underbrace{o_{\tau_s}, \ldots, o_{\tau_e}}_{=1}, \underbrace{o_{\tau_e+1}, \ldots, o_{T-1}}_{\in\{4,5\}}, \underbrace{o_T}_{=7}. \tag{1}$$

That is, the patient underwent regular screenings ($x_{1:\tau_s-1} = 0$) and appeared healthy ($o_{1:\tau_s-1} \in \{2,3\}$) up to and including time $\tau_s - 1$. During the periods τ_s to τ_e, coverage for screenings was delayed or denied ($x_{\tau_s:\tau_e} = 1$; see red font in (1)), and as a result, no screenings were performed ($o_{\tau_s:\tau_e} = 1$). Once coverage was re-approved in period $\tau_e + 1$ ($x_{\tau_e+1} = 0$), the patient was diagnosed with cancer (either in-situ or invasive), leading to the initiation of treatment ($o_{\tau_e+1} \in \{4,5\}$). Unfortunately, the patient passed away at time T.

We define PN to be the probability that a patient would not have died if screening had been covered in every period (intervention policy $\widetilde{x}_{1:T} = 0$), given the observed data $(o_{1:T}, x_{1:T})$. That is,

$$\text{PN} := \mathbb{P}_{\widetilde{\mathbf{M}}}(H_T^{\widetilde{x}_{1:T}} \neq h^*), \tag{2}$$

where $\widetilde{\mathbf{M}} := \mathbf{M} \mid (o_{1:T}, x_{1:T})$ denotes the *posterior model* and $H_T^{\widetilde{x}_{1:T}}$ represents the patient's *counterfactual state* at time T under the intervention policy $\widetilde{x}_{1:T}$. We note that $H_T^{\widetilde{x}_{1:T}}$ has a minor notational inconsistency with Y_x in Example 1 (subscript vs. superscript). Recalling h^* denotes the "death" state, we see from (2) and the definition of $\widetilde{\mathbf{M}}$ that we can also write

$$\begin{aligned} \text{PN} &= \mathbb{P}_{\mathbf{M}}(H_T^{\widetilde{x}_{1:T}} \neq \text{death} \mid o_{1:T}, x_{1:T}) \\ &= \mathbb{P}_{\mathbf{M}}(H_T^{\widetilde{x}_{1:T}} \neq \text{death} \mid o_{1:T}, x_{1:T}, H_T = \text{death}) \end{aligned} \tag{3}$$

since our setup assumes $H_t = \text{death}$ iff $O_t = \text{death}$, and we observe $o_T = \text{death}$. The more explicit expression in (3) is interesting for two reasons. First, we note an alternative version of PN could be defined if we don't condition on $o_{1:T}$ in (2) and (3). This version of PN would be more consistent with the definition of PN from static models where we don't condition on particular characteristics ($o_{1:T}$ in this case) of the individual. But that would not be interesting in our legal setting where it's important to condition on the entire observed *path* of observations. Thus, our definition (2) can be viewed as a form of *conditional* PN, representing the proportion of units with a given profile (treatment vector $x_{1:T}$ and emission sequence $o_{1:T}$, who later died) that would have survived had they instead been treated with $\widetilde{x}_{1:T}$. Second, there are many variations of PN that are possible in our dynamic setting. For example, even if don't condition on $o_{1:T}$, there are 2^T possible paths $x_{1:T}$ – each corresponding to a different screening policy – and each one would yield a different PN.

However, once we observe $x_{1:T}$ and $o_{1:T}$ for a given individual, then (2) becomes the appropriate PN for this individual. That said, other variations of PN for this same individual could also be defined. For example, we might also be interested in the counterfactual probability of surviving for a fixed number of periods *beyond* time T given $x_{1:T}$ and $o_{1:T}$. We note that, with minor tweaks, the modeling and optimization approach we propose can handle all such variations.

A Note on Our Terminology. Up to this point and indeed throughout the paper, we use the term "posterior distribution" to reflect the fact that we are working with a conditional probability distribution, i.e., the distribution we obtain from conditioning on the observed evidence. While "conditional distribution" might be more accurate, it is standard (e.g., see Bishop (2006) or Barber (2012)) to use "posterior distribution" in the GHMM and graphical modeling literature. We make this point in order to emphasize that posterior distributions in this paper should not be confused with the posterior distributions that would arise if we'd adopted a Bayesian modeling approach by placing priors over the GHMM parameters $(\mathbf{p}, \mathbf{E}, \mathbf{Q})$.

4. Solution Approach

We now present our solution approach to bound PN. In §4.1, we outline our algorithm for simulating sample paths of the hidden states, conditioned on the observed data. These samples are then used to define the posterior model in §4.2. Finally, in §4.3, we formulate the polynomial optimization problems to bound PN and explain our approach to solving them.

4.1. Sampling Hidden Paths from the Posterior Distribution

We are interested in understanding the dynamics of the posterior model $\widetilde{\mathbf{M}} = \mathbf{M} \mid (o_{1:T}, x_{1:T})$. To this end, we first simulate B Monte Carlo samples from the distribution of $H_{1:T} \mid (o_{1:T}, x_{1:T})$. We denote these samples by $h_{1:T}(b)$ for $b = 1, \ldots, B$. This sampling is carried out using a filtering algorithm, which computes the *forward probabilities* for the GHMM:

$$\alpha(h_t) := \mathbb{P}(h_t, o_{1:t}, x_{1:t}) \ \forall t \in [T], \ h_t \in \mathbb{H}.$$

Calculating the $\alpha(h_t)$'s is easy, as they admit the following characterization (proof in §EC.1).

Lemma 1. *For $t > 1$, the forward probabilities obey the following recursion:*

$$\alpha(h_t) = e_{h_t x_t o_t} \sum_{h_{t-1} \in \mathbb{H}} q_{h_{t-1} o_{t-1} h_t} \times \alpha(h_{t-1}) \ \forall h_t \in \mathbb{H}. \tag{4}$$

The recursion breaks at $t = 1$ with $\alpha(h_1) = p_{h_1} \times e_{h_1 x_1 o_1}$ for all $h_1 \in \mathbb{H}$.

We leverage the recursion in (4) to efficiently compute the forward probabilities via dynamic programming (see Algorithm 3 in §EC.1), and these probabilities enable us to sample efficiently from the posterior $H_{1:T} \mid (o_{1:T}, x_{1:T})$. We present our sampling procedure in Algorithm 1 and establish its correctness in Proposition 1, which we prove in §EC.1.

Proposition 1. *Algorithm 1 outputs samples from the posterior distribution of $H_{1:T} \mid (o_{1:T}, x_{1:T})$.*

Algorithm 1 `sample_hidden_paths`$(\mathbf{p}, \mathbf{E}, \mathbf{Q}, o_{1:T}, x_{1:T}, B)$

$\alpha =$ `compute_alpha`$(\mathbf{p}, \mathbf{E}, \mathbf{Q}, o_{1:T}, x_{1:T})$ % see Algorithm 3 (§EC.1)
for $b = 1 : B$ **do**
  $h_T(b) \sim \text{Categorical}([\alpha(T,h)]_h)$ % "Categorical" refers to the categorical distribution
  $h_t(b) \sim \text{Categorical}([\alpha(t,h) \times q_{ho_t h_{t+1}(b)}]_h)$ for $t = T-1, \ldots, 1$
end for
return $[h_{1:T}(b)]_b$

4.2. The Posterior Model and Pairwise Marginals

We now use the posterior samples $[h_{1:T}(b)]_b$ from Algorithm 1 to characterize the posterior model $\widetilde{\mathbf{M}} = \mathbf{M} \mid (o_{1:T}, x_{1:T})$. Our key observation is that it is possible to characterize $\widetilde{\mathbf{M}}$ *conditional* on a sample path $h_{1:T}(b)$. Denote by $\widetilde{\mathbf{M}}(b) \equiv (\widetilde{\mathbf{p}}(b), [\widetilde{\mathbf{E}}_t(b)]_t, [\widetilde{\mathbf{Q}}_t(b)]_t)$ the posterior model corresponding to the sample $h_{1:T}(b)$, i.e., $\widetilde{\mathbf{M}} \mid (H_{1:T} = h_{1:T}(b))$. Similar to the primitives $(\mathbf{p}, \mathbf{E}, \mathbf{Q})$ in §3.1, $(\widetilde{\mathbf{p}}(b), [\widetilde{\mathbf{E}}_t(b)]_t, [\widetilde{\mathbf{Q}}_t(b)]_t)$ correspond to initial state, emission, and transition distributions. Given the GHMM structure in Figure 1, it is easy to see that for period 1, the hidden state under the posterior model equals the posterior sample $h_1(b)$ w.p. 1, i.e., $\widetilde{p}_{h_1(b)}(b) = 1$ and $\widetilde{p}_h(b) = 0$ for all $h \neq h_1(b)$. In contrast with $\mathbf{E}$ and $\mathbf{Q}$, $\widetilde{\mathbf{E}}_t(b) := [\widetilde{e}_{thxi}(b)]_{h,x,i}$ and $\widetilde{\mathbf{Q}}_t(b) := [\widetilde{q}_{thih'}(b)]_{h,i,h'}$ are time-dependent:

$$\widetilde{e}_{thxi}(b) = \mathbb{P}(O_{hx} = i \mid O_{h_t(b)x_t} = o_t) \tag{5a}$$

$$\widetilde{q}_{thih'}(b) = \mathbb{P}(H_{hi} = h' \mid H_{h_t(b)o_t} = h_{t+1}(b)). \tag{5b}$$

The O_{hx} and H_{hi} notation is formally introduced in §3.1. The dependence on t arises through the observed data (o_t, x_t) and the posterior samples $(h_t(b), h_{t+1}(b))$. As a result, for each b, $\widetilde{\mathbf{M}}(b)$ is a time-dependent GHMM. Under the SCM subclass that is consistent with the GHMM conditional-independence structure (see §EC.3), standard graphical model arguments imply it is unnecessary to condition on the entire path $(h_{1:T}(b), o_{1:T})$ in (5). In particular, this rules out unobserved cross-time or cross-mechanism dependence beyond what is carried by the GHMM state, so local conditioning suffices. (Formally, in §EC.3.1, we impose SCM independence restrictions that ensure the SCM inherits the GHMM conditional-independence structure.) Moreover, the posterior primitives in (5) are defined for an arbitrary policy, which will ultimately be set to $\widetilde{x}_{1:T}$ (from §3.2).

If $[\widetilde{\mathbf{E}}_t(b)]_t$ and $[\widetilde{\mathbf{Q}}_t(b)]_t$ were known, we could simulate $\widetilde{\mathbf{M}}(b)$ to compute a Monte Carlo estimate of PN by averaging over the B posterior paths. However, $[\widetilde{\mathbf{E}}_t(b)]_t$ and $[\widetilde{\mathbf{Q}}_t(b)]_t$ are unknown. A straightforward application of Bayes' theorem to (5) implies

$$\widetilde{e}_{thxi}(b) = \frac{\mathbb{P}(O_{hx} = i, O_{h_t(b)x_t} = o_t)}{\mathbb{P}(O_{h_t(b)x_t} = o_t)} = \frac{\theta_{hx,h_t(b)x_t}(i, o_t)}{e_{h_t(b)x_t o_t}}$$

$$\widetilde{q}_{thih'}(b) = \frac{\mathbb{P}(H_{hi} = h', H_{h_t(b)o_t} = h_{t+1}(b))}{\mathbb{P}(H_{h_t(b)o_t} = h_{t+1}(b))} = \frac{\pi_{hi,h_t(b)o_t}(h', h_{t+1}(b))}{q_{h_t(b)o_t h_{t+1}(b)}},$$

where the *pairwise marginals* (or the *response function distributions*) are defined as follows:

$$\theta_{\bar{h}\bar{x},hx}(\bar{i},i) := \mathbb{P}(O_{\bar{h}\bar{x}} = \bar{i}, O_{hx} = i) \qquad \forall(\bar{h},\bar{x},\bar{i},h,x,i) \tag{6a}$$

$$\pi_{\bar{h}\bar{i},hi}(\bar{h}',h') := \mathbb{P}(H_{\bar{h}\bar{i}} = \bar{h}', H_{hi} = h') \quad \forall(\bar{h},\bar{i},\bar{h}',h,i,h'). \tag{6b}$$

Similar to the "joint distribution" in Example 1, these pairwise marginals are unknown because the GHMM primitives do not uniquely determine them. However, (6) expresses the *time-dependent* emission and transition distributions in terms of *time-independent* variables $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$. This transformation into the $(\boldsymbol{\theta},\boldsymbol{\pi})$-space allows us to effectively characterize and bound PN. (In (6), we have implicitly assumed $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ to be time-independent, which is very reasonable given the marginals in §3.1 are time-independent.)

§EC.3 formalizes the SCM subclass that underlies the conditional-independence reduction used above and clarifies that optimizing over this SCM class is equivalent to optimizing over the response function distributions $(\boldsymbol{\theta},\boldsymbol{\pi})$ in (6). We therefore work directly in $(\boldsymbol{\theta},\boldsymbol{\pi})$-space, which parameterizes the remaining (unidentified) dependence among the potential outcomes while keeping the GHMM marginals fixed. In §6 and §7, we additionally instantiate two canonical dependence structures – independence and comonotonicity – as benchmark SCMs and report their implied PN values. Since these structures correspond to feasible SCMs, their PN values must lie between our lower and upper bounds. These benchmarks are particularly useful in §7 for gauging the loss introduced by the scalable relaxation that we introduce there.

4.3. Polynomial Optimization Formulations for Bounding PN

We now formulate our polynomial optimization problems where we treat the unknowns $(\boldsymbol{\theta},\boldsymbol{\pi})$ as decision variables and maximize (minimize) the PN to obtain an upper (lower) bound.

Objective. Recall from (2) that $\text{PN} = \mathbb{P}_{\widetilde{\mathbf{M}}}(H_T^{\widetilde{x}_{1:T}} \neq h^*)$, which maps to setting the policy to $\widetilde{x}_{1:T}$ in the posterior model $\widetilde{\mathbf{M}}$. This results in the counterfactual hidden path to be different than the posterior hidden path captured by $[h_{1:T}(b)]_b$. There are two sources of randomness in PN: (i) the true hidden path $H_{1:T}$ (captured by $[h_{1:T}(b)]_b$) and (ii) the posterior model $\widetilde{\mathbf{M}}$ after conditioning on $H_{1:T}$ (captured by $\widetilde{\mathbf{M}}(b)$). Lemma 2 decomposes PN using these two uncertainties.

Lemma 2. *We have*

$$PN = 1 - \lim_{B\to\infty} \frac{1}{B}\sum_{b=1}^{B} \mathbb{P}_{\widetilde{\mathbf{M}}(b)}(H_T^{\widetilde{x}_{1:T}} = h^*). \tag{7}$$

Proofs for this sub-section are in §EC.2. We next express $\mathbb{P}_{\widetilde{\mathbf{M}}(b)}(H_t^{\widetilde{x}_{1:T}})$ in terms of $(\boldsymbol{\theta},\boldsymbol{\pi})$.

Lemma 3. $\mathbb{P}_{\widetilde{\mathbf{M}}(b)}(\bar{h}_t) := \mathbb{P}_{\widetilde{\mathbf{M}}(b)}(H_t^{\widetilde{x}_{1:T}} = \bar{h}_t)$ *obeys the following recursion over* $t \in \{T,\dots,2\}$*:*

$$\mathbb{P}_{\widetilde{\mathbf{M}}(b)}(\bar{h}_t) = \sum_{\bar{h}_{t-1},\bar{o}_{t-1}} \frac{\pi_{\bar{h}_{t-1}\bar{o}_{t-1},h_{t-1}(b)o_{t-1}}(\bar{h}_t, h_t(b))}{q_{h_{t-1}(b)o_{t-1}h_t(b)}} \times \frac{\theta_{\bar{h}_{t-1}\bar{x}_{t-1},h_{t-1}(b)x_{t-1}}(\bar{o}_{t-1}, o_{t-1})}{e_{h_{t-1}(b)x_{t-1}o_{t-1}}} \times \mathbb{P}_{\widetilde{\mathbf{M}}(b)}(\bar{h}_{t-1}).$$

The recursion breaks at $t = 1$ *with* $\mathbb{P}_{\widetilde{\mathbf{M}}(b)}(\bar{h}_1) = 1$ *if* $\bar{h}_1 = h_1(b)$ *and 0 otherwise.*

Lemmas 2 and 3 allow us to express PN in terms of the various primitives, all of which are known (or can be sampled) except for $(\boldsymbol{\theta},\boldsymbol{\pi})$. As soon as we fix $(\boldsymbol{\theta},\boldsymbol{\pi})$, we can estimate PN. However, it is unclear apriori how we should fix $(\boldsymbol{\theta},\boldsymbol{\pi})$. We, therefore, take an agnostic view and compute bounds over PN. The upper (lower) bound is computed by maximizing (minimizing) $\text{PN}(\boldsymbol{\theta},\boldsymbol{\pi} \mid [h_{1:T}(b)]_b)$ over the set of feasible $(\boldsymbol{\theta},\boldsymbol{\pi})$, where we have written $\text{PN}(\boldsymbol{\theta},\boldsymbol{\pi} \mid [h_{1:T}(b)]_b)$ for PN in (7) in order to explicitly recognize its dependence on the B sampled posterior paths $[h_{1:T}(b)]_{b=1}^B$ as well as $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$. Denoting by $\mathcal{F}$ the set of feasible $(\boldsymbol{\theta},\boldsymbol{\pi})$ (discussed below in constraints), we define

$$\text{PN}^{\text{ub}}(B) := \max_{(\boldsymbol{\theta},\boldsymbol{\pi})\in\mathcal{F}} \text{PN}(\boldsymbol{\theta},\boldsymbol{\pi} \mid [h_{1:T}(b)]_b) \tag{8a}$$

$$\text{PN}^{\text{lb}}(B) := \min_{(\boldsymbol{\theta},\boldsymbol{\pi})\in\mathcal{F}} \text{PN}(\boldsymbol{\theta},\boldsymbol{\pi} \mid [h_{1:T}(b)]_b). \tag{8b}$$

Both optimizations in (8) are sample-average approximations as we use $[h_{1:T}(b)]_b$. Thus, $\text{PN}^{\text{ub}}(B)$ and $\text{PN}^{\text{lb}}(B)$ are estimates of the "true" PN^{ub} and PN^{lb}. However, given $[h_{1:T}(b)]_b$ are IID samples, the following consistency result is immediate (cf. Proposition 5.2 in Shapiro et al. (2021)).

