

Introduction to Anticipatory Oncology: From Hidden Disease States to Personalized Medicine

Philip De Melo*, Marie St Rose and Mildred Jackson

Department of Nursing and Allied Health, Norfolk State University, United States of America.

Abstract

Cancer develops through continuous and complex biological processes, yet oncology primarily relies on measurements obtained during discrete clinical encounters. Imaging, pathology, biomarkers, molecular tests, and clinical observations provide only partial and sometimes conflicting evidence of the underlying disease process. Although conventional artificial intelligence can predict clinical outcomes from observed variables, it does not necessarily reconstruct the hidden biological state that produces them.

This paper introduces anticipatory oncology as a new paradigm centered on the reconstruction of latent, or hidden, cancer states. It presents latent state artificial intelligence (LAT-AI), as an integrative framework that combines clinical, imaging, pathological, biomarker, and molecular evidence to estimate a clinically interpretable representation of a patient's underlying disease state. By distinguishing observed measurements from the latent disease process, LAT-AI may: 1) Identify ambiguous or discordant evidence, 2) quantify uncertainty, 3) detect clinically meaningful changes, and 4) improve understanding of disease progression.

Anticipatory oncology shifts the central question from “What outcome is associated with the patient's observed data?” to “What hidden disease state most plausibly generated these observations, and how is that state changing over time?” The framework is intended to complement rather than replace biopsy, pathology, established diagnostic procedures, or clinical judgment. Its goal is to advance precision and personalized medicine by providing interpretable, uncertainty aware, and dynamically updated estimates of hidden cancer states. The paper also examines validation, fairness, data quality, and clinical governance as essential requirements for translating latent state reconstruction into safe and clinically useful decision support.

Keywords

Anticipatory Oncology, Breast Cancer, Latent state, Hidden progression of cancer, Personalized medicine.

Corresponding Author Information

Philip De Melo

Department of Nursing and Allied Health, Norfolk State University, United States of America, ORCID: 0009-0003-3607-137x

Received: September 02, 2026; **Accepted:** October 03, 2026; **Published:** October 14, 2026

Copyright: © 2026 ASRJS. This is an openaccess article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Citation: Philip De Melo, Marie St Rose, Mildred Jackson. Introduction to Anticipatory Oncology: From Hidden Disease States to Personalized Medicine. J Can Res Rep. 2026; 2(4):1-10.

Introduction

Cancer is a complex biological process that changes over time through interactions among tumor cells, the surrounding tissue, the immune system, treatment effects, genetic alterations, and the

characteristics of the individual patient. Despite this dynamic nature, oncology continues to depend largely on observations collected at separate clinical encounters. Imaging studies, biopsies, laboratory tests, biomarkers, molecular profiles, and clinical examinations

provide valuable information, but each capture only a limited part of the disease at a particular moment. The true biological condition of the cancer remains only partially observable.

This creates an important challenge. Clinicians can observe tumor size, biomarker concentrations, pathological features, treatment responses, and symptoms, but they cannot directly observe the complete internal disease state that produces these findings. Two patients may have similar observable measurements while having substantially different underlying biological conditions and future disease trajectories. Conversely, apparently conflicting measurements may represent different views of the same evolving process rather than simple measurement errors. The central problem is therefore not only to predict an outcome from available data, but also to reconstruct the hidden disease state responsible for the observed evidence.

Most conventional predictive models estimate the probability of a predefined outcome, such as disease progression, treatment response, recurrence, or mortality. These models can be clinically valuable, but their predictions are generally based on statistical associations between observed variables and known outcomes. They do not necessarily explain what is occurring within the underlying disease process. A patient may be assigned a high probability of progression without the model identifying the hidden biological state that generated this risk. This distinction is important because outcome prediction and disease state reconstruction answer different clinical questions.