Proposition 2. *$(PN^{lb}(B), PN^{ub}(B))$ converges to (PN^{lb}, PN^{ub}) w.p. 1 as $B \to \infty$.*

We can also characterize the asymptotics of $(\text{PN}^{\text{lb}}(B), \text{PN}^{\text{ub}}(B))$ via results in the sample-average approximation theory and we refer the reader to §5.1.2 of Shapiro et al. (2021).

Constraints. We now discuss the feasibility set $\mathcal{F}$ used in (8). Recall from (6) that we use $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ to denote the pairwise marginal distributions over $[O_{hx}]_{h,x}$ and $[H_{hi}]_{h,i}$, respectively. Only the pairwise marginals are required to define our objective function but the full joint distributions of $[O_{hx}]_{h,x}$ and $[H_{hi}]_{h,i}$ are required to define $\mathcal{F}$ and (with a slight abuse of notation) we will also use $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$, respectively, to denote them. To simplify notation, let $k \equiv (h,x)$ and $m \equiv (h,i)$. Hence,

$$O_k \equiv O_{hx},\ e_{ki} \equiv e_{hxi}$$
$$H_m \equiv H_{hi},\ q_{mh'} \equiv q_{hih'}.$$

We have $k \in [K]$ and $m \in [M]$, where $K := |\mathbb{H}||\mathbb{X}|$ and $M := |\mathbb{H}||\mathbb{O}|$. The K- and M- dimensional joint PMFs for all $i_1,\ldots,i_K \in \mathbb{O}$ and $h_1,\ldots,h_M \in \mathbb{H}$ are defined as

$$\theta_{1,\ldots,K}(i_1,\ldots,i_K) := \mathbb{P}(O_1 = i_1,\ldots,O_K = i_K) \tag{9a}$$

$$\pi_{1,\ldots,M}(h_1,\ldots,h_M) := \mathbb{P}(H_1 = h_1,\ldots,H_M = h_M). \tag{9b}$$

We only have *one* joint $\theta_{1,\ldots,K}$ for K random variables, compared to *multiple* pairwise marginals $[\theta_{k\ell}]_{k,\ell}$. Each of the joint probabilities in (9) and pairwise probabilities are treated as decision variables in our optimization problems and must satisfy specific constraints. First, the 1-dimensional marginals of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ must equal the given $[e_{ki}]_{k,i}$ and $[q_{mh}]_{m,h}$ values that define the GHMM:

$$\sum_{\{i_1,\ldots,i_K\}\setminus\{i_k\}} \theta_{1,\ldots,K}(i_1,\ldots,i_K) = e_{1i_k}\ \forall i_k \in \mathbb{O}, k \in [K] \tag{10a}$$

$$\sum_{\{h_1,\ldots,h_M\}\setminus\{h_m\}} \pi_{1,\ldots,M}(h_1,\ldots,h_M) = q_{1h_m} \;\; \forall h_m \in \mathbb{H}, m \in [M]. \tag{10b}$$

Recall that $\mathbf{Q}$ and $\mathbf{E}$ in the RHS of (10) are known. Moreover, as $\mathbf{Q}$ and $\mathbf{E}$ themselves define 1-dimensional probability distributions and thus sum to 1, (10) ensures the same will be true of each of the joint PMFs. Second, we must link the pairwise marginals to the joints:

$$\theta_{k\ell}(i_k, i_\ell) = \sum_{\{i_1,\ldots,i_K\}\setminus\{i_k,i_\ell\}} \theta_{1,\ldots,K}(i_1,\ldots,i_K) \tag{11a}$$

$$\pi_{mn}(h_m, h_n) = \sum_{\{h_1,\ldots,h_M\}\setminus\{h_m,h_n\}} \pi_{1,\ldots,M}(h_1,\ldots,h_M). \tag{11b}$$

(11a) holds for all $i_k, i_\ell \in \mathbb{O}$ and $k, \ell \in [K]$ s.t. $k < \ell$ whereas (11b) holds for all $h_m, h_n \in \mathbb{H}$ and $m, n \in [M]$ s.t. $m < n$. The "$k < \ell$" and "$m < n$" conditions avoid unnecessary duplication. In fact, we do not need to define all pairwise marginals as decision variables but only for "$k < \ell$" and "$m < n$". Finally, we need to ensure non-negativity:

$$\boldsymbol{\theta}, \boldsymbol{\pi} \geq 0, \tag{12}$$

where we now use $(\boldsymbol{\theta}, \boldsymbol{\pi})$ to denote all the decision variables (joint and pairwise).

Let $\mathcal{F}$ be the feasible region over $(\boldsymbol{\theta}, \boldsymbol{\pi})$ defined by the constraints (10), (11), and (12). Observe that $\mathrm{PN}(\boldsymbol{\theta}, \boldsymbol{\pi})$ is a polynomial in $(\boldsymbol{\theta}, \boldsymbol{\pi})$ (cf. Lemmas 2 and 3) and the constraints in $\mathcal{F}$ are linear. Thus, each of the problems in (8) is a polynomial program (Anjos and Lasserre 2011). Denoting by PN^* the PN under the true (unknown) $(\boldsymbol{\theta}, \boldsymbol{\pi})$, we obtain the following bounds.

Proposition 3. *$PN^{lb} \leq PN^* \leq PN^{ub}$.*

Algorithm 2 `compute_bounds`$(\mathbf{p}, \mathbf{E}, \mathbf{Q}, o_{1:T}, x_{1:T}, B, \widetilde{x}_{1:T})$

1: $[h_{1:T}(b)]_b =$ `sample_hidden_paths`$(\mathbf{p}, \mathbf{E}, \mathbf{Q}, o_{1:T}, x_{1:T}, B)$ % see Algorithm 1
2: $\mathrm{PN}^{\mathrm{ub}}(B) = \max_{(\boldsymbol{\theta},\boldsymbol{\pi})\in\mathcal{F}} \mathrm{PN}(\boldsymbol{\theta}, \boldsymbol{\pi} \mid [h_{1:T}(b)]_b)$
3: $\mathrm{PN}^{\mathrm{lb}}(B) = \min_{(\boldsymbol{\theta},\boldsymbol{\pi})\in\mathcal{F}} \mathrm{PN}(\boldsymbol{\theta}, \boldsymbol{\pi} \mid [h_{1:T}(b)]_b)$
4: **return** $(\mathrm{PN}^{\mathrm{lb}}(B), \mathrm{PN}^{\mathrm{ub}}(B))$

We summarize our approach in Algorithm 2, which computes the bounds $(\mathrm{PN}^{\mathrm{lb}}(B), \mathrm{PN}^{\mathrm{ub}}(B))$. Line 1 (sampling) can be executed efficiently (see §4.1) and we discuss three computational considerations behind solving the polynomial optimizations of Lines 2 and 3. First, these optimizations are non-trivial non-convex mathematical programs. To solve these, we utilize the `BARON` solver (Sahinidis 2023), which employs a polyhedral branch-and-cut approach to achieve global optima (Tawarmalani and Sahinidis 2005). This solver performed well in our numerical experiments (§6 and §7), providing global optimality certificates within an $\epsilon = 0.01$ absolute tolerance via duality.

Second, in terms of the problem size, it follows from (6) and (9) that we have *at most* $|\mathbb{H}|^4|\mathbb{O}|^2 + |\mathbb{H}|^2|\mathbb{O}|^2|\mathbb{X}|^2$ pairwise variables and $|\mathbb{O}|^{|\mathbb{H}||\mathbb{X}|} + |\mathbb{H}|^{|\mathbb{H}||\mathbb{O}|}$ joint variables. Similarly, it follows from (10) and (11) that the feasible region $\mathcal{F}$ is defined by *at most* $|\mathbb{O}||\mathbb{H}||\mathbb{X}| + |\mathbb{H}|^2|\mathbb{O}| + |\mathbb{O}|^2|\mathbb{H}|^2|\mathbb{X}|^2 + |\mathbb{H}|^4|\mathbb{O}|^2$ constraints. However, these are merely upper bounds, and we can exploit the sparsity inherent in the underlying application (along with the "$k < \ell$" and "$m < n$" variable and constraint elimination discussed below (11)) to drastically reduce the problem size. For instance, in our breast cancer application, we have $(|\mathbb{H}|, |\mathbb{O}|, |\mathbb{X}|) = (7, 7, 2)$, with the above formulae giving over 10^{41} variables and 10^5 constraints. After exploiting sparsity (discussed in §6), these numbers reduce to 16,124 variables and 610 constraints, making the problem computationally tractable. Furthermore, incorporating domain-specific knowledge can further reduce these numbers, as we discuss in §5.

Third, a naive expansion of the recursion in Lemma 3 results in an exponential number of terms in T, leading to memory issues for moderate to large T. In §7, we demonstrate how this exponential dependence on T can be eliminated through a reformulation of the optimization problem, which introduces polynomial (non-linear) constraints. Despite this added complexity, the reformulation, combined with a problem relaxation, enabled us to compute high-quality bounds on PN in our breast cancer case study for T as large as 100 (see §7.3). In contrast, the original formulation encounters memory issues for T as small as 11, which nonetheless can be practically useful.

5. Embedding Domain-Specific Knowledge

We now illustrate how domain-specific knowledge can be integrated into our framework. One example of such knowledge is *pathwise monotonicity*, or monotonicity (e.g., Pearl 2009), which is particularly relevant in medical contexts. Pathwise monotonicity reflects the expectation that a counterfactual outcome should not be worse than the observed outcome if the counterfactual treatment or policy is better than the treatment or the policy that was actually applied. For instance, in our breast cancer application, suppose a patient has "low severity" cancer (state h) in period t, the cancer is undetected (observation i), and the patient remains in a "low severity" state in period $t+1$. In the counterfactual scenario, if the cancer is detected in period t (observation $\bar{i}$), pathwise monotonicity would require that the patient's state in period $t+1$ cannot transition to a worse state than "low severity". Formally, $\mathbb{P}(H_{h\bar{i}} = \bar{h}' \mid H_{hi} = h) = 0$ for all states $\bar{h}'$ worse than h.

To enforce this condition, we set the corresponding decision variable $\pi_{h\bar{i},hi}(\bar{h}', h)$ to zero, either by adding a linear constraint or by removing the variable from the optimization. Since $\mathbb{P}(H_{h\bar{i}} = \bar{h}' \mid H_{hi} = h) = \pi_{h\bar{i},hi}(\bar{h}', h)/\mathbb{P}(H_{hi} = h)$, and the denominator is a constant (specifically, the transition probability q_{hih}), this approach ensures that monotonicity is upheld. More generally, monotonicity of the form $\mathbb{P}(H_{\bar{h}\bar{i}} = \bar{h}' \mid H_{hi} = h') = 0$ can be enforced by setting $\pi_{\bar{h}\bar{i},hi}(\bar{h}', h')$ to zero. In our breast cancer case study (§6), we embed such knowledge by carefully enumerating all $(h, i, h', \bar{h}, \bar{i}, \bar{h}')$ combinations where monotonicity intuitively holds (details in §EC.5.3).

Incorporating such constraints naturally leads to tighter bounds on PN, which is practically advantageous. Denoting the bounds obtained by adding pathwise monotonicity constraints to the optimizations in (8) as $\text{PN}^{\text{ub}}_{\text{pm}}$ and $\text{PN}^{\text{lb}}_{\text{pm}}$, we have the following result.

Proposition 4. $PN^{lb} \leq PN^{lb}_{pm} \leq PN^{ub}_{pm} \leq PN^{ub}$.

It is important to note that the resulting sharper bounds are valid only if the domain-specific knowledge is accurate. Monotonicity is just one example, and other forms of domain-specific knowledge can also be incorporated into our framework. One such example is the recently proposed concept of *counterfactual stability* (Oberst and Sontag 2019), which, as demonstrated in §A, can be handled similarly via linear constraints or by removing the corresponding decision variables.

Next, we present our data-driven breast cancer case study, where we show that incorporating such domain-specific knowledge can result in *much* tighter bounds on PN.

6. Case Study

We now present our numerical experiments for the breast cancer case study, where we estimate PN for a patient who ultimately died of breast cancer after some of her screenings were incorrectly denied. The intervention under consideration is to approve the screening in every period, thereby addressing the consequences of delay-and-deny practices.

We described the breast cancer GHMM in §3.1. Given patient-level data $(o_{1:T}, x_{1:T})$, we aim to bound PN as defined in (2). Specifically, we consider two different observed paths consistent with the scenario described in (1). The first path is as follows:

$$\underbrace{o_1}_{=2}, \underbrace{o_2, \ldots, o_{T-2}, o_{T-1}}_{=1}, \underbrace{o_T}_{=7}$$
$$\underbrace{x_1}_{=0}, \underbrace{x_2, \ldots, x_{T-2}, x_{T-1}}_{=1}, \underbrace{x_T}_{=0}.$$

Here, we observe a negative test result in period 1, after which screening was not performed for $T-2$ periods (indicated in red font), corresponding to incorrect denial of coverage. The patient ultimately died from breast cancer in period T. Based on the calibrated primitives (discussed in §6.1), the hidden state just prior to death must be $h_{T-1} = 3$ (undiagnosed invasive cancer) since a direct transition from state 2 (undiagnosed in-situ cancer) to state 7 (death) is not possible. Moreover, a direct transition from state 3 (undiagnosed invasive) to state 7 (death) is plausible and has a probability of ≈ 0.15 (discussed in §6.1).

The second path is similar but includes one critical difference: screening was performed in period $T-1$, during which invasive cancer was detected:

$$\underbrace{o_1}_{=2}, \underbrace{o_2, \ldots, o_{T-2}}_{=1}, \underbrace{o_{T-1}}_{=5}, \underbrace{o_T}_{=7}$$
$$\underbrace{x_1}_{=0}, \underbrace{x_2, \ldots, x_{T-2}}_{=1}, \underbrace{x_{T-1}, x_T}_{=0}.$$

In this path, the transition from invasive cancer to death occurred *whilst under treatment*, with a probability of ≈ 0.01 (discussed in §6.1), which is significantly smaller than the ≈ 0.15 probability in Path 1. The occurrence of this low-probability transition suggests that the cancer in Path 2 was particularly aggressive.

In both paths, the intervention policy is $\widetilde{x}_{1:T} = 0$, meaning that screening was approved in every period. In this case study, PN quantifies the likelihood that the patient would not have died had the medical screening not been denied, given that she did die following the denial. This measure directly quantifies the extent to which the patient's death can be attributed to delay-and-deny practices, providing valuable evidence for lawsuits and policy debates, as highlighted in §1.

In §6.1, we outline the model primitives and their calibration to real-world data. This is followed by a detailed description of our experimental setup in §6.2 and the presentation of our main results in §6.3. For brevity, we omit the sensitivity analysis but note that our findings are robust to the calibration. Also, though our numerics are calibrated to real-world data, this is not a "real" case study but it serves to demonstrate our framework in action.

6.1. Model Primitives and Calibration

We introduced the primitives $(\mathbf{p}, \mathbf{E}, \mathbf{Q})$ of the breast cancer GHMM in §3.1. The time periods correspond to mammogram intervals and are set to six months (Ayer et al. 2012). The model includes $|\mathbb{H}| = 7$ states, $|\mathbb{O}| = 7$ emissions, and $|\mathbb{X}| = 2$ actions. The primitives $(\mathbf{p}, \mathbf{E}, \mathbf{Q})$ are calibrated to real-world data using a mix of sources, following approaches similar to those in existing academic literature (e.g., Ayer et al. (2012)) and models used in practice (e.g., UWBCS (2013)).

For brevity, we defer the calibration of $\mathbf{p}$ and $\mathbf{E}$ to §EC.5.1 and focus on the calibration of $\mathbf{Q}$ here. The transition matrix $\mathbf{Q} = [q_{hih'}]_{h,i,h'}$ is defined by $q_{hih'} = \mathbb{P}(H_{t+1} = h' \mid H_t = h, O_t = i)$. Before detailing the calibration, we highlight the sparse structure of $\mathbf{Q}$. To illustrate this, we define the transition matrix for each emission i as $\mathbf{Q}(i) := [q_{hih'}]_{h,h'}$, where each $\mathbf{Q}(i)$ is a 7×7 matrix. These matrices exhibit the following structure:

$$\mathbf{Q}(1) = \begin{bmatrix} \mathsf{q}_{11} & \mathsf{q}_{12} & \mathsf{q}_{13} & & & & \\ & \mathsf{q}_{22} & \mathsf{q}_{23} & & & & \mathsf{q}_{27} \\ & & \mathsf{q}_{33} & & & & \mathsf{q}_{37} \\ & & & \mathsf{q}_{44} & \mathsf{q}_{45} & \mathsf{q}_{46} & \mathsf{q}_{47} \\ & & & & \mathsf{q}_{55} & \mathsf{q}_{56} & \mathsf{q}_{57} \\ & & & & & 1 & \\ & & & & & & 1 \end{bmatrix} \quad \mathbf{Q}(2) = \begin{bmatrix} \mathsf{q}_{11} & \mathsf{q}_{12} & \mathsf{q}_{13} & & \\ & \mathsf{q}_{22} & \mathsf{q}_{23} & & \mathsf{q}_{27} \\ & & \mathsf{q}_{33} & & \mathsf{q}_{37} \\ & & & & \\ & & & & \\ & & & & \\ & & & & \end{bmatrix}$$

$$\mathbf{Q}(3) = \begin{bmatrix} \mathsf{q}_{11} & \mathsf{q}_{12} & \mathsf{q}_{13} & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{bmatrix} \quad \mathbf{Q}(4) = \begin{bmatrix} & \mathsf{q}_{24} & \mathsf{q}_{25} & \mathsf{q}_{26} & \bar{\mathsf{q}}_{27} \\ & & & & \\ & \mathsf{q}_{44} & \mathsf{q}_{45} & \mathsf{q}_{46} & \mathsf{q}_{47} \\ & & & & \\ & & & & \\ & & & & \\ & & & & \end{bmatrix} \quad \mathbf{Q}(5) = \begin{bmatrix} & & & & \\ & & & \mathsf{q}_{35} & \mathsf{q}_{36} & \bar{\mathsf{q}}_{37} \\ & & & & \\ & & & \mathsf{q}_{55} & \mathsf{q}_{56} & \mathsf{q}_{57} \\ & & & & \\ & & & & \end{bmatrix}.$$

A few comments are in order. First, for brevity, we do not display $\mathbf{Q}(6)$ and $\mathbf{Q}(7)$, noting that each has exactly one non-zero entry: $\mathbf{Q}_{66}(6)=1$ and $\mathbf{Q}_{77}(7)=1$. Second, an empty row in $\mathbf{Q}(i)$ indicates an impossible (h,i) combination. For example, if the observed emission is $i=3$ (negative biopsy), the underlying patient state must be healthy, i.e., $h \notin \{2,3,4,5,6,7\}$. Consequently, rows 2 through 7 are empty in $\mathbf{Q}(3)$. Third, note that some parameters overlap across $[\mathbf{Q}(i)]_i$. For instance, q_{11} represents the probability that a healthy patient remains healthy, regardless of whether the emission is 1 (no test), 2 (negative test), or 3 (positive test but negative biopsy). Thus, q_{11} appears in all three matrices $\mathbf{Q}(1)$, $\mathbf{Q}(2)$, and $\mathbf{Q}(3)$. Naturally, if the emission is 4, 5, 6, or 7, the patient cannot be in a healthy state. As a result, the corresponding entries in $\mathbf{Q}(4)$, $\mathbf{Q}(5)$, $\mathbf{Q}(6)$, and $\mathbf{Q}(7)$ are absent, and the first row is entirely empty – another instance of an impossible (h,i) combination. Fourth, some rows in $\mathbf{Q}(i)$ have only partial entries, with all others being zero. For example, if a patient is in state 1 (healthy), their state cannot transition to 4 (diagnosed in-situ with treatment started), 5 (diagnosed invasive with treatment started), 6 (recovered), or 7 (death). Therefore, $\mathsf{q}_{14}=\mathsf{q}_{15}=\mathsf{q}_{16}=\mathsf{q}_{17}=0$, and these transitions are not shown in $\mathbf{Q}(1)$, $\mathbf{Q}(2)$, or $\mathbf{Q}(3)$. Fifth, we use a "bar" in $\bar{\mathsf{q}}_{27}$ (in $\mathbf{Q}(4)$) and $\bar{\mathsf{q}}_{37}$ (in $\mathbf{Q}(5)$) to distinguish them from q_{27} (in $\mathbf{Q}(1)$ and $\mathbf{Q}(2)$) and q_{37} (in $\mathbf{Q}(1)$ and $\mathbf{Q}(2)$). For example, q_{37} represents the probability of transitioning from invasive cancer to death when the cancer was not detected (and hence, no treatment was provided). In contrast, $\bar{\mathsf{q}}_{37}$ represents the probability of transitioning from invasive cancer to death when the cancer was detected and treatment was administered. Naturally, we expect $\bar{\mathsf{q}}_{37} \le \mathsf{q}_{37}$. Next, we discuss calibrating $\mathbf{Q}$ to real data, by iterating over each state $h \in \{1,\ldots,7\}$.