Anticipatory oncology is proposed as a new paradigm for addressing this limitation. It seeks to reconstruct the most plausible current state of cancer from incomplete, heterogeneous, and sometimes conflicting observations. It then evaluates how that state may be changing as additional evidence becomes available. The word anticipatory emphasizes the effort to recognize clinically meaningful changes before they become fully apparent through conventional measurements or symptoms. The purpose is not merely to forecast a distant outcome, but to improve understanding of the disease as it develops. In addition, anticipatory oncology is intended to support clinicians rather than replace them. Its estimates must be interpreted alongside pathology, imaging, established diagnostic procedures, clinical expertise, and patient preferences. Before clinical implementation, LAT-AI models must undergo rigorous evaluation of discrimination, calibration, reliability, interpretability, and clinical usefulness. They must also be validated across independent populations and examined for differences in performance among racial, ethnic, socioeconomic, geographic, and clinically relevant groups.

Furthermore, anticipatory oncology represents a transition from predicting outcomes based only on visible measurements to reconstructing the hidden processes that generate those measurements. Its central question is not simply what will happen to the patient, but what underlying disease state most plausibly explains the available evidence and how that state is changing

over time. By combining latent state reconstruction, uncertainty quantification, continuous updating, clinical interpretation, and fairness assessment, this paradigm may provide a stronger foundation for precision and personalized cancer care.

Latent state artificial intelligence, referred to as LAT-AI, provides a methodological foundation for anticipatory oncology. A latent state is an underlying condition that cannot be measured directly but can be inferred from observable evidence. In oncology, this hidden state may represent the current degree of malignant activity, biological aggressiveness, treatment sensitivity, likelihood of progression, or another clinically meaningful characteristic of the disease. LAT-AI integrates information from multiple sources to estimate this state and quantify the uncertainty associated with that estimate.

This distinction between observations and latent states is essential. An imaging result is an observation rather than the disease itself. A biomarker concentration is also an observation influenced by the disease, the patient, the measurement process, and other biological factors. Pathology provides highly important evidence, but even pathological findings describe sampled tissue and must be interpreted within a broader clinical context. LAT-AI does not treat any isolated measurement as a complete representation of the patient. Instead, it evaluates how the available observations collectively support different interpretations of the hidden disease state.

Bayesian inference offers one possible mathematical framework for this reconstruction. It allows prior clinical knowledge to be combined with new patient evidence and enables the estimated disease state to be revised as additional data become available. The model may begin with an initial estimate based on demographic information, clinical history, imaging, pathology, biomarkers, and molecular findings. As new laboratory results, imaging studies, treatment responses, or symptoms are recorded, the estimate can be updated. The result is a dynamic representation of the disease rather than a prediction based on a single clinical snapshot.

Uncertainty is a central element of this framework. Clinical data are frequently incomplete, noisy, inconsistent, or collected at irregular intervals. A model that produces only a single probability without reporting the reliability of that estimate may create a false sense of confidence. LAT-AI should therefore provide both an estimate of the latent state and a measure of uncertainty. High uncertainty may indicate insufficient information, conflicting evidence, unusual patient characteristics, or a case that differs substantially from the data used to develop the model. In such situations, the appropriate result may be a recommendation for additional evaluation rather than a more confident prediction.

The reconstruction of hidden disease states may also improve the interpretation of discordant evidence. For example, pathology may confirm malignancy while imaging, biomarkers, and other clinical indicators provide only moderate support. The correct conclusion

remains that malignancy is present because the established diagnostic evidence must retain its clinical authority. However, the disagreement among the available information may reveal unusual disease presentation, incomplete data, measurement limitations, or a need for closer review. LAT-AI can help identify and explain such discordance without attempting to overrule biopsy or clinical judgment.

This framework has important implications for precision and personalized medicine. Precision medicine uses measurable biological and clinical characteristics to identify the interventions most likely to benefit particular groups of patients. Personalized medicine applies this evidence within the broader circumstances of an individual patient, including health status, treatment history, preferences, risks, and goals. Reconstruction of the hidden disease state can contribute to both approaches by providing a more integrated and dynamically updated understanding of the patient than any single observation can offer.

Fairness is especially important because latent state estimates may be affected by differences in access to care, frequency of testing, data completeness, diagnostic quality, and representation within development datasets. A model may appear technically accurate while producing greater uncertainty or lower reliability for populations that have historically been underserved. Equity assessment must therefore be incorporated into model design, validation, and interpretation rather than treated as a secondary analysis.

Literature Review

Cancer is a dynamic and heterogeneous disease shaped by continuous genomic, cellular, and microenvironmental evolution. The TRACERx program demonstrated that longitudinal genomic sampling could reveal clonal diversification, treatment-driven selection, and emerging resistance during disease progression [1,2]. Pan-cancer whole-genome analysis further established the complexity of cancer evolution by identifying recurrent driver alterations, structural variants, and mutational processes across multiple tumor types [3]. Genetic heterogeneity and tumors complicates diagnosis, prognosis, and treatment selection [4,5]. Phylogenetic analysis of circulating tumor DNA offers a minimally invasive approach for tracking this evolution and detecting residual or recurrent disease [6].

Artificial intelligence provides powerful methods for extracting clinically meaningful information from complex oncologic data. Deep neural networks have achieved expert-level performance in classifying skin lesions [7], demonstrated promising accuracy in breast cancer screening [8] and predicted microsatellite instability directly from histopathology images [9]. Machine-learning methods have also produced highly accurate breast cancer classifications from structured clinical data [10]. Nevertheless, models based on a single data source may overlook important biological relationships. Multimodal frameworks such as pathomic fusion combine histopathologic and genomic features to improve

diagnostic and prognostic modeling [11].

Most existing AI systems emphasize classification at a single time point, whereas cancer develops through a longitudinal process. Joint models of repeated clinical measurements and time-to-event outcomes provide a statistical foundation for linking changing biomarkers with risks of progression, recurrence, or death [12]. Digital twins extend this concept by creating patient-specific computational representations that can be updated as new observations become available and used to simulate alternative disease and treatment trajectories [13]. Evolutionary treatment models have shown that adapting therapy to tumor dynamics may prolong disease control compared with continuously applying maximum-tolerated treatment [14].

Building on these developments, anticipatory oncology proposes reconstructing a patient's hidden disease state from incomplete and potentially discordant clinical, imaging, pathologic, biomarker, and molecular evidence [15]. Rather than relying exclusively on visible tumor characteristics or static classifications, this framework seeks to estimate current disease severity, quantify uncertainty, and anticipate plausible future progression. Such an approach may help identify cases requiring additional diagnostic confirmation and support more personalized surveillance and treatment planning.

Clinical usefulness, however, depends on more than predictive discrimination. Poorly calibrated models may generate probabilities that do not correspond to patients' observed risks, making calibration essential for clinical decision-making [16]. Transparent reporting through transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) and TRIPOD+AI [17,18], systematic evaluation of bias through a prediction model risk of bias assessment tool (PROBAST) [19] and prospective assessment under consolidated standards of reporting trials-artificial intelligence (CONSORT-AI) and standard protocol items recommendations for interventional trials-artificial intelligence (SPIRIT-AI) [20,21] are therefore necessary. External validation, workflow integration, data quality, interpretability, and ongoing monitoring also remain critical considerations for clinical implementation [22-24].

Finally, AI systems may reproduce or amplify disparities when training data, outcome definitions, or healthcare utilization patterns differ across populations [25,26]. Fairness must therefore be evaluated across racial, socioeconomic, geographic, and clinically relevant groups rather than assumed from overall model performance [27]. The literature collectively supports a transition from static cancer classification toward longitudinal, multimodal, uncertainty-aware, and equity-centered modeling. Anticipatory oncology represents a logical next step in this transition by combining cancer evolution, Bayesian inference, and responsible AI to estimate hidden disease states and support earlier, more individualized clinical decisions [15].

While the earlier study [15] mathematically formulated LAT-AI,

the current paper provides a deeper examination of anticipatory oncology and its clinical and methodological foundations. In particular, it extends the framework to evaluate the quality, completeness, and concordance of breast cancer data and to estimate the probabilities of multiple plausible cancer trajectories. Rather than predicting a single inevitable outcome, the proposed approach quantifies uncertainty across alternative trajectories and updates their probabilities as new patient-specific evidence becomes available.