State 1 (healthy). We are interested in $(\mathsf{q}_{11},\mathsf{q}_{12},\mathsf{q}_{13})$. Let's focus on $(\mathsf{q}_{12},\mathsf{q}_{13})$ as $\mathsf{q}_{11}=1-\mathsf{q}_{12}-\mathsf{q}_{13}$. For q_{12} and q_{13}, we use the in-situ and invasive incidence rates, respectively, from Tables 4.12 and 4.11 of NIH (2020) (all races, females). The reported numbers are per year, and we should divide by 2 to convert to a 6-month scale: $\mathsf{q}_{12}=\frac{1}{2}\times\frac{33.0}{100000}$ and $\mathsf{q}_{13}=\frac{1}{2}\times\frac{128.5}{100000}$. Consistent with the 20-80 split in (p_2,p_3) (§EC.5.1), we have $\mathsf{q}_{13}\approx 4\mathsf{q}_{12}$.

State 2 (undiagnosed in-situ cancer). We are interested in $(\mathsf{q}_{22},\mathsf{q}_{23},\mathsf{q}_{27})$ (if cancer is not detected) and $(\mathsf{q}_{24},\mathsf{q}_{25},\mathsf{q}_{26},\bar{\mathsf{q}}_{27})$ (if cancer is detected). First, consider $(\mathsf{q}_{22},\mathsf{q}_{23},\mathsf{q}_{27})$. Table 4.13 of NIH (2020) and Page 26 of UWBCS (2013) indicate that there is no death from in-situ cancer, so $\mathsf{q}_{27}=0$. We assume q_{23} equals the invasive incidence rate q_{13}, giving $\mathsf{q}_{23}=\mathsf{q}_{13}$ and $\mathsf{q}_{22}=1-\mathsf{q}_{23}-\mathsf{q}_{27}=1-\mathsf{q}_{13}$. Next, consider $(\mathsf{q}_{24},\mathsf{q}_{25},\mathsf{q}_{26},\bar{\mathsf{q}}_{27})$. Since $\mathsf{q}_{27}=0$ and $\bar{\mathsf{q}}_{27}\le\mathsf{q}_{27}$ (recall comment #5 above), we set $\bar{\mathsf{q}}_{27}=0$. As all in-situ cancer patients survive if treated, no transitions to invasive cancer occur when in-situ cancer is detected, so $\mathsf{q}_{25}=0$. Finally, we have $\mathsf{q}_{24}+\mathsf{q}_{26}=1$. The split between q_{24} and q_{26} is irrelevant in terms of survival, as there is no positive probability path from state 4 to death (as will become clear when we discuss state 4 below).

State 3 (undiagnosed invasive cancer). We are interested in $(\mathsf{q}_{33}, \mathsf{q}_{37})$ (if cancer is not detected) and $(\mathsf{q}_{35}, \mathsf{q}_{36}, \bar{\mathsf{q}}_{37})$ (if cancer is detected). First, consider $(\mathsf{q}_{33}, \mathsf{q}_{37})$. Here, q_{37} represents the probability of dying from invasive breast cancer under no treatment. According to Johnstone et al. (2000), the 5-year and 10-year survival rates for invasive breast cancer patients (without treatment) are 18.4% and 3.6%, respectively. Using the 5-year rate, we calibrate as follows: $(1-\mathsf{q}_{37})^{10} = 0.184 \implies \mathsf{q}_{37} \approx 15.6\%$. Similarly, using the 10-year rate: $(1-\mathsf{q}_{37})^{20} = 0.036 \implies \mathsf{q}_{37} \approx 15.3\%$. The consistency between these estimates supports the assumption of time-invariance. Minimizing the sum of squared errors over both data points: $\min_{\mathsf{q}_{37}\in[0,1]}\{((1-\mathsf{q}_{37})^{10} - 0.184)^2 + ((1-\mathsf{q}_{37})^{20} - 0.036)^2\}$, we obtain $\mathsf{q}_{37} = 0.1554$. Naturally, $\mathsf{q}_{33} = 1 - \mathsf{q}_{37}$. Next, consider $(\mathsf{q}_{35}, \mathsf{q}_{36}, \bar{\mathsf{q}}_{37})$. Here, q_{36} and $\bar{\mathsf{q}}_{37}$ represent the probabilities of recovering and dying from invasive cancer under treatment. Using Table 4.14 of NIH (2020), we calibrate using 10 data points corresponding to the year 2007 (see Figure 2): $\mathsf{q}_{36} = 0.0459$ and $\bar{\mathsf{q}}_{37} = 0.0113$. As a sanity check, note that $\bar{\mathsf{q}}_{37} < \mathsf{q}_{37}$. Finally, $\mathsf{q}_{35} = 1 - \mathsf{q}_{36} - \bar{\mathsf{q}}_{37}$.

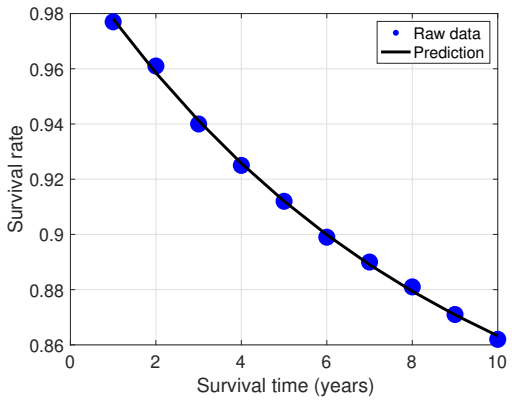


Figure 2 Calibration of $(\mathsf{q}_{36}, \bar{\mathsf{q}}_{37})$. Under our time-homogeneous Markov model, starting in the state of invasive breast cancer (under treatment), the survival rate after x years is given by: $1 - \bar{\mathsf{q}}_{37} \sum_{i=0}^{2x-1} (1 - \mathsf{q}_{36} - \bar{\mathsf{q}}_{37})^i$. We estimate $(\mathsf{q}_{36}, \bar{\mathsf{q}}_{37})$ by minimizing the sum of squared errors over the ten blue data points in the plot. This yields $(\mathsf{q}_{36}, \bar{\mathsf{q}}_{37}) = (0.0459, 0.0113)$. The predicted survival rate based on this fit is shown as the black curve in the plot.

States 4 and 5 (diagnosed in-situ and invasive cancer). We are interested in $(\mathsf{q}_{44}, \mathsf{q}_{45}, \mathsf{q}_{46}, \mathsf{q}_{47})$ and $(\mathsf{q}_{55}, \mathsf{q}_{56}, \mathsf{q}_{57})$. Under our Markov model (which is "memoryless"), it seems reasonable to set $(\mathsf{q}_{44}, \mathsf{q}_{45}, \mathsf{q}_{46}, \mathsf{q}_{47}) = (\mathsf{q}_{24}, \mathsf{q}_{25}, \mathsf{q}_{26}, \bar{\mathsf{q}}_{27})$ and $(\mathsf{q}_{55}, \mathsf{q}_{56}, \mathsf{q}_{57}) = (\mathsf{q}_{35}, \mathsf{q}_{36}, \bar{\mathsf{q}}_{37})$. This follows because, in a memoryless Markov model, the timing of treatment initiation does not affect the transition probabilities. For example, consider a patient in state 2 (undiagnosed in-situ cancer), where the cancer is detected and treatment begins. Here, q_{25} represents the probability of transitioning to invasive cancer. Now consider a different patient already in state 4 (diagnosed in-situ cancer) where treatment is ongoing. In this case, q_{45} represents the same probability of transitioning to invasive cancer. Since the Markov model does not retain memory of when treatment started, q_{25} and q_{45} are equivalent. This reasoning extends analogously to other parameters.

States 6 (recovery) and 7 (death). These two states are absorbing and hence, $\mathsf{q}_{66} = \mathsf{q}_{77} = 1$.

6.2. Experimental Setup

For both Path 1 and Path 2 defined above, we vary $T \in \{4, \ldots, 10\}$, where larger values of T indicate slower cancer progression. PN bounds are computed using our framework (Algorithm 2), implemented in `MATLAB` (MATLAB 2021). The feasibility set $\mathcal{F}$ over $(\boldsymbol{\theta}, \boldsymbol{\pi})$ corresponds to (10), (11), and (12). We used the `MATLAB-BARON` interface (Sahinidis 2023) with `CPLEX` (IBM 2017) as the "LP/MIP solver" to solve the polynomial optimization problems. Each problem instance was solved to global optimality within minutes to hours (depending on T), with an "absolute termination tolerance" of 0.01 on an Intel Xeon E5 processor with 16 GB RAM. Problems with $T = 10$ took the longest time on average (approximately 2 hours). For each optimization, $B = 100$ samples were generated using Algorithm 1, requiring less than 1 second. We found $B = 100$ sufficient for stable sample-average approximations. This stability was verified by running the optimizations across 20 different random seeds (for each (path, T) pair) and confirming that the standard deviations of the optimal objectives were small. We also solved the polynomial optimization problems with $B = 500$ and observed the resulting solutions to be nearly identical to those with $B = 100$.

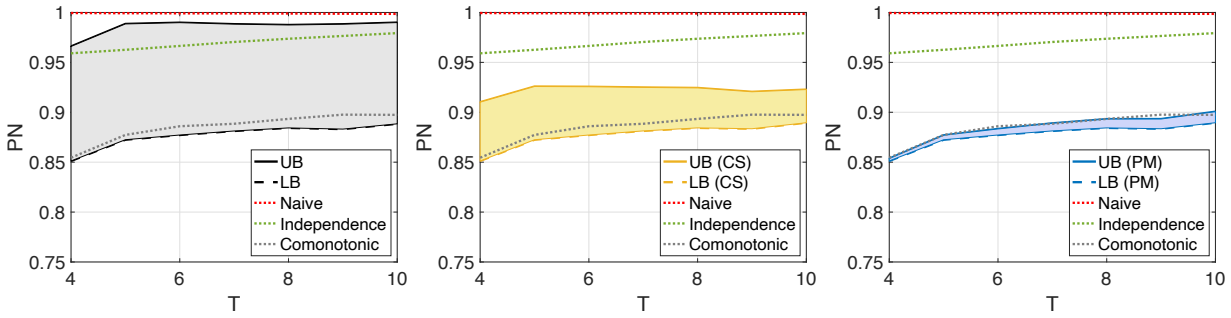
As noted below Algorithm 2, the sparse structure of $\mathbf{E}$ and $\mathbf{Q}$ significantly reduces the size of the problem. For instance, when considering the pairwise decision variables $\pi_{\bar{h}\bar{i},hi}(\bar{h}', h')$, we can eliminate those variables corresponding to impossible (h, i, h') or $(\bar{h}, \bar{i}, \bar{h}')$ combinations; see §6.1. This observation also applies to all joint variables. By carefully removing such variables and the corresponding constraints (details in §EC.5.2), the problem size is reduced from 10^{41} variables and 10^5 constraints to 16,124 variables and 610 constraints, respectively.

We note in passing that, for each of our optimization problems, we included the constraint that the objective value (PN), being a probability, must lie in $[0, 1]$. While this constraint is technically redundant, it proved helpful in speeding up solver convergence for certain problem instances, possibly by reducing the search space since the solver does not inherently know that the objective is a probability. In those instances, convergence was achieved within a few hours when the constraint was imposed, compared to no convergence even after 24 hours without it.

Hence, while we use an off-the-shelf solver, implementing our approach requires careful effort. Furthermore, although one can work on developing a custom solver for our class of optimization problems (polynomial objective with linear constraints), it is unclear if that effort is warranted given `BARON` worked very well in our use cases.

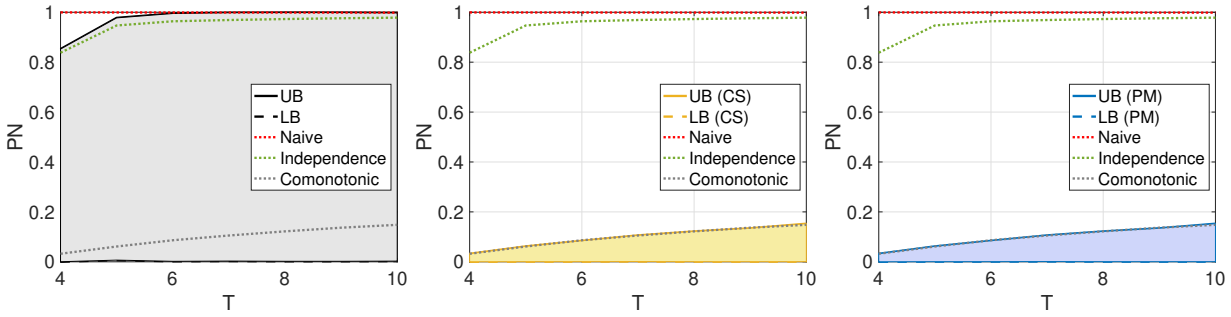
6.3. Main Results and Discussion

We display our PN bounds as a function of T for Paths 1 and 2 in Figures 3 and 4, respectively. Alongside the baseline bounds $(\text{PN}^{\text{lb}}, \text{PN}^{\text{ub}})$ computed via our optimization framework of §4 (`UB`



(a) UB/LB (b) UB/LB with CS (c) UB/LB with PM

Figure 3 PN bounds and estimates for Path 1 as we vary $T \in \{4, \ldots, 10\}$. Observe that the `LB`, `LB (CS)`, and `LB (PM)` curves coincide (forming the lowest curve in each sub-figure). The bounds are averaged over 20 seeds, with a standard deviation (s.d.) smaller than 0.01 for each reported value. Due to this small magnitude, we omit the ± 1 s.d. bars to reduce visual clutter in the sub-figures. We used 10^4 Monte Carlo samples to compute the `naive` estimate and the PN estimates for the `independence` and `comonotonic` SCMs; these simulations took $\approx$ 2 minutes.



(a) UB/LB (b) UB/LB with CS (c) UB/LB with PM

Figure 4 PN bounds and estimates for Path 2 as we vary $T \in \{4, \ldots, 10\}$. Similar to Figure 3, the `LB`, `LB (CS)`, and `LB (PM)` curves coincide (forming the lowest curve in each sub-figure). Additionally, for Path 2, the `UB (CS)` and `UB (PM)` curves also coincide.

and `LB`), we also show the bounds obtained by incorporating domain-specific knowledge, such as counterfactual stability (`UB (CS)` and `LB (CS)`) and pathwise monotonicity (`UB (PM)` and `LB (PM)`) as described in §5. As highlighted in §4.2, we also compute PN estimates based on the `independence` and `comonotonic` SCMs, which amount to fixing a particular SCM and are calculated via Monte Carlo simulation, with further details provided in §EC.4.

In addition, we provide a `naive` estimate, which disregards the information in the observed path and is equivalent to the off-policy evaluation estimate $\mathbb{P}_{\mathbf{M}}(H_T^{\widetilde{x}_{1:T}} \neq h^*)$. Since the probability of developing and dying from breast cancer within any 5-year interval ($T = 10$) is quite low, the `naive` estimate is nearly indistinguishable from 1 in both figures. Moreover, it is path-independent and falls outside the valid bounds (`UB` and `LB`) in Figures 3a and 4a.

For Path 1, we obtain relatively tight bounds, with PN consistently above 0.85, indicating that regardless of the underlying SCM, the patient would have survived at time T in the counterfactual

world with high probability. Even without additional knowledge such as counterfactual stability or pathwise monotonicity, the gap between the lower and upper bounds is within approximately 10 percentage points. This gap narrows to around 5 percentage points when counterfactual stability is imposed and further reduces to about 1 percentage point when pathwise monotonicity is applied. Notably, the `comonotonic` estimate closely aligns with the pathwise monotonicity bounds, and the PN estimates for the `comonotonic` and `independence` SCMs approximately span the range of the bounds (Figure 3a). This suggests that valuable insights into the range of counterfactual query values can be gained by simulating SCMs based on standard copulas when solving the polynomial optimization problems becomes computationally challenging.

We note that the LB and UB under counterfactual stability do not coincide, indicating the existence of infinitely many SCMs that are consistent with counterfactual stability. This follows from the fact that any convex combination of two copulas is itself a copula. This result resolves the open question posed by Oberst and Sontag (2019) regarding whether SCMs other than the Gumbel-max SCM can satisfy counterfactual stability.

For Path 2, we observe that the lower bounds are close to 0, even when counterfactual stability or pathwise monotonicity is imposed. This aligns with the fact that despite being diagnosed in period $T-1$ (and subsequently treated), the patient eventually died, suggesting that the cancer was "aggressive". The bounds without counterfactual stability or pathwise monotonicity are relatively loose, reflecting the lack of knowledge to reason effectively in a counterfactual scenario. However, when knowledge is injected via counterfactual stability or pathwise monotonicity, the bounds become significantly tighter. Notably, the upper bound on Path 2, under either counterfactual stability or pathwise monotonicity, never exceeds 0.2. This highlights the considerable strength of such knowledge in refining the counterfactual bounds.

We believe such bounds have practical implications. For instance, if the bounds suggest that the patient would likely have survived (or not survived) in the counterfactual world had the scans been permitted, this information could be invaluable in legal contexts. Importantly, a wide gap between the lower and upper bounds does not indicate a weakness of our framework but rather reflects a lack of knowledge about the underlying SCM or insufficient information in the observed data. In such cases, incorporating domain-specific knowledge can reduce the space of feasible SCMs and yield tighter bounds.

Our experiments have so far considered $T \leq 10$, as memory issues arise for $T > 10$ due to the exponential growth of terms (in T) in the polynomial program's objective (as discussed towards the end of §4.3). In §7, we show that a reformulation and relaxation of the feasible region significantly improve scalability, enabling high-quality bounds for T up to 100 within a few hours of computation.

7. Computational Scalability

We begin in §7.1, where we show how the optimization can be reformulated to eliminate the exponential dependence on T. Next, in §7.2, we discuss a relaxation of the optimization, which significantly reduces the dimensionality of the problem. Finally, in §7.3, we present numerical experiments demonstrating how these two techniques together enable high-quality bounds for T as large as 100 within a few hours of compute time.

7.1. Reformulation

It is straightforward to see that a naive expansion of PN (as per Lemmas 2 and 3) leads to a number of terms that grows exponentially with T. This is undesirable, as it causes memory issues for $T > 10$ in our numerics presented in §6. However, this exponential dependence can be eliminated through a reformulation of the optimization problem:

1. Define $\mathbb{P}_{\widetilde{\mathbf{M}}(b)}(\bar{h}_t)$ from Lemma 3 as a decision variable $\gamma^t_{\bar{h}_t b}$ for all $(t, \bar{h}_t, b) \in [T] \times \mathbb{H} \times [B]$.
2. Add the Lemma 3 equations as constraints (for each $(t, \bar{h}_t, b) \in [T] \times \mathbb{H} \times [B]$).
3. The objective now is simply the expression in Lemma 2.

The reformulated maximization problem is as follows:

$$\max_{(\boldsymbol{\theta},\boldsymbol{\pi})\in\mathcal{F},\boldsymbol{\gamma}} \left\{1 - \frac{1}{B}\sum_{b=1}^{B} \gamma^T_{h^* B}\right\} \tag{13a}$$

$$\text{s.t. } \gamma^t_{hb} = \sum_{h'\in\mathbb{H}}\sum_{o'\in\mathbb{O}} \frac{\pi_{h'o',h_{t-1}(b)o_{t-1}}(h, h_t(b))}{q_{h_{t-1}(b)o_{t-1}h_t(b)}} \times \frac{\theta_{h'\widetilde{x}_{t-1},h_{t-1}(b)x_{t-1}}(o', o_{t-1})}{e_{h_{t-1}(b)x_{t-1}o_{t-1}}} \times \gamma^{t-1}_{h'b} \qquad \forall t > 1,\ h \in \mathbb{H},\ b \in [B] \tag{13b}$$

$$\gamma^1_{hb} = 1 \qquad h = h_1(b),\ \forall b \in [B] \tag{13c}$$

$$\gamma^1_{hb} = 0 \qquad \forall h \neq h_1(b),\ b \in [B]. \tag{13d}$$

For the reformulated minimization, we simply change the "max" to a "min" in the objective function (13a). As before, the feasibility set $\mathcal{F}$ over $(\boldsymbol{\theta}, \boldsymbol{\pi})$ is defined by (10), (11), and (12). This set can also incorporate additional constraints, such as counterfactual stability or pathwise monotonicity, or correspond to the lower-dimensional space over the pairwise marginals, as discussed in §7.2. Clearly, (13) has a linear objective and polynomial constraints, making it a polynomial program. While the objective no longer has an exponential dependence on T, this reformulation introduces $|\mathbb{H}|TB$ additional decision variables and polynomial constraints to the original formulation in (8). Hence, the size of (13) scales with both T and B. However, we found it to scale much more gracefully (with respect to T) than the original formulation, as we discuss in §7.3 below.

7.2. Approximation

The optimization problems in (8) operate over the joint PMFs of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ ("joint optimization"). As noted near the end of §4.3, the problem size can grow exponentially with the primitives $|\mathbb{H}|$, $|\mathbb{X}|$, and $|\mathbb{O}|$, requiring $|\mathbb{O}|^{|\mathbb{H}||\mathbb{X}|} + |\mathbb{H}|^{|\mathbb{H}||\mathbb{O}|}$ decision variables to represent these joint distributions. While application-specific sparsity can sometimes mitigate this exponential growth (as we demonstrate in the breast cancer application), we now explore the use of tractable approximations to the optimization problems in (8) that could be beneficial in more general settings.

Our key observation, which follows directly from Lemmas 2 and 3, is that the objective functions in (8) depend on the joint PMFs of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ only through their *pairwise* marginals. The joint PMFs (and their associated decision variables) are included to ensure that the pairwise marginal decision variables in the objectives correspond to valid probability distributions for $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$. This reflects the fact that, while each pairwise marginal is a valid bivariate distribution, it does not necessarily imply the existence of a full multivariate distribution consistent with all of the pairwise marginals.

This observation suggests that we can easily derive a relaxed version of the problem by omitting the decision variables for the joint distributions and retaining only those for the pairwise marginals. The constraint set for the relaxed problem is straightforward to specify. Since the pairwise variables represent 2-dimensional PMFs, they must satisfy basic probability axioms. Specifically, they must be non-negative and consistent with their known 1-dimensional marginals, ensuring

$$\sum_{h'} \pi_{\bar{h}\bar{i},hi}(\bar{h}',h') = q_{\bar{h}\bar{i}\bar{h}'} \;\forall(\bar{h},\bar{i},\bar{h}',h,i) \tag{14a}$$

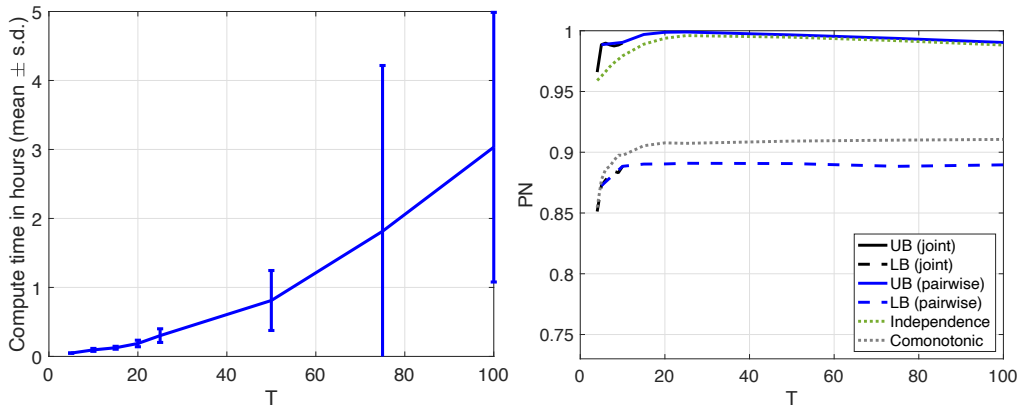
$$\sum_{\bar{h}'} \pi_{\bar{h}\bar{i},hi}(\bar{h}',h') = q_{hih'} \;\forall(\bar{h},\bar{i},h,i,h') \tag{14b}$$

$$\sum_{i} \theta_{\bar{h}\bar{x},hx}(\bar{i},i) = e_{\bar{h}\bar{x}\bar{i}} \;\forall(\bar{h},\bar{x},\bar{i},h,x) \tag{14c}$$

$$\sum_{\bar{i}} \theta_{\bar{h}\bar{x},hx}(\bar{i},i) = e_{hxi} \;\forall(\bar{h},\bar{x},h,x,i). \tag{14d}$$

These constraints are analogous to (10) in §4.3. Indeed, if (10) is satisfied, then so is (14). However, the reverse does not hold. Therefore, optimizing over the feasibility space defined by (14) and non-negativity, denoted as $\mathcal{F}'$ ("pairwise optimization"), is a relaxation of the problem where we optimize over $\mathcal{F}$. It is worth noting that one can go beyond the 2-dimensional pairwise marginals and incorporate higher-dimensional marginals, which would yield a closer approximation to $\mathcal{F}$. However, this improvement comes at the cost of increased time and space complexity.

Despite being a strict relaxation, we found the pairwise optimization to produce high-quality solutions and to be highly scalable; see §7.3. This scalability is primarily due to the lower dimensionality of the decision variables. Specifically, the pairwise optimization involves *at most* $|\mathbb{H}|^4|\mathbb{O}|^2 + |\mathbb{H}|^2|\mathbb{O}|^2|\mathbb{X}|^2$ decision variables. After exploiting the sparsity in the breast cancer application (along



(a) Compute time (b) Probability of necessity (PN)

Figure 5 Evaluating the computational performance of our techniques in §7.1 and §7.2 on Path 1 from §6. The compute times in sub-figure (a) correspond to the blue curves in sub-figure (b) ("pairwise"), which are the focus of this section. All results are averaged over 20 seeds, with vertical bars representing ±1 standard deviations. In sub-figure (b), the UB and LB curves for the "joint" optimization (black color) match those in Figure 3a and extend only up to $T = 10$ due to memory limitations for $T > 10$. To reduce clutter, standard deviation bars are omitted in sub-figure (b), but the maximum standard deviation value is noted to be less than 0.01.

with the "$k < \ell$" and "$m < n$" variable elimination discussed below (11)), the pairwise optimization reduces to just 1,082 decision variables, compared to 16,124 in the joint optimization. Although the pairwise optimization has more constraints than the joint optimization, the difference is modest (2,085 vs. 610). These numbers correspond to the objective formulation in §4.3 rather than the reformulation in §7.1. The reformulation in §7.1 adds $|\mathbb{H}|TB$ decision variables and $|\mathbb{H}|TB$ constraints to both the pairwise and joint optimizations.

7.3. Computational Performance

We now compute upper and lower bounds on PN using (a) the reformulation described in §7.1 and (b) the relaxed constraint set defined by the pairwise marginals as discussed in §7.2. For brevity, we focus on Path 1 from §6. The implementation details remain consistent with §6, including the use of `MATLAB-BARON` with `CPLEX` as the "LP/MIP solver", an absolute termination tolerance of 0.01, $B = 100$ samples for sample-average approximation, and averaging over 20 seeds. To evaluate the scalability of our approach, we experiment with $T \in \{5, 10, 15, 20, 25, 50, 75, 100\}$. Note that $T = 100$ represents an order of magnitude larger horizon than the maximum value ($T = 10$) considered in §6. Our results are presented in Figure 5, with Figure 5a illustrating scalability and Figure 5b demonstrating solution quality.

In Figure 5a, we display the compute time as a function of T for the reformulated pairwise optimization problems. Compute time refers to the total time required to compute both LB and UB. The solver was allowed to run until convergence to global optimality on a single core with at

most 16 GB RAM. For $T = 100$, the average compute time was 3 hours (averaged over 20 seeds), with a minimum of 1.2 hours and a maximum of 8.4 hours, demonstrating the scalability of our approach. It is worth noting that after eliminating redundant variables and constraints, exploiting sparsity, and applying the reformulation in §7.1 along with the pairwise approximation in §7.2, the $T = 100$ and $B = 100$ optimization contains $71,082$ decision variables, 2085 linear constraints, and $70,000$ polynomial constraints.

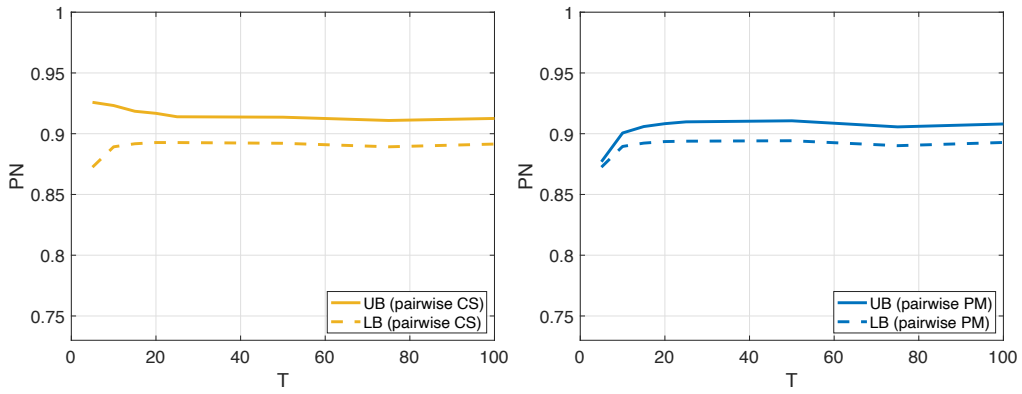
In Figure 5b, we display the PN values as a function of T. The "joint" UB and LB curves are the same as the ones in Figure 3a, and only go as far as $T = 10$ because of the aforementioned memory issues for $T > 10$. The "pairwise" UB and LB curves are the focus of this section, and our goal here is to evaluate the quality of the "pairwise" bounds (blue curves), which we do in two ways. First, as we are able to solve the joint optimization for $T \leq 10$, we can use the "joint" bounds as benchmarks for the "pairwise" bounds. As may be seen from the figure, the joint and pairwise bounds are very close to each other (for values of $T \leq 10$). In particular, the joint and pairwise lower bounds coincide and equal 0.8725 and 0.8884 for $T = 5$ and $T = 10$, respectively. The pairwise and joint upper bounds also coincide and equal 0.9885 for $T = 5$. The only difference between the two is the upper bound for $T = 10$ with values of 0.9904 and 0.9899, respectively. We therefore conclude that the pairwise bounds provide a very good approximation to the joint bounds, at least when $T \leq 10$.

Second, for $T > 10$, we utilize the fact that we can simulate the independence and comonotonic SCMs (conditional on the observations and original policy) (§EC.4), which, by definition, are feasible solutions to the joint optimization. Consequently, we know that maximizing (minimizing) over the joint distribution will produce an upper (lower) bound that is no lower (higher) than the independence (comonotonic) curves in the figure. For example, the gap between the independence and pairwise UB curves (for $T > 10$) is never greater than 0.01, indicating a maximum loss of 0.01 when restricting to pairwise marginals. Similarly, the maximum gap between the comonotonic and pairwise LB curves is approximately 0.02. Thus, the pairwise bounds serve as a high-quality approximation to the joint bounds, even for $T > 10$.

We can also embed counterfactual stability and pathwise monotonicity into the pairwise optimization, as both constraints apply to the pairwise variables. The resulting bounds are shown in Figure 6. Naturally, these bounds are tighter than the pairwise bounds presented in Figure 5b. In particular, the UB becomes significantly tighter, while the LB remains largely unchanged.

8. Concluding Remarks

We propose a principled framework for bounding the probability of necessity (PN) in the context of health insurance delay-and-deny practices. By leveraging GHMMs to model disease progression



(a) Counterfactual stability (b) Pathwise monotonicity

Figure 6 PN bounds obtained by embedding counterfactual stability and pathwise monotonicity constraints in the pairwise optimization. All results are averaged over bounds computed from 20 different seeds, with each seed generating $B = 100$ paths. To reduce clutter, standard deviation bars are omitted; however, the maximum standard deviation value is noted to be less than 0.01.

and screening and combining them with structural causal modeling and optimization techniques, we provide a powerful framework for bounding counterfactual probabilities in complex, dynamic systems. Our approach contributes to the partial identification literature by addressing the challenges of hidden and dynamic states within GHMMs. Specifically, we demonstrate how to construct a natural class of plausible causal models and compute PN bounds through polynomial optimization along with sample average approximation. By incorporating domain-specific knowledge, such as pathwise monotonicity and counterfactual stability, our framework achieves tighter bounds, enhancing its practical utility. The results from our breast cancer case study highlight the practical relevance of PN bounds, particularly in legal and policy contexts where understanding the impact of delay-and-deny practices is critical. These bounds can provide insights into whether a patient might have survived under different policies, offering valuable evidence for legal cases and policy evaluations. Further, our framework's scalability – achieving high-quality bounds for horizons up to $T = 100$ – demonstrates its applicability to large-scale problems.

While our focus has been on PN, our framework can be extended to other counterfactual probabilities, such as the *probability of sufficiency* (PS) or the *probability of necessity and sufficiency* (PNS). Additionally, it can accommodate intermediate rewards, enabling applications beyond mortality to broader outcomes, such as *quality-adjusted life years* (QALYs). The framework is also general enough to address scenarios beyond healthcare as long as the underlying system can be captured via a dynamic latent-state model. Overall, our framework provides a rigorous and flexible approach to counterfactual analysis in dynamic latent-state models.

Acknowledgments

The authors thank the *M&SOM* Frontiers in Operations panel for a highly constructive review process, with special thanks to Huseyin Topaloglu (Department Editor) and the anonymous Associate Editor and referee. The authors also thank Jim Smith for feedback on an early version of this work, Nick Sahinidis for assistance with `BARON`-related issues, and the participants of the 2025 MSOM Healthcare SIG – especially the discussant, Kraig Delana – for their helpful comments.

References

Alagoz O, Ayer T, Erenay FS (2010) Operations Research Models for Cancer Screening. *Lung Cancer* 215:161–840.

Anjos MF, Lasserre JB (2011) *Handbook on Semidefinite, Conic and Polynomial Optimization*, volume 166 (Springer Science & Business Media).

Ayer T, Alagoz O, Stout NK (2012) A POMDP Approach to Personalize Mammography Screening Decisions. *Operations Research* 60(5):1019–1034.

Balke A, Pearl J (1994) Counterfactual Probabilities: Computational Methods, Bounds and Applications. *Uncertainty in Artificial Intelligence*, 46–54 (San Francisco (CA)).

Barber D (2012) *Bayesian Reasoning and Machine Learning* (Cambridge University Press).

Bertsimas D, Imai K, Li ML (2022) Distributionally Robust Causal Inference with Observational Data. *arXiv preprint arXiv:2210.08326* .

Biniek J, Sroczynski N, Tricia N (2024) Use of Prior Authorization in Medicare Advantage Exceeded 46 Million Requests in 2022 [URL], *KFF*.

Bishop CM (2006) *Pattern Recognition and Machine Learning*. Information Science and Statistics (Springer), ISBN 978-0-387-31073-2.

Buesing L, Weber T, Zwols Y, Heess N, Racaniere S, Guez A, Lespiau JB (2019) Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search. *International Conference on Learning Representations*.

Cai Z, Kuroki M (2005) Variance Estimators for Three "Probabilities of Causation". *Risk Analysis* 25(6):1611–1620.

Chaffin J, Wernau J (2024) Doctors Say Dealing With Health Insurers Is Only Getting Worse [URL], *The Wall Street Journal*.

Cox DR (1972) Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 34(2):187–220.

Duarte G, Finkelstein N, Knox D, Mummolo J, Shpitser I (2024) An Automated Approach to Causal Inference in Discrete Settings. *Journal of the American Statistical Association* 119(547):1778–1793.

Eberhardt F, Kaynar N, Siddiq A (2025) Discovering Causal Models with Optimization: Confounders, Cycles, and Instrument Validity. *Management Science* 71(4):3283–3302.

Eckholm E (1993) $89 Million Awarded Family Who Sued H.M.O. [URL], *The New York Times*.

Freed M, Biniek J, Damico A, Neuman T (2024) Medicare Advantage in 2024 [URL], *KFF*.

Greenland S (1999) Relation of Probability of Causation to Relative Rsk and Doubling Dose: A Methodologic Error That Has Become a Social Problem. *American Journal of Public Health* 89(8):1166–1169.

Grimm C (2022) Some Medicare Advantage Organization Denials of Prior Authorization Requests Raise Concerns About Beneficiary Access to Medically Necessary Care [URL], *Office of Inspector General*.

Haugh M, Singal R (2023) Counterfactual Analysis in Dynamic Latent State Models. *International Conference on Machine Learning* (PMLR).

Haugh M, Singal R (2026) A Counterfactual Analysis of the Dishonest Casino. *Journal of Causal Inference* (Accepted) .

Huff D (2024) Judge Rules HMSA Contracts are 'Unconscionable' in Lawsuit from Doctors and Patients [URL], *Hawaii News Now*.

IBM (2017) ILOG CPLEX Optimizer Version 12.8 .

Johnstone PA, Norton MS, Riffenburgh RH (2000) Survival of Patients with Untreated Breast Cancer. *Journal of Surgical Oncology* 73(4):273–277.

Kaplan EL, Meier P (1958) Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* 53(282):457–481.

Lagakos SW, Mosteller F (1986) Assigned Shares in Compensation for Rdiation-related Cancers. *Risk Analysis* 6(3):345–357.

Maillart LM, Ivy JS, Ransom S, Diehl K (2008) Assessing Dynamic Breast Cancer Screening Policies. *Operations Research* 56(6):1411–1427.

Manski CF (1990) Nonparametric Bounds on Treatment Effects. *The American Economic Review* 80(2):319–323.

Mathews A (2024) After UnitedHealthcare CEO Killing, Doctors Speak Out [URL], *The Wall Street Journal*.

MATLAB (2021) *MATLAB Version 9.10.0 (R2021b)* (Natick, Massachusetts: The MathWorks Inc.).

Miller TC, Rucker P, Armstrong D (2023) "Not Medically Necessary": Inside the Company Helping America's Biggest Health Insurers Deny Coverage for Care [URL], *ProPublica*.

Mole B (2023) UnitedHealth Uses AI Model with 90% Error Rate to Deny Care, Lawsuit Alleges [URL], *Ars Technica*.

Mueller S, Li A, Pearl J (2022) Causes of Effects: Learning Individual Responses from Population Data. *International Joint Conference on Artificial Intelligence*, 2712–2718.

NHS (2024) Martha's Rule [URL], *National Health Service*.

NIH (2020) SEER Cancer Statistics Review (CSR) 1975-2017 [URL] .

Oberst M, Sontag D (2019) Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models. *International Conference on Machine Learning*, 4881–4890.

Pearl J (2009) *Causality* (Cambridge University Press), 2nd edition.

Rabin RC (2023) When Should Women Get Regular Mammograms? At 40, U.S. Panel Now Says [URL], *The New York Times*.

Robins JM (1994) Correcting for Non-Compliance in Randomized Trials Using Structural Nested Mean Models. *Communications in Statistics - Theory and Methods* 23(8):2379–2412.

Sahinidis NV (2023) *BARON 2023.1.5: Global Optimization of Mixed-Integer Nonlinear Programs,* User's Manual.

Shapiro A, Dentcheva D, Ruszczynski A (2021) *Lectures on Stochastic Programming: Modeling and Theory* (SIAM).

Sklar A (1959) Fonctions De Répartition à N Dimensions Et Leurs Marges. *Publ. Inst. Statist. Univ. Paris* 8:229–231.

Sonnenberg FA, Beck JR (1993) Markov Models in Medical Decision Making: A Practical Guide. *Medical Decision Making* 13(4):322–338.

Sprague BL, Trentham-Dietz A (2009) Prevalence of Breast Carcinoma In Situ in the United States. *Journal of the American Medical Association* 302(8):846.

Stockton A (2024) 'What's My Life Worth?' The Big Business of Denying Medical Care [URL], *The New York Times*.

Tawarmalani M, Sahinidis NV (2005) A Polyhedral Branch-and-cut Approach to Global Optimization. *Mathematical Programming* 103(2):225–249.

Nitta vs HMSA (2024) Case ID: CAAP-24-0000079 [URL], *Hawaii Medical Association*.

Tian J, Pearl J (2000) Probabilities of Causation: Bounds and Identification. *Annals of Mathematics and Artificial Intelligence* 8:287–313.

Tsirtsis S, De A, Rodriguez M (2021) Counterfactual Explanations in Sequential Decision Making Under Uncertainty. Ranzato M, Beygelzimer A, Dauphin Y, Liang P, Vaughan JW, eds., *Advances in Neural Information Processing Systems*, volume 34, 30127–30139 (Curran Associates, Inc.).

UWBCS (2013) University of Wisconsin Breast Cancer Simulation Model [URL] .

Zhang J, Denton BT, Balasubramanian H, Shah ND, Inman BA (2012) Optimization of Prostate Biopsy Referral Decisions. *Manufacturing & Service Operations Management* 14(4):529–547.

Zhang J, Tian J, Bareinboim E (2022) Partial Counterfactual Identification from Observational and Experimental Data. *International Conference on Machine Learning*, 26548–26558 (PMLR).

Appendix A: Counterfactual Stability

Counterfactual stability, proposed by Oberst and Sontag (2019), has become a popular approach to perform counterfactual analysis in certain settings. Before embedding it via linear constraints in our framework, we illustrate it using the simple probabilistic model of Example 1, involving two variables, X and Y. Suppose we observe $Y = y$ under policy $X = x$. With $Y_x := Y \mid (X = x)$, counterfactual stability requires that the counterfactual outcome under an interventional policy $\widetilde{x}$ (denoted by $Y_{\widetilde{x}} \mid (Y_x = y)$) cannot be y' (for $y' \neq y$) if $\mathbb{P}(Y_{\widetilde{x}} = y)/\mathbb{P}(Y_x = y) \geq \mathbb{P}(Y_{\widetilde{x}} = y')/\mathbb{P}(Y_x = y')$. Simply put, counterfactual stability states that if y was observed and becomes relatively more likely than y' under the intervention, then the counterfactual outcome cannot be y'.

Oberst and Sontag (2019) show that the Gumbel-max SCM obeys counterfactual stability and can be efficiently simulated, providing *one* estimate of a counterfactual quantity (e.g., PN) that obeys counterfactual stability. We now demonstrate that, similar to pathwise monotonicity, it is possible to characterize the *entire* space of SCMs that satisfy counterfactual stability.

Consider our GHMM setting. Suppose the patient is in state h, with emission i, and transitions to state h', corresponding to the realization $H_{hi} = h'$. For $\bar{h}' \neq h'$, assume $\mathbb{P}(H_{\bar{h}\bar{i}} = h')/\mathbb{P}(H_{hi} = h') \geq \mathbb{P}(H_{\bar{h}\bar{i}} = \bar{h}')/\mathbb{P}(H_{hi} = \bar{h}')$, or equivalently, $q_{\bar{h}\bar{i}h'}/q_{hih'} \geq q_{\bar{h}\bar{i}\bar{h}'}/q_{hi\bar{h}'}$. If counterfactual stability holds, this implies $\pi_{\bar{h}\bar{i},hi}(\bar{h}', h') = 0$. Thus, for hidden state transitions, invoking counterfactual stability is equivalent to adding the following linear constraints for all $(h, i, h', \bar{h}, \bar{i}, \bar{h}')$:

$$\pi_{\bar{h}\bar{i},hi}(\bar{h}', h') = 0 \text{ if } \frac{q_{\bar{h}\bar{i}h'}}{q_{hih'}} \geq \frac{q_{\bar{h}\bar{i}\bar{h}'}}{q_{hi\bar{h}'}}.$$

Similarly, counterfactual stability can be modeled for emissions by adding the following linear constraints for all $(h, x, i, \bar{h}, \bar{x}, \bar{i})$:

$$\theta_{\bar{h}\bar{x},hx}(\bar{i}, i) = 0 \text{ if } \frac{e_{\bar{h}\bar{x}i}}{e_{hxi}} \geq \frac{e_{\bar{h}\bar{x}\bar{i}}}{e_{hx\bar{i}}}.$$

As with pathwise monotonicity, instead of adding these constraints, we can remove the corresponding decision variables from the optimization. This allows us to characterize the space of all structural causal models that satisfy counterfactual stability. Enforcing counterfactual stability results in tighter bounds; however, the bounds may not be "legitimate" if the true $(\boldsymbol{\theta}, \boldsymbol{\pi})$ does not adhere to counterfactual stability.

Bounding Counterfactual Outcomes of Health Insurance Delay-and-Deny Practices (E-companion)

Martin B. Haugh and Raghav Singal

Appendix EC.1: Further Details on Sampling Hidden Paths

Lemma 1. *For $t>1$, the forward probabilities obey the following recursion:*

$$\alpha(h_t) = e_{h_t x_t o_t} \sum_{h_{t-1}\in\mathbb{H}} q_{h_{t-1}o_{t-1}h_t} \times \alpha(h_{t-1}) \;\forall h_t \in \mathbb{H}. \tag{4}$$

The recursion breaks at $t=1$ with $\alpha(h_1) = p_{h_1} \times e_{h_1 x_1 o_1}$ for all $h_1 \in \mathbb{H}$.

Proof. We begin with $t=1$ and note that for all $h_1 \in \mathbb{H}$, $\alpha(h_1) := \mathbb{P}(o_1 \mid h_1, x_1) \times \mathbb{P}(h_1 \mid x_1) \times \mathbb{P}(x_1) = \mathbb{P}(o_1 \mid h_1, x_1) \times \mathbb{P}(h_1) \times \mathbb{P}(x_1) = e_{h_1 x_1 o_1} \times p_{h_1}$, where we use the fact that the policy $x_{1:T}$ is deterministic. For $t>1$, note that for all $h_t \in \mathbb{H}$,

$$\begin{aligned}
\alpha(h_t) &= \sum_{h_{t-1}} \mathbb{P}\left(h_t, h_{t-1}, o_{1:t-1}, o_t, x_{1:t}\right) \\
&= \sum_{h_{t-1}} \mathbb{P}\left(o_t \mid h_t, h_{t-1}, o_{1:t-1}, x_{1:t}\right) \times \mathbb{P}\left(h_t \mid h_{t-1}, o_{1:t-1}, x_{1:t}\right) \\
&\qquad \times \mathbb{P}(x_t \mid h_{t-1}, o_{1:t-1}, x_{1:t-1}) \times \mathbb{P}\left(h_{t-1}, o_{1:t-1}, x_{1:t-1}\right) \\
&= \sum_{h_{t-1}} \mathbb{P}\left(o_t \mid h_t, x_t\right) \times \mathbb{P}\left(h_t \mid h_{t-1}, o_{t-1}\right) \times \mathbb{P}(x_t) \times \mathbb{P}\left(h_{t-1}, o_{1:t-1}, x_{1:t-1}\right) \\
&= \mathbb{P}(x_t) \times \mathbb{P}\left(o_t \mid h_t, x_t\right) \sum_{h_{t-1}} \mathbb{P}\left(h_t \mid h_{t-1}, o_{t-1}\right) \alpha(h_{t-1}) = e_{h_t x_t o_t} \sum_{h_{t-1}} q_{h_{t-1}o_{t-1}h_t} \times \alpha(h_{t-1}),
\end{aligned}$$

where we again use the fact that the policy $x_{1:T}$ is deterministic. This completes the proof. □

Efficiently Computing the Forward Probabilities. See Algorithm 3. For numerical stability, we compute these probabilities on the log scale. In particular, the recursion in (4) (Lemma 1) can be expressed as follows (by taking log on both sides):

$$\log(\alpha(h_t)) = \log(e_{h_t x_t o_t}) + \log \sum_{h_{t-1}} \underbrace{q_{h_{t-1}o_{t-1}h_t} \times \alpha(h_{t-1})}_{=\exp\left\{\log(q_{h_{t-1}o_{t-1}h_t}) + \log(\alpha(h_{t-1}))\right\}},$$

where the "$\log\sum\exp$" term can be evaluated using a standard logsumexp function.

Proposition 1. *Algorithm 1 outputs samples from the posterior distribution of $H_{1:T} \mid (o_{1:T}, x_{1:T})$.*

Proof. Observe that

$$\begin{aligned}
\mathbb{P}\left(h_{1:T} \mid o_{1:T}, x_{1:T}\right) &= \mathbb{P}\left(h_1 \mid h_{2:T}, o_{1:T}, x_{1:T}\right) \times \mathbb{P}\left(h_{T-1} \mid h_T, o_{1:T}, x_{1:T}\right) \times \mathbb{P}\left(h_T \mid o_{1:T}, x_{1:T}\right) \\
&= \mathbb{P}\left(h_1 \mid h_2, o_{1:T}, x_{1:T}\right) \times \mathbb{P}\left(h_{T-1} \mid h_T, o_{1:T}, x_{1:T}\right) \times \mathbb{P}\left(h_T \mid o_{1:T}, x_{1:T}\right).
\end{aligned}$$

We can therefore sample sequentially via the following two steps:

Algorithm 3 `compute_alpha`$(\mathbf{p}, \mathbf{E}, \mathbf{Q}, o_{1:T}, x_{1:T})$

$\log(\alpha(1,h)) = \log(p_h) + \log(e_{hx_1o_1})$ for all $h \in \mathbb{H}$ % initialize period 1
for $t = 2:T$ **do**
  for $h \in \mathbb{H}$ **do**
    `t_minus_terms`$(h') = \log(q_{h'o_{t-1}h}) + \log(\alpha(t-1,h'))$ for all $h' \in \mathbb{H}$
    $\log(\alpha(t,h)) = \log(e_{hx_to_t}) + \text{logsumexp}($`t_minus_terms`$)$
  end for
end for
return $\exp(\log(\alpha))$

1. Draw h_T from $\mathbb{P}(h_T \mid o_{1:T}, x_{1:T})$.
2. For $t = T-1, \ldots 1$, draw h_t from $\mathbb{P}(h_t \mid h_{t+1}, o_{1:T}, x_{1:T})$.

For Step 1, we can use $\alpha(h_T)$ since $\mathbb{P}(h_T \mid o_{1:T}, x_{1:T}) \propto \alpha(h_T)$. For Step 2, observe that for $t < T$,

$$\begin{aligned}
\mathbb{P}(h_t \mid h_{t+1}, o_{1:T}, x_{1:T}) &\propto \mathbb{P}(h_t, h_{t+1} \mid o_{1:T}, x_{1:T}) \\
&\propto \alpha(h_t) \times \mathbb{P}(h_{t+1}, x_{t+1} \mid h_t, o_t) && \text{[see (EC.1) below]} \\
&= \alpha(h_t) \times \mathbb{P}(h_{t+1} \mid h_t, o_t) \times \mathbb{P}(x_{t+1}) \\
&= \alpha(h_t) \times q_{h_to_th_{t+1}}.
\end{aligned}$$

For the second proportionality, note the *pairwise marginal* $\mathbb{P}(h_t, h_{t+1} \mid o_{1:T}, x_{1:T})$ is proportional to

$$\begin{aligned}
&\propto \mathbb{P}(o_{1:t}, o_{t+1}, o_{t+2:T}, x_{1:t}, x_{t+1}, x_{t+2:T}, h_{t+1}, h_t) \\
&= \mathbb{P}(o_{t+2:T}, x_{t+2:T} \mid o_{1:t}, o_{t+1}, x_{1:t}, x_{t+1}, h_{t+1}, h_t) \times \mathbb{P}(o_{1:t}, o_{t+1}, x_{1:t}, x_{t+1}, h_{t+1}, h_t) \\
&= \mathbb{P}(o_{t+2:T}, x_{t+2:T} \mid h_{t+1}, o_{t+1}) \times \mathbb{P}(o_{t+1} \mid o_{1:t}, h_{t+1}, h_t, x_{1:t}, x_{t+1}) \times \mathbb{P}(o_{1:T}, h_{t+1}, h_t, x_{1:t}, x_{t+1}) \\
&= \mathbb{P}(o_{t+2:T}, x_{t+2:T} \mid h_{t+1}, o_{t+1}) \times \mathbb{P}(o_{t+1} \mid h_{t+1}, x_{t+1}) \times \mathbb{P}(h_{t+1}, x_{t+1} \mid o_{1:T}, x_{1:t}, h_t) \times \mathbb{P}(o_{1:t}, x_{1:t}, h_t) \\
&= \mathbb{P}(o_{t+2:T}, x_{t+2:T} \mid h_{t+1}, o_{t+1}) \times \mathbb{P}(o_{t+1} \mid h_{t+1}, x_{t+1}) \times \mathbb{P}(h_{t+1}, x_{t+1} \mid h_t, o_t) \times \mathbb{P}(o_{1:t}, x_{1:t}, h_t) \\
&= \mathbb{P}(o_{t+2:T}, x_{t+2:T} \mid h_{t+1}, o_{t+1}) \times \mathbb{P}(o_{t+1} \mid h_{t+1}, x_{t+1}) \times \mathbb{P}(h_{t+1}, x_{t+1} \mid h_t, o_t) \times \alpha(h_t). && \text{(EC.1)}
\end{aligned}$$

This completes the proof. □

Appendix EC.2: Proofs of Lemmas 2 and 3

Lemma 2. *We have*

$$PN = 1 - \lim_{B \to \infty} \frac{1}{B} \sum_{b=1}^{B} \mathbb{P}_{\widetilde{\mathbf{M}}(b)}(H_T^{\widetilde{x}_{1:T}} = h^*). \tag{7}$$

Proof. Observe that

$$\text{PN} = 1 - \mathbb{P}_{\widetilde{\mathbf{M}}}(H_T^{\widetilde{x}_{1:T}} = h^*) \qquad \text{[by definition]}$$

$$
\begin{aligned}
&= 1 - \mathbb{E}_{\widetilde{\mathbf{M}}}[\mathbb{I}\{H_T^{\widetilde{x}_{1:T}} = h^*\}] && [\mathbb{P}(Y=y) = \mathbb{E}[\mathbb{I}\{Y=y\}]] \\
&= 1 - \mathbb{E}_{H_{1:T}}[\mathbb{E}_{\widetilde{\mathbf{M}}|H_{1:T}}[\mathbb{I}\{H_T^{\widetilde{x}_{1:T}} = h^*\}]] && [\text{law of total expectation}] \\
&= 1 - \lim_{B\to\infty} \frac{1}{B}\sum_{b=1}^{B} \mathbb{P}_{\widetilde{\mathbf{M}}(b)}(H_T^{\widetilde{x}_{1:T}} = h^*). && [\text{law of large numbers}]
\end{aligned}
$$

The proof is now complete. □

Lemma 3. *$\mathbb{P}_{\widetilde{\mathbf{M}}(b)}(\bar{h}_t) := \mathbb{P}_{\widetilde{\mathbf{M}}(b)}(H_t^{\widetilde{x}_{1:T}} = \bar{h}_t)$ obeys the following recursion over $t \in \{T, \dots, 2\}$:*

$$
\mathbb{P}_{\widetilde{\mathbf{M}}(b)}(\bar{h}_t) = \sum_{\bar{h}_{t-1}, \bar{o}_{t-1}} \frac{\pi_{\bar{h}_{t-1}\bar{o}_{t-1}, h_{t-1}(b)o_{t-1}}(\bar{h}_t, h_t(b))}{q_{h_{t-1}(b)o_{t-1}h_t(b)}} \times \frac{\theta_{\bar{h}_{t-1}\bar{x}_{t-1}, h_{t-1}(b)x_{t-1}}(\bar{o}_{t-1}, o_{t-1})}{e_{h_{t-1}(b)x_{t-1}o_{t-1}}} \times \mathbb{P}_{\widetilde{\mathbf{M}}(b)}(\bar{h}_{t-1}).
$$

The recursion breaks at $t=1$ with $\mathbb{P}_{\widetilde{\mathbf{M}}(b)}(\bar{h}_1) = 1$ if $\bar{h}_1 = h_1(b)$ and 0 otherwise.

Proof. For $t \in \{T, T-1, \dots, 2\}$, observe that

$$
\begin{aligned}
\mathbb{P}_{\widetilde{\mathbf{M}}(b)}(\bar{h}_t) &= \sum_{\bar{h}_{t-1}\in\mathbb{H}} \sum_{\bar{o}_{t-1}\in\mathbb{O}} \mathbb{P}_{\widetilde{\mathbf{M}}(b)}(\bar{h}_t, \bar{h}_{t-1}, \bar{o}_{t-1}) \\
&= \sum_{\bar{h}_{t-1}\in\mathbb{H}} \sum_{\bar{o}_{t-1}\in\mathbb{O}} \mathbb{P}_{\widetilde{\mathbf{M}}(b)}(\bar{h}_t \mid \bar{h}_{t-1}, \bar{o}_{t-1}) \mathbb{P}_{\widetilde{\mathbf{M}}(b)}(\bar{o}_{t-1} \mid \bar{h}_{t-1}) \mathbb{P}_{\widetilde{\mathbf{M}}(b)}(\bar{h}_{t-1}) \\
&= \sum_{\bar{h}_{t-1}\in\mathbb{H}} \sum_{\bar{o}_{t-1}\in\mathbb{O}} \widetilde{q}_{t-1\bar{h}_{t-1}\bar{o}_{t-1}\bar{h}_t}(b) \times \widetilde{e}_{t-1\bar{h}_{t-1}\bar{x}_{t-1}\bar{o}_{t-1}}(b) \times \mathbb{P}_{\widetilde{\mathbf{M}}(b)}(\bar{h}_{t-1}) \\
&= \sum_{\bar{h}_{t-1}\in\mathbb{H}} \sum_{\bar{o}_{t-1}\in\mathbb{O}} \frac{\pi_{\bar{h}_{t-1}\bar{o}_{t-1}, h_{t-1}(b)o_{t-1}}(\bar{h}_t, h_t(b))}{q_{h_{t-1}(b)o_{t-1}h_t(b)}} \times \frac{\theta_{\bar{h}_{t-1}\bar{x}_{t-1}, h_{t-1}(b)x_{t-1}}(\bar{o}_{t-1}, o_{t-1})}{e_{h_{t-1}(b)x_{t-1}o_{t-1}}} \times \mathbb{P}_{\widetilde{\mathbf{M}}(b)}(\bar{h}_{t-1}).
\end{aligned}
$$

The base case ($t=1$) holds since the counterfactual hidden state in period 1 equals the posterior sample $h_1(b)$ (as discussed in §4.2). The proof is now complete. □

Appendix EC.3: The Structural Causal Model and Response Distributions

We define the SCM in §EC.3.1 and model it via response function distributions in §EC.3.2.

EC.3.1. The Structural Causal Model (SCM)

A generic SCM for the GHMM from Figure 1 is illustrated in Figure EC.1. Consider, for example, the observation O_t at any time t. It is modeled as a stochastic function of its parents (H_t, X_t). To capture this stochasticity, we introduce an exogenous noise vector $\mathbf{V}_t := [V_{thx}]_{h,x}$, consisting of $|\mathbb{H}||\mathbb{X}|$ U$[0,1]$ random variables. The vector representation of $\mathbf{V}_t$ (rather than a scalar) allows us to account for the fact that each $O_{thx} := O_t \mid (H_t = h, X_t = x)$ defines a *distinct* random variable (or *potential outcome*) for all (h, x). These random variables may be independent or exhibit positive or negative dependence. To model these scenarios, we associate each O_{thx} with a corresponding

noise variable V_{thx}. The *dependence structure* among $[V_{thx}]_{h,x}$ determines the dependence structure among $[O_{thx}]_{h,x}$. The structural equation for O_t is given by

$$O_t = f(H_t, X_t, \mathbf{V}_t) = \sum_{h,x} f_{hx}(V_{thx}) \mathbb{I}\{H_t = h, X_t = x\}, \tag{EC.2a}$$

where $f_{hx}(\cdot)$ is the inverse CDF derived from the emission distribution $[e_{hxi}]_i$. This function generates $O_t \mid (H_t = h, X_t = x)$ using the noise variable $V_{thx} \sim \mathrm{U}[0,1]$.

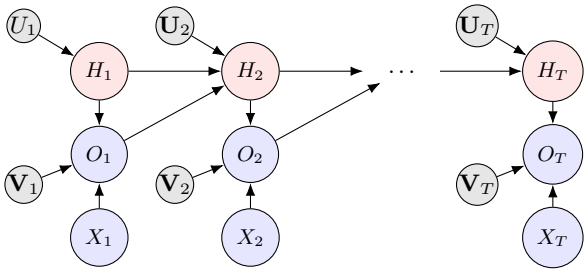


Figure EC.1 The SCM underlying the GHMM (Figure 1). The only difference between the SCM and the GHMM is the addition of exogenous noise nodes $[\mathbf{U}_t, \mathbf{V}_t]_t$.

Similarly, for $t > 1$, each $H_{thi} := H_t \mid (H_{t-1} = h, O_{t-1} = i)$ is a distinct random variable (or potential outcome) for all (h, i). For time $t = 1$, there are no parents (H_0, O_0) for H_1, so a single $\mathrm{U}[0,1]$ noise variable U_1 suffices to generate H_1. For $t > 1$, we associate each H_{thi} with its own $\mathrm{U}[0,1]$ noise variable U_{thi}, such that

$$H_t = g(H_{t-1}, O_{t-1}, \mathbf{U}_t) = \sum_{h,i} g_{hi}(U_{thi}) \mathbb{I}\{H_{t-1} = h, O_{t-1} = i\} \tag{EC.2b}$$

where $g_{hi}(\cdot)$ is the inverse CDF derived from the transition distribution $[q_{hih'}]_{h'}$. Here, $\mathbf{U}_t := [U_{thi}]_{h,i}$ consists of $|\mathbb{H}||\mathbb{O}|$ $\mathrm{U}[0,1]$ random variables.

The exogenous noise vectors $\{\mathbf{U}_t\}_{t=1}^T$ and $\{\mathbf{V}_t\}_{t=1}^T$ are assumed to be mutually independent across time and type. (Type = emissions vs transitions.) Specifically, (i) $\mathbf{U}_t$ is independent of $\mathbf{U}_{t'}$ for all $t \neq t'$, (ii) $\mathbf{V}_t$ is independent of $\mathbf{V}_{t'}$ for all $t \neq t'$, and (iii) $\mathbf{U}_t$ is independent of $\mathbf{V}_{t'}$ for all t, t'. Such independence ensures that the SCM is consistent with the conditional-independence structure implied by the GHMM in Figure 1. This is exactly the restriction invoked in §4.2 to justify the local conditioning in (5). This assumption, akin to the commonly invoked *sequential ignorability* assumption in dynamic models, significantly reduces the space of feasible SCMs by prohibiting direct links between the $\mathbf{U}_t$'s and $\mathbf{V}_t$'s in Figure EC.1. Given the established validity of GHMMs for modeling disease progression (as discussed in earlier sections), this assumption seems very reasonable. Nonetheless, while this independence assumption defines a natural subclass of SCMs compatible with the underlying GHMM (an assumption on which our framework to bound

PN critically relies), there are other SCMs that would also be compatible with the underlying GHMM. For example, Haugh and Singal (2026) provide a simple example of such an SCM in the context of their dishonest casino model. The class of SCMs we do consider is still very rich, however, since we allow dependence among the components of $\mathbf{V}_t = [V_{thx}]_{h,x}$ and among the components of $\mathbf{U}_t = [U_{thi}]_{h,i}$.

The representation in (EC.2) enables us to model $[O_{thx}]_{h,x}$ ($[H_{thi}]_{h,i}$) and capture any dependence structure among these random variables by specifying the joint multivariate distribution of $\mathbf{V}_t$ ($\mathbf{U}_t$). Since the univariate marginals of $\mathbf{V}_t$ ($\mathbf{U}_t$) are known to be U$[0,1]$, specifying their multivariate distribution amounts to defining the dependence structure or the *copula* of $\mathbf{V}_t$ ($\mathbf{U}_t$). For example, if the V_{thx}'s are mutually independent (the *independence copula*) and $H_t = h'$, then inferring the conditional distribution of $V_{th'x'}$ provides no information about the V_{thx}'s for $(h,x) \neq (h',x')$. Alternatively, if $V_{thx} = V_{th'x'}$ for all (h,x) and (h',x'), this models perfect positive dependence (the *comonotonic copula*), where inferring the conditional distribution of $V_{th'x'}$ simultaneously determines the conditional distribution of all V_{thx}'s.

We emphasize that the exogenous vectors $\mathbf{U}_t$ and $\mathbf{V}_t$ are critical for *counterfactual* analysis, as different joint distributions among the components of $\mathbf{U}_t$ and among the components of $\mathbf{V}_t$ can yield significantly different values of PN. However, if counterfactual analysis is not the goal and we only care about the joint distribution of a (subset of) $(O_{1:T}, H_{1:T})$, the analysis depends on $\mathbf{U}_t$ and $\mathbf{V}_t$ only through their known univariate marginals. In summary, the SCM is fully specified only when the dependence structure or copula is defined for each $\mathbf{U}_t$ and $\mathbf{V}_t$. Finally, the emissions and the state transitions in the GHMM are time-homogeneous. It is, therefore, natural to also assume the copulas underlying $\mathbf{V}_t$ and $\mathbf{U}_t$ are time-homogeneous.

EC.3.2. Modeling the SCM via Response Distributions

While specifying the SCM in terms of the $\mathbf{U}_t$'s and $\mathbf{V}_t$'s via Figure EC.1 is conceptually useful, it is often more convenient to work with a direct but equivalent construction of the SCM. This alternative approach is well-established in the causal inference literature, where related concepts include *canonical partitions* and *principal strata* (see, e.g., Duarte et al. 2024, Zhang et al. 2022).

Consider, for example, the relationship between $\mathbf{V}_t := [V_{thx}]_{h,x}$ and $[O_{thx}]_{h,x}$. From (EC.2), we know that the joint distribution of $[O_{thx}]_{h,x}$ is entirely determined by the joint distribution of $\mathbf{V}_t$. However, since our primary interest lies in the joint distribution of $[O_{thx}]_{h,x}$ itself, it is more natural to model this distribution *directly*, rather than indirectly via $\mathbf{V}_t$. Indeed, infinitely many joint distributions of $\mathbf{V}_t$ can produce the same joint distribution of $[O_{thx}]_{h,x}$. This follows from Sklar's Theorem in copula theory (Sklar 1959).

Accordingly, in §4, we directly modeled the distributions of $[O_{hx}]_{h,x}$ and $[H_{hi}]_{h,i}$ using the $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ variables (response distributions). Specifically, we defined the pairwise marginals in (6) and the

full joint PMFs in (9). Nonetheless, the copula framework remains valuable in §7 where we leverage the independence and comonotonic copulas to evaluate the degree of sub-optimality introduced by our approximation for computational scalability.

Appendix EC.4: Using Copulas to Estimate the Probability of Necessity

Building on the SCM presented in §EC.3.1, we now describe the Monte Carlo simulation algorithms that we use in §6 and §7 for estimating the PN values for the independence (§EC.4.1) and comonotonic (§EC.4.2) SCMs. As stated in §EC.3.1, the SCM is fully specified only when the dependence structure or copula is defined for each $\mathbf{U}_t$ and $\mathbf{V}_t$. When we refer to the independence (comonotonic) SCM below and indeed in the main body of the paper, what we have in mind is that the dependence structure of each $\mathbf{U}_t$ and $\mathbf{V}_t$ is given by the independence (comonotonic) copula. Of course, it would be easy to obtain other SCMs by allowing the $\mathbf{U}_t$'s to have one dependence structure, e.g., the independence copula, and the $\mathbf{V}_t$'s to have a different dependence structure, e.g., the monotonic copula. Indeed, each $\mathbf{U}_t$ and $\mathbf{V}_t$ could have their own different copulas and we would still obtain a legitimate SCM. There are countless combinations and so we will focus on just two SCMs here: the independence SCM where each $\mathbf{U}_t$ and $\mathbf{V}_t$ has the independence copula, and the comonotonic SCM where each $\mathbf{U}_t$ and $\mathbf{V}_t$ has the comonotonic copula.

EC.4.1. Counterfactual Simulations Under the Independence SCM

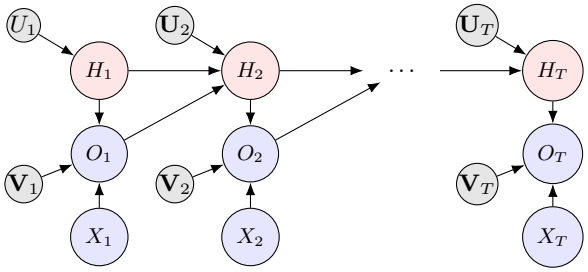


Figure EC.2 The same structural causal model as in Figure EC.1.

For convenience, we replicate Figure EC.1 from §EC.3.1, which now appears as Figure EC.2. Recall that $(o_{1:T}, x_{1:T})$ represents the observed data, and $\widetilde{x}_{1:T}$ denotes the intervention policy. As in §4.3, we begin with the posterior samples $[h_{1:T}(b)]_{b=1}^{B}$, corresponding to the random path $H_{1:T} \mid (o_{1:T}, x_{1:T})$. These samples can be generated efficiently, as established in §4.1. For each b, our goal is to convert the sampled path $h_{1:T}(b)$ into a counterfactual path, denoted by $\widetilde{h}_{1:T}(b)$. As noted in §4.2, the counterfactual hidden state in period 1 equals the posterior sample, i.e., $\widetilde{h}_1(b) = h_1(b)$. To sample $\widetilde{h}_2(b)$, we first need to sample the counterfactual emission $\widetilde{o}_1(b)$ (cf. Figure EC.2). With the copula underlying $\mathbf{V}_1$ being the independence copula, it follows that

$$\widetilde{o}_1(b) = \begin{cases} o_1 & \text{if } x_1 = \widetilde{x}_1 \text{ and } h_1(b) = \widetilde{h}_1(b) \\ \text{sample from the emission distribution } [e_{\widetilde{h}_1(b)\widetilde{x}_1 i}]_i & \text{otherwise.} \end{cases}$$

The counterfactual emission $\widetilde{o}_1(b)$ allows us to sample the counterfactual state $\widetilde{h}_2(b)$, which again leverages the fact that the copula underlying $\mathbf{U}_2$ is the independence copula:

$$\widetilde{h}_2(b) = \begin{cases} h_2(b) & \text{if } h_1(b) = \widetilde{h}_1(b) \text{ and } o_1 = \widetilde{o}_1(b) \\ \text{sample from the transition distribution } [q_{\widetilde{h}_1(b)\widetilde{o}_1(b)h'}]_{h'} & \text{otherwise.} \end{cases}$$

We then generate the period 2 counterfactual emission $\widetilde{o}_2(b)$ in a similar manner, and the process repeats iteratively until the end of the horizon T. The complete procedure is summarized in Algorithm 4.

Algorithm 4 Counterfactual simulations under the independence SCM

Require: $(\mathbf{E}, \mathbf{Q})$, $(o_{1:T}, x_{1:T})$, $[h_{1:T}(b)]_{b=1}^{B}$, $\widetilde{x}_{1:T}$

for $b = 1$ to B **do**
  $\widetilde{h}_1(b) = h_1(b)$
  for $t = 1$ to $T-1$ **do**
    if $x_t = \widetilde{x}_t$ and $h_t(b) = \widetilde{h}_t(b)$ **then**
      $\widetilde{o}_t(b) = o_t$
    else
      $\widetilde{o}_t(b) \sim \text{Categorical}([e_{\widetilde{h}_t(b)\widetilde{x}_t i}]_i)$
    end if
    if $h_t(b) = \widetilde{h}_t(b)$ and $o_t = \widetilde{o}_t(b)$ **then**
      $\widetilde{h}_{t+1}(b) = h_{t+1}(b)$
    else
      $\widetilde{h}_{t+1}(b) \sim \text{Categorical}([q_{\widetilde{h}_t(b)\widetilde{o}_t(b)h'}]_{h'})$
    end if
  end for
end for
return $[\widetilde{h}_{1:T}(b)]_b$

EC.4.2. Counterfactual Simulations Under the Comonotonic SCM

Before formally describing how to estimate PN using the comonotonic SCM, we provide some intuition for our approach via a simple example consistent with the setup of Example 1. Consider two variables, X and Y, where $X \in \{0, 1\}$ represents the medical treatment, and $Y \in \{\text{bad}, \text{better}, \text{best}\}$ represents the patient outcome. The outcome $Y_x := Y \mid (X = x)$ obeys the following distribution: $Y_0 \sim \{\text{bad}, \text{better}, \text{best}\}$ w.p. $\{0.2, 0.3, 0.5\}$ and $Y_1 \sim \{\text{bad}, \text{better}, \text{best}\}$ w.p. $\{0.2, 0.2, 0.6\}$. The underlying SCM is shown in Figure EC.3. Note that X is deterministic and does not have an exogenous noise variable pointing to it.

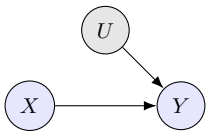


Figure EC.3 The comonotonic copula SCM for the "simple" example. The noise node is represented by the scalar $U \sim \mathrm{U}[0,1]$ rather than a vector $\mathbf{U}$. The structural equation is $Y = f(X,U)$, which we denote by $f_X(U)$, the *inverse transform* function corresponding to Y_x. That is, $f_0(u) =$ bad, better, and best if $u \in [0,0.2]$, $u \in [0.2,0.5]$, and $u \in [0.5,1]$, resp. Similarly, $f_1(u) =$ bad, better, and best if $u \in [0,0.2]$, $u \in [0.2,0.4]$, and $u \in [0.4,1]$, resp.

Consider a patient whose outcome Y was "better" under no treatment ($x=0$). Given the prior $U \sim \mathrm{U}[0,1]$, the posterior distribution is $U \mid (Y_0 = \text{better}) \sim \mathrm{U}[0.2,0.5]$. Now, suppose we are interested in the counterfactual outcome under the intervention $\widetilde{x}=1$, i.e., the random variable $\widetilde{Y} := Y_1 \mid (Y_0 = \text{better})$. Using the posterior $\mathrm{U}[0.2,0.5]$ for U and the functional form of $f_1(\cdot)$ (as defined in the caption of Figure EC.3), we observe that the interval $[0.2,0.4]$ maps to "better" and $[0.4,0.5]$ maps to "best". Consequently, $\widetilde{Y}$ is "better" w.p. 2/3 and "best" w.p. 1/3.

We now generalize this approach to the GHMM. We first fix an ordering of the states (set $\mathbb{H}$) and the emissions (set $\mathbb{O}$), e.g., from "best" to "worst". Denote by $r_H(h)$ the rank of state h with respect to this ordering and by $r_O(i)$ the rank of emission i. Furthermore, let $r_H^{-1}(r)$ and $r_O^{-1}(r)$ denote the inverse functions corresponding to $r_H(h)$ and $r_O(i)$, respectively. That is, $r_H^{-1}(r)$ returns the state with rank r and $r_O^{-1}(r)$ returns the emission with rank r. Also, for each (h,i) pair, observe that $[q_{hih'}]_{h'}$ denotes the transition distribution (which maps to the random variable H_{hi}). Corresponding to this distribution, we define the rank-ordered CDF as follows:

$$Q_{hih'} := \sum_{h'':r_H(h'')\le r_H(h')} q_{hih''} \ \forall h'. \tag{EC.3a}$$

Similarly, for each (h,x) pair, observe that $[e_{hxi}]_i$ denotes the emission distribution (which maps to the random variable O_{hx}). Corresponding to this distribution, we define the rank-ordered CDF as follows:

$$E_{hxi} := \sum_{j:r_O(j)\le r_O(i)} e_{hxj} \ \forall i. \tag{EC.3b}$$

Also, define $Q_{hi0} = E_{hx0} = 0$ for all (h,i) and (h,x). We discuss these orderings for the breast cancer application in §EC.5.4.

As in §EC.4.1, we begin with the samples $[h_{1:T}(b)]_{b=1}^B$, which we generate from the posterior distribution of $H_{1:T} \mid (o_{1:T}, x_{1:T})$. For each sample path $h_{1:T}(b)$, our goal is to generate a counterfactual path $\widetilde{h}_{1:T}(b)$. As noted in §4.2, irrespective of the copula choice, the counterfactual hidden state in period 1 equals the posterior sample, i.e., $\widetilde{h}_1(b) = h_1(b)$. To generate $\widetilde{o}_1(b)$, consider the SCM in Figure EC.4. The exogenous noise nodes are now scalar $\mathrm{U}[0,1]$ random variables rather than

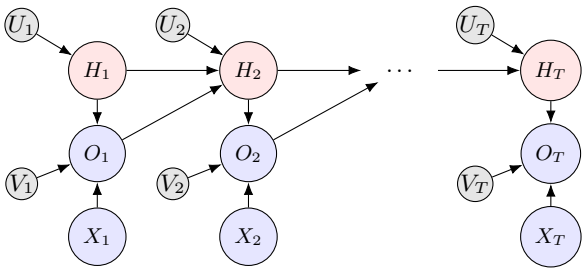


Figure EC.4 The comonotonic SCM underlying the GHMM. The key observation is that the noise nodes are now scalar U[0, 1] random variables as opposed to random vectors, i.e., (U_t, V_t) as opposed to $(\mathbf{U}_t, \mathbf{V}_t)$.

random vectors with U[0, 1] marginals. This is a direct implication of the comonotonic copula (as discussed in §EC.3.1). It follows from the structural equation (EC.2a) that $\widetilde{o}_1(b)$ satisfies

$$\widetilde{o}_1(b) = f(\widetilde{h}_1(b), \widetilde{x}_1, V_1) = f_{\widetilde{h}_1(b)\widetilde{x}_1}(V_1), \tag{EC.4a}$$

where $f_{hx}(\cdot)$ is the inverse transform function corresponding to the rank-ordered CDF $[E_{hxi}]_i$ (recall (EC.3b)). Hence, all we need to sample $\widetilde{o}_1(b)$ is the posterior distribution of V_1, where the "posterior" corresponds to conditioning on $O_{th_1(b)x_1} = o_1$ where $O_{thx} := O_t \mid (H_t = h, X_t = x)$ was defined in §EC.3.1. (We must keep the t notation here since the posterior dynamics are time-dependent.) Given the prior $V_1 \sim \mathrm{U}[0,1]$, the posterior of V_1 satisfies

$$V_1 \mid (O_{1h_1(b)x_1} = o_1) \sim \mathrm{U}[E_{h_1(b)x_1o_1^-}, E_{h_1(b)x_1o_1}], \tag{EC.4b}$$

where $o^- := r_O^{-1}(r_O(o) - 1)$ is the emission ranked just below o. Hence, we can efficiently sample V_1 from its posterior, and this V_1 sample can be used to generate $\widetilde{o}_1(b)$ (via (EC.4a)).

We can sample $\widetilde{h}_2(b)$ similarly. By the structural equation (EC.2b), $\widetilde{h}_2(b)$ satisfies

$$\widetilde{h}_2(b) = g(\widetilde{h}_1(b), \widetilde{o}_1(b), U_2) = g_{\widetilde{h}_1(b)\widetilde{o}_1(b)}(U_2), \tag{EC.5a}$$

where $g_{hi}(\cdot)$ is the inverse transform function corresponding to the rank-ordered CDF $[Q_{hih'}]_{h'}$ (recall (EC.3a)). Hence, all we need to sample $\widetilde{h}_2(b)$ is the posterior distribution of U_2, where the "posterior" corresponds to conditioning on $H_{2h_1(b)o_1} = h_2(b)$ where $H_{thi} := H_t \mid (H_{t-1} = h, O_{t-1} = i)$ was defined in §EC.3.1. (Once again, we must keep the t notation here since the posterior dynamics are time-dependent.) Given the prior $U_2 \sim \mathrm{U}[0,1]$, its posterior satisfies

$$U_2 \mid (H_{2h_1(b)o_1} = h_2(b)) \sim \mathrm{U}[Q_{h_1(b)o_1h_2(b)^-}, Q_{h_1(b)o_1h_2(b)}], \tag{EC.5b}$$

where $h^- := r_H^{-1}(r_H(h) - 1)$ is the state ranked just below h. Hence, we can efficiently sample U_2 from its posterior, and this U_2 sample can be used to generate $\widetilde{h}_2(b)$ (via (EC.5a)).

We can now generate the period 2 counterfactual emission $\widetilde{o}_2(b)$ in a similar manner and the process repeats until we reach the end of horizon T. We summarize the procedure in Algorithm 5.

Algorithm 5 Counterfactual simulations under the comonotonic SCM

Require: $(\mathbf{E}, \mathbf{Q})$, $(o_{1:T}, x_{1:T})$, $[h_{1:T}(b)]_{b=1}^{B}$, $\widetilde{x}_{1:T}$, $r_H(\cdot)$, $r_O(\cdot)$

for $b = 1$ to B **do**
  $\widetilde{h}_1(b) = h_1(b)$
  for $t = 1$ to $T-1$ **do**
    $V_t \sim \mathrm{U}[E_{h_t(b)x_t o_t^-}, E_{h_t(b)x_t o_t}]$ % posterior sample of V_t (see (EC.4b))
    $\widetilde{o}_t(b) = f_{\widetilde{h}_t(b)\widetilde{x}_t}(V_t)$ % counterfactual emission (see (EC.4a))
    $u_{t+1} \sim \mathrm{U}[Q_{h_t(b)o_t h_{t+1}(b)^-}, Q_{h_t(b)o_t h_{t+1}(b)}]$ % posterior sample of U_{t+1} (see (EC.5b))
    $\widetilde{h}_{t+1}(b) = g_{\widetilde{h}_t(b)\widetilde{o}_t(b)}(u_{t+1})$ % counterfactual state (see (EC.5a))
  end for
end for
return $[\widetilde{h}_{1:T}(b)]_b$

Remark EC.1. It should be intuitively clear that the comonotonic copula implies pathwise monotonicity. Our discussion of the simple example at the beginning of this sub-appendix further illustrates this point. In that example, a patient's outcome Y under no treatment ($x = 0$) was "better". The counterfactual outcome $\widetilde{Y}$ for this patient under treatment ($\widetilde{x} = 1$) in the SCM with the comonotonic copula was either "better" (w.p. 2/3) or "best" (w.p. 1/3), thereby satisfying pathwise monotonicity.

Appendix EC.5: Further Details on the Case Study

We discuss further details on the breast cancer model primitives and their calibration in §EC.5.1, followed by showing how we exploit sparsity to reduce the number of decision variables (§EC.5.2). We then provide details on the pathwise monotonicity constraints and the comonotonic copula in §EC.5.3 and §EC.5.4, respectively.

EC.5.1. Calibration of Initial State Distribution p and Emission Distribution E

We have $\mathbf{p} := (p_1, \ldots, p_7)$. Typically, breast cancer screening begins around age 40, with a prevalence among females aged 40-49 of 1.0183% (Table 4.24 of NIH (2020), all races, females): $p_2 + p_3 = 0.010183$. In-situ cancer accounts for 20% of new breast cancer diagnoses (Sprague and Trentham-Dietz 2009), giving $p_2 = 0.2 \times 0.010183$ and $p_3 = 0.8 \times 0.010183$. It is natural to set $p_4 = p_5 = p_6 = p_7 = 0$, leading to $p_1 = 1 - p_2 - p_3 = 1 - 0.010183$.

We have $\mathbf{E} := [e_{hxi}]_{h,x,i}$, where $e_{hxi} := \mathbb{P}(O_t = i \mid H_t = h, X_t = x)$. Before discussing the calibration, we highlight the sparse structure of $\mathbf{E}$. To illustrate this, we define the matrix $\mathbf{E}(x) := [e_{hxi}]_{hi}$ for each action x (with each row summing to 1), and observe the following structure:

$$\mathbf{E}(0) = \begin{bmatrix} 0 & \mathsf{e}_{12} & 1-\mathsf{e}_{12} & & & & & \\ 0 & 1-\mathsf{e}_{24} & & \mathsf{e}_{24} & & & & \\ 0 & 1-\mathsf{e}_{35} & & & \mathsf{e}_{35} & & & \\ 0 & & & 1 & & & & \\ 0 & & & & 1 & & & \\ 0 & & & & & 1 & & \\ 0 & & & & & & 1 & \\ 0 & & & & & & & 1 \end{bmatrix} \quad \mathbf{E}(1) = \begin{bmatrix} 1 & 0 & 0 & & & & & \\ 1 & 0 & 0 & & & & & \\ 1 & 0 & 0 & & & & & \\ 0 & 0 & 0 & 1 & & & & \\ 0 & 0 & 0 & & 1 & & & \\ 0 & 0 & 0 & & & 1 & & \\ 0 & 0 & 0 & & & & 1 & \end{bmatrix}.$$

A few comments are in order. First, for $x = 1$ (no screening), the emission matrix $\mathbf{E}(1)$ is extremely sparse, with entries in $\{0, 1\}$. For example, when the hidden state is 1 (healthy), 2 (undiagnosed in-situ cancer), or 3 (undiagnosed invasive cancer), no signal is observed (emission equals 1) w.p. 1. When the hidden state is 4, 5, 6, or 7, the emission matches the hidden state w.p. 1.

Second, for $x = 0$ (screening), the emission matrix $\mathbf{E}(0)$ is also sparse. If the patient is healthy (row 1), the test result is negative (true negative) w.p. e_{12} and positive (false positive) w.p. $1 - \mathsf{e}_{12}$. When the patient has undiagnosed in-situ cancer (row 2), it is detected (true positive) w.p. e_{24} and missed (false negative) w.p. $1 - \mathsf{e}_{24}$. Similarly, e_{35} represents the probability of detecting invasive cancer (true positive) and has the same interpretation as e_{24}. As with $x = 1$, when the hidden state is 4, 5, 6, or 7, the emission matches the hidden state w.p. 1.

Having discussed the structure of $\mathbf{E}$, we now calibrate it to real data. There are three parameters: e_{12}, e_{24}, and e_{35}. While these parameters can be age-specific, we simplify by using average values. The parameter e_{12} represents the *specificity* of the screening (i.e., the probability of a true negative) and is calibrated using Table 3 of Ayer et al. (2012), giving $\mathsf{e}_{12} = 0.9$. The parameter e_{24} represents the in-situ *sensitivity* (i.e., the probability of a true positive for in-situ cancer) and is also calibrated using Table 3 of Ayer et al. (2012), giving $\mathsf{e}_{24} = 0.8$. Finally, e_{35} represents the *sensitivity* for invasive cancer (i.e., the probability of a true positive for invasive cancer). Following Ayer et al. (2012), we assume $\mathsf{e}_{35} = \mathsf{e}_{24} = 0.8$.

EC.5.2. Reducing the Number of Joint Decision Variables by Exploiting Sparsity

Recall from §4.3 the following setup. Let $k \equiv (h, x)$ and $m \equiv (h, i)$ so that $O_k \equiv O_{hx}$, $e_{ki} \equiv e_{hxi}$, $H_m \equiv H_{hi}$, and $q_{mh'} \equiv q_{hih'}$. We have $k \in [K]$ and $m \in [M]$, where $K := |\mathbb{H}||\mathbb{X}|$ and $M := |\mathbb{H}||\mathbb{O}|$. The K- and M-dimensional joint PMFs for all $i_1, \ldots, i_K \in \mathbb{O}$ and $h_1, \ldots, h_M \in \mathbb{H}$ are defined as

$$\theta_{1,\ldots,K}(i_1, \ldots, i_K) := \mathbb{P}(O_1 = i_1, \ldots, O_K = i_K) \tag{9a}$$

$$\pi_{1,\ldots,M}(h_1, \ldots, h_M) := \mathbb{P}(H_1 = h_1, \ldots, H_M = h_M). \tag{9b}$$

As discussed towards the end of §4.3, it follows from (9) that there are *at most* $|\mathbb{O}|^{|\mathbb{H}||\mathbb{X}|} + |\mathbb{H}|^{|\mathbb{H}||\mathbb{O}|}$ joint variables. We now show that these are merely upper bounds and that the sparsity inherent in the breast cancer application can be exploited to drastically reduce these numbers.

State h	Policy x	Range of O_{hx}	Range cardinality
1	0	$\{2,3\}$	2
1	1	$\{1\}$	1
2	0	$\{2,4\}$	2
2	1	$\{1\}$	1
3	0	$\{2,5\}$	2
3	1	$\{1\}$	1
4	0	$\{4\}$	1
4	1	$\{4\}$	1
5	0	$\{5\}$	1
5	1	$\{5\}$	1
6	0	$\{6\}$	1
6	1	$\{6\}$	1
7	0	$\{7\}$	1
7	1	$\{7\}$	1

Table EC.1 Range of the 14 random variables $[O_{hx}]_{h,x}$ corresponding to $\theta_{1,\ldots,K}$.

Consider the $\theta_{1,\ldots,K}$ decision variables, which represent the joint PMF of $[O_k]_k$, where $k \equiv (h,x)$. To reduce the number of variables, we first determine the valid (h,x) pairs, rather than naively considering all $(h,x) \in \mathbb{H} \times \mathbb{X}$. Recall that the state h is encoded as follows: (1) healthy, (2) undiagnosed in-situ cancer, (3) undiagnosed invasive cancer, (4) diagnosed in-situ cancer, (5) diagnosed invasive cancer, (6) recovery, and (7) death. Furthermore, $x = 0$ indicates a mammogram is performed, while $x = 1$ indicates it is not performed. It is straightforward to verify that all 14 combinations of $(h,x) \in \mathbb{H} \times \mathbb{X}$ are valid, meaning none of the corresponding decision variables can be removed. Turning to the observations, they are encoded as follows: (1) no screening, (2) negative screening result, (3) positive mammogram followed by a negative biopsy, (4) diagnosed in-situ cancer, (5) diagnosed invasive cancer, (6) recovery, and (7) death. Table EC.1 summarizes the range of all $7 \times 2 = 14$ random variables $[O_{hx}]_{h,x}$. Multiplying the cardinalities from the last column of Table EC.1 reveals that only eight $\theta_{1,\ldots,K}$ decision variables need to be considered, significantly fewer than the naive upper bound of $|\mathbb{O}|^{|\mathbb{H}||\mathbb{X}|} = 7^{14}$.

The same logic applies to the $\pi_{1,\ldots,M}$ decision variables, which represent the joint PMF of $[H_m]_m$, where $m \equiv (h,i)$. For the $\pi_{1,\ldots,M}$ decision variables, even the first step proves useful as some (h,i) pairs are invalid. For example, if $h = 1$ (healthy), then $i \notin \{4,5,6,7\}$, as these emissions correspond to diagnosed cancer, recovery, or death. This initial step reduces the number of $[H_{hi}]_{h,i}$ random variables from $|\mathbb{H}||\mathbb{O}| = 49$ to 13 valid pairs. The second step trims the range of each of these 13 random variables. Table EC.2 documents these reductions, showing how the number of $\pi_{1,\ldots,M}$ decision variables is reduced from the naive upper bound of $|\mathbb{H}|^{|\mathbb{H}||\mathbb{O}|} = 7^{49}$ to 15,552, which equals the product of the cardinalities presented in the last column of Table EC.2.

EC.5.3. Details on the Pathwise Monotonicity Constraints

Pathwise monotonicity can be enforced via linear constraints. We discussed this in §5 and now provide a comprehensive description of all pathwise monotonicity constraints embedded in our

State h	Observation i	Range of H_{hi}	Range cardinality
1	1	$\{1,2,3\}$	3
1	2	$\{1,2,3\}$	3
1	3	$\{1,2,3\}$	3
2	1	$\{2,3\}$	2
2	2	$\{2,3\}$	2
2	4	$\{4,6\}$	2
3	1	$\{3,7\}$	2
3	2	$\{3,7\}$	2
3	5	$\{5,6,7\}$	3
4	4	$\{4,6\}$	2
5	5	$\{5,6,7\}$	3
6	6	$\{6\}$	1
7	7	$\{7\}$	1

Table EC.2 Range of the 13 random variables $[H_{hi}]_{h,i}$ corresponding to $\pi_{1,\ldots,M}$. Only 13 (h,i) pairs are shown as the other 36 are not valid.

breast cancer numerics. To recap from §5, suppose the patient has "low severity" cancer (state h) in period t, the cancer has not been detected (observation i), and the patient's state remains in "low severity" in period $t+1$. In the counterfactual world, if the cancer is detected in period t (observation $\bar{i}$), pathwise monotonicity requires that the cancer cannot transition to a worse state than "low severity" in period $t+1$. Formally, this implies $\mathbb{P}(H_{h\bar{i}} = \bar{h}' \mid H_{hi} = h) = 0$ for all states $\bar{h}'$ worse than h. There are multiple such cases to consider. We enforce all pathwise monotonicity constraints by setting the corresponding $\pi_{\bar{h}\bar{i},hi}(\bar{h}', h')$ variables to 0, as $\pi_{\bar{h}\bar{i},hi}(\bar{h}', h') = \mathbb{P}(H_{hi} = h')\mathbb{P}(H_{\bar{h}\bar{i}} = \bar{h}' \mid H_{hi} = h')$.

Hence, to provide details on all pathwise monotonicity constraints we enforce, it suffices to enumerate the $(h, i, h', \bar{h}, \bar{i}, \bar{h}')$ combinations for which the $\pi_{\bar{h}\bar{i},hi}(\bar{h}', h')$ variables are set to 0. To achieve this, we iterate over each state $h \in \{1, \ldots, 7\}$. (Note that for pathwise monotonicity, there are no $(h, x, i, \bar{h}, \bar{x}, \bar{i})$ combinations for which we set the $\theta_{\bar{h}\bar{x},hx}(\bar{i}, i)$ variables equal to 0.)

Healthy ($h = 1$). Pathwise monotonicity is enforced for the following combinations:

- If (h, i, h') equals (healthy, whatever emission, healthy), then the counterfactual state $\bar{h}'$ can not be in-situ, invasive, or death if $\bar{h}$ is healthy. That is, $h = 1$, $i \in \mathbb{O}$, $h' = 1$, $\bar{h} = 1$, $\bar{i} \in \mathbb{O}$, and $\bar{h}' \in \{2, 3, 4, 5, 7\}$.
- If (h, i, h') equals (healthy, whatever emission, in-situ), then the counterfactual state $\bar{h}'$ can not be healthy, invasive, or death if $\bar{h}$ is healthy. That is, $h = 1$, $i \in \mathbb{O}$, $h' = 2$, $\bar{h} = 1$, $\bar{i} \in \mathbb{O}$, and $\bar{h}' \in \{1, 3, 5, 7\}$.
- If (h, i, h') equals (healthy, whatever emission, invasive), then the counterfactual state $\bar{h}'$ can not be healthy, in-situ, or death if $\bar{h}$ is healthy. That is, $h = 1$, $i \in \mathbb{O}$, $h' = 3$, $\bar{h} = 1$, $\bar{i} \in \mathbb{O}$, and $\bar{h}' \in \{1, 2, 4, 7\}$.

- If (h, i, h') equals (healthy, whatever emission, death), then the counterfactual state $\bar{h}'$ can not be healthy, in-situ, or invasive if $\bar{h}$ is healthy. That is, $h = 1$, $i \in \mathbb{O}$, $h' = 7$, $\bar{h} = 1$, $\bar{i} \in \mathbb{O}$, and $\bar{h}' \in \{1, 2, 3, 4, 5\}$.

Undiagnosed In-situ ($h = 2$). Pathwise monotonicity is enforced for the following combinations:

- If (h, i, h') equals (in-situ, undetected, in-situ), then the counterfactual state $\bar{h}'$ can not be invasive or death if $\bar{h}$ is healthy or in-situ. That is, $h = 2$, $i \in \{1, 2, 3\}$, $h' = 2$, $\bar{h} \in \{1, 2, 4\}$, $\bar{i} \in \mathbb{O}$, and $\bar{h}' \in \{3, 5, 7\}$.
- If (h, i, h') equals (in-situ, detected, in-situ), then the counterfactual state $\bar{h}'$ can not be invasive or death if $\bar{h}$ is in-situ and detected. That is, $h = 2$, $i = 4$, $h' = 4$, $\bar{h} \in \{2, 4\}$, $\bar{i} = 4$, and $\bar{h}' \in \{3, 5, 7\}$.
- If (h, i, h') equals (in-situ, undetected, invasive), then the counterfactual state $\bar{h}'$ can not be death if $\bar{h}$ is healthy or in-situ. That is, $h = 2$, $i \in \{1, 2, 3\}$, $h' = 3$, $\bar{h} \in \{1, 2, 4\}$, $\bar{i} \in \mathbb{O}$, and $\bar{h}' = 7$.
- If (h, i, h') equals (in-situ, detected, invasive), then the counterfactual state $\bar{h}'$ can not be death if $\bar{h}$ is in-situ and detected. That is, $h = 2$, $i = 4$, $h' = 5$, $\bar{h} \in \{2, 4\}$, $\bar{i} = 4$, and $\bar{h}' = 7$.
- If (h, i, h') equals (in-situ, detected, recovered), then the counterfactual state $\bar{h}'$ can not be in-situ, invasive, or death if $\bar{h}$ is in-situ and detected. That is, $h = 2$, $i = 4$, $h' = 6$, $\bar{h} \in \{2, 4\}$, $\bar{i} = 4$, and $\bar{h}' \in \{2, 3, 4, 5, 7\}$.
- If (h, i, h') equals (in-situ, undetected, death), then the counterfactual state $\bar{h}'$ can not be in-situ, invasive, or recovered if $\bar{h}$ is in-situ and undetected. That is, $h = 2$, $i \in \{1, 2, 3\}$, $h' = 7$, $\bar{h} = 2$, $\bar{i} \in \{1, 2, 3\}$, and $\bar{h}' \in \{2, 3, 4, 5, 6\}$.
- If (h, i, h') equals (in-situ, detected, death), then the counterfactual state $\bar{h}'$ can not be in-situ, invasive, or recovered if $\bar{h}$ is in-situ and detected. That is, $h = 2$, $i = 4$, $h' = 7$, $\bar{h} \in \{2, 4\}$, $\bar{i} = 4$, and $\bar{h}' \in \{2, 3, 4, 5, 6\}$.

Undiagnosed Invasive ($h = 3$). Pathwise monotonicity enforced for the following combinations:

- If (h, i, h') equals (invasive, undetected, invasive), then the counterfactual state $\bar{h}'$ can not be death if $\bar{h}$ is healthy, in-situ, or invasive. That is, $h = 3$, $i \in \{1, 2, 3\}$, $h' = 3$, $\bar{h} \in \{1, 2, 3, 4, 5\}$, $\bar{i} \in \mathbb{O}$, and $\bar{h}' = 7$.
- If (h, i, h') equals (invasive, detected, invasive), then the counterfactual state $\bar{h}'$ can not be death if $\bar{h}$ is invasive and detected. That is, $h = 3$, $i = 5$, $h' = 5$, $\bar{h} \in \{3, 5\}$, $\bar{i} = 5$, and $\bar{h}' = 7$.
- If (h, i, h') equals (invasive, detected, recovered), then the counterfactual state $\bar{h}'$ can not be invasive or death if $\bar{h}$ is invasive and detected. That is, $h = 3$, $i = 5$, $h' = 6$, $\bar{h} \in \{3, 5\}$, $\bar{i} = 5$, and $\bar{h}' \in \{3, 5, 7\}$.

- If (h,i,h') equals (invasive, undetected, death), then the counterfactual state $\bar{h}'$ can not be invasive or recovered if $\bar{h}$ is invasive and undetected. That is, $h=3$, $i\in\{1,2,3\}$, $h'=7$, $\bar{h}\in\{3,5\}$, $\bar{i}\in\{1,2,3\}$, and $\bar{h}'\in\{3,5,6\}$.
- If (h,i,h') equals (invasive, detected, death), then the counterfactual state $\bar{h}'$ can not be invasive or recovered if $\bar{h}$ is invasive and detected. That is, $h=3$, $i=5$, $h'=7$, $\bar{h}\in\{3,5\}$, $\bar{i}=5$, and $\bar{h}'\in\{3,5,6\}$.

Diagnosed In-situ ($h=4$). Pathwise monotonicity is enforced for the following combinations:

- If (h,i,h') equals (in-situ, detected, in-situ), then the counterfactual state $\bar{h}'$ can not be invasive, recovered, or death if $\bar{h}$ is in-situ and detected. That is, $h=4$, $i=4$, $h'=4$, $\bar{h}\in\{2,4\}$, $\bar{i}=4$, and $\bar{h}'\in\{5,6,7\}$.
- If (h,i,h') equals (in-situ, detected, invasive), then the counterfactual state $\bar{h}'$ can not be in-situ, recovered, or death if $\bar{h}$ is in-situ and detected. That is, $h=4$, $i=4$, $h'=5$, $\bar{h}\in\{2,4\}$, $\bar{i}=4$, and $\bar{h}'\in\{4,6,7\}$.
- If (h,i,h') equals (in-situ, detected, recovery), then the counterfactual state $\bar{h}'$ can not be in-situ, invasive, or death if $\bar{h}$ is in-situ and detected. That is, $h=4$, $i=4$, $h'=6$, $\bar{h}\in\{2,4\}$, $\bar{i}=4$, and $\bar{h}'\in\{4,5,7\}$.
- If (h,i,h') equals (in-situ, detected, death), then the counterfactual state $\bar{h}'$ can not be in-situ, invasive, or recovered if $\bar{h}$ is in-situ and detected. That is, $h=4$, $i=4$, $h'=7$, $\bar{h}\in\{2,4\}$, $\bar{i}=4$, and $\bar{h}'\in\{4,5,6\}$.

Diagnosed Invasive ($h=5$). Pathwise monotonicity is enforced for the following combinations:

- If (h,i,h') equals (invasive, detected, invasive), then the counterfactual state $\bar{h}'$ can not be recovered or death if $\bar{h}$ is invasive and detected. That is, $h=5$, $i=5$, $h'=5$, $\bar{h}\in\{3,5\}$, $\bar{i}=5$, and $\bar{h}'\in\{6,7\}$.
- If (h,i,h') equals (invasive, detected, recovery), then the counterfactual state $\bar{h}'$ can not be invasive or death if $\bar{h}$ is invasive and detected. That is, $h=5$, $i=5$, $h'=6$, $\bar{h}\in\{3,5\}$, $\bar{i}=5$, and $\bar{h}'\in\{5,7\}$.
- If (h,i,h') equals (invasive, detected, death), then the counterfactual state $\bar{h}'$ can not be invasive or recovery if $\bar{h}$ is invasive and detected. That is, $h=5$, $i=5$, $h'=7$, $\bar{h}\in\{3,5\}$, $\bar{i}=5$, and $\bar{h}'\in\{5,6\}$.

Recovery ($h=6$) and Death ($h=7$). Pathwise monotonicity enforced for no combinations.

EC.5.4. Details on the Comonotonic Copula

We discussed the counterfactual simulation under the comonotonic copula for a general GHMM in §EC.4.2. In this section, we connect that discussion to the breast cancer application. To do so, it

suffices to define the rank functions $r_H(\cdot)$ (for states) and $r_O(\cdot)$ (for emissions). For states, there are two possible orderings that seem "natural" (from "best" to "worst"):

- $(1, 6, 4, \mathbf{2}, \mathbf{5}, 3, 7)$
- $(1, 6, 4, \mathbf{5}, \mathbf{2}, 3, 7)$.

Recalling the $\mathbf{Q}(i)$ notation from §6, note that columns 2 and 5 are never "active" simultaneously in any row of $\mathbf{Q}(i)$ (for any i). Thus, the choice of ordering between the two options above is inconsequential, and we can select either. Suppose we choose the first ordering. This defines the rank function, e.g., $r_H(6) = 2$, meaning the rank of state 6 is 2. For the inverse function, $r_H^{-1}(2) = 6$.

Defining $r_O(\cdot)$ for the breast cancer setting is unclear, but fortunately, it does not matter. To see why, consider the following path of interest, which generalizes both Paths 1 and 2 from §6:

$$\underbrace{o_1, \ldots, o_{\tau_s - 1}}_{\in\{2,3\}}, \underbrace{o_{\tau_s}, \ldots, o_{\tau_e}}_{=1}, \underbrace{o_{\tau_e+1}, \ldots, o_{T-1}}_{\in\{4,5\}}, \underbrace{o_T}_{=7}.$$

For the first $\tau_s - 1$ periods, observe that the counterfactual emission $\widetilde{o}_t(b)$ equals the observed emission o_t for each b. This is because the intervention policy $\widetilde{x}_t$ equals the observed policy x_t for $t \leq \tau_s - 1$. Now, consider periods τ_s to τ_e, during which the screening was not done, i.e., $x_{\tau_s:\tau_e} = 1$. Hence, the corresponding emissions $o_t = 1$ w.p. 1 (see the matrix $\mathbf{E}(1)$ in §EC.5.1). This means that the emissions do not contain any information regarding the underlying noise variables $V_{\tau_s:\tau_e}$ (see Figure EC.4) and hence, their posterior equals their prior, which is U$[0, 1]$. As such, for $t \in \{\tau_s, \ldots, \tau_e\}$, we can sample $\widetilde{o}_t(b)$ using the categorical distribution over the probability vector $[e_{\widetilde{h}_t(b)\widetilde{x}_t i}]_i$. Note that we can use $\widetilde{o}_{\tau_s - 1}(b)$ to sample $\widetilde{h}_{\tau_s}(b)$, which we can then use to sample $\widetilde{o}_{\tau_s}(b)$, and so on (until we have sampled $\widetilde{h}_{\tau_e+1}(b)$). Now, consider $t = \tau_e + 1$. We know $o_t \in \{4, 5\}$:

- If $o_t = 4$, then $h_t(b) = 2$ and $\widetilde{h}_t(b) \in \{2, 4, 6\}$ (cf. pathwise monotonicity).
  - —If $\widetilde{h}_t(b) \in \{4, 6\}$, then $\widetilde{o}_t(b) = \widetilde{h}_t(b)$ (since rows 4 and 6 of $\mathbf{E}(0)$ have 1 on the diagonal).
  - —Else, if $\widetilde{h}_t(b) = 2$ $(= h_t(b))$, then $\widetilde{o}_t(b) = o_t = 4$.
- Else, if $o_t = 5$, then $h_t(b) = 3$ and $\widetilde{h}_t(b) \in \{3, 4, 5, 6\}$ (cf. pathwise monotonicity).
  - —If $\widetilde{h}_t(b) \in \{4, 5, 6\}$, then $\widetilde{o}_t(b) = \widetilde{h}_t(b)$ (since rows 4, 5, 6 of $\mathbf{E}(0)$ have 1 on the diagonal).
  - —If $\widetilde{h}_t(b) = 3$ $(= h_t(b))$, then $\widetilde{o}_t(b) = o_t = 5$.

Finally, for $t \geq \tau_e + 2$, we know $h_t(b) \geq 4$ and that the corresponding rows in $\mathbf{E}(0)$ are 0-1. Hence, the posterior of V_t equals the prior and we can sample $\widetilde{o}_t(b)$ using the categorical distribution over the probability vector $[e_{\widetilde{h}_t(b)\widetilde{x}_t i}]_i$. By construction, the comonotonic copula will obey pathwise monotonicity and hence, will ensure that in the counterfactual world, the patient does not die before period T.