Clinical Significance

Quantifying uncertainty across alternative cancer trajectories has important clinical significance because patients with similar current findings may experience substantially different patterns of progression. Instead of producing a single deterministic forecast, LAT-AI assigns probabilities to several plausible outcomes, such as stable disease, slow progression, rapid progression, treatment response, and recurrence.

The degree of uncertainty surrounding these probabilities can help clinicians distinguish well-supported predictions from cases involving incomplete, inconsistent, or discordant evidence. A high probability of rapid progression may justify closer surveillance or additional diagnostic evaluation, whereas broadly distributed probabilities may indicate that the available evidence is insufficient to support a confident decision. By updating these probabilities as new imaging, pathology, biomarker, genomic, and treatment-response data become available, LAT-AI may support timely risk reassessment and personalized care planning while preserving clinical judgment as the basis for final decisions.

A Step to Personalized Medicine

Anticipatory oncology represents an important step toward personalized medicine because it shifts the focus from population-level predictions and static disease classifications to continuously updated, patient-specific assessments. Traditional prognostic models often assign patients to broad risk groups using information collected at a single time point. However, breast cancer progression is dynamic and may differ considerably among patients with similar diagnoses, tumor stages, or molecular subtypes. LAT-AI addresses this limitation by integrating each patient's clinical history, imaging findings, pathology, biomarkers, molecular characteristics, treatment exposure, and longitudinal responses to reconstruct the current hidden disease state.

This patient-specific state can then serve as the starting point for estimating the probabilities of several plausible cancer trajectories, including stable disease, treatment response, slow progression, rapid progression, recurrence, or metastatic spread. Importantly, LAT-AI also quantifies the uncertainty associated with these estimates. A narrowly concentrated probability distribution may indicate relatively strong agreement among the available data, whereas a broad distribution may reveal incomplete information, measurement variability, or discordance among diagnostic sources. Clinicians could use this distinction to determine when a prediction

is sufficiently reliable to inform care and when additional imaging, biomarker testing, pathology review, or molecular analysis may be necessary.

As new evidence becomes available, the estimated disease state and trajectory probabilities can be updated. This creates a learning process in which surveillance and treatment planning evolve with the patient rather than remaining fixed at diagnosis. By combining individualized risk estimation with explicit evaluation of uncertainty and data quality, anticipatory oncology may support more appropriate surveillance intervals, earlier recognition of unfavorable changes, and better-informed treatment decisions.

It does not replace biopsy, validated clinical guidelines, or professional judgment. Instead, it provides an important uncertainty-aware analytical framework that may significantly assist clinicians in selecting the most appropriate next step based on each patient's individual disease pattern.

Formulation of Anticipatory Oncology

In this section we follow [15] and formulate the anticipatory oncology. Anticipatory oncology can be formulated as a probabilistic latent state framework in which the true biological condition of a disease is not observed directly but is reconstructed from clinical, imaging, pathological, biomarker, and molecular evidence. The fundamental objective is to estimate the current hidden disease state, quantify uncertainty in that estimate, determine how the state changes over time, and calculate the probability of clinically meaningful progression.

Patient observations

Anticipatory oncology can be formulated as a probabilistic latent state framework in which the true biological condition of a cancer is not observed directly but is reconstructed from clinical, imaging, pathological, biomarker, and molecular evidence. The mathematical objective is to estimate the current hidden disease state, quantify uncertainty in that estimate, determine how the state changes over time, and calculate the probability of clinically meaningful progression.

For patient i , let measurements be collected at times

$$t_{i1}, t_{i2}, \dots, t_{iT_i}.$$

At time t , the observed data vector is

$$y_{i,t} = \begin{pmatrix} y_{i,t}^{(1)} \\ \vdots \\ y_{i,t}^{(p)} \end{pmatrix}$$

where p is the number of observed variables.

For breast cancer, the observation vector may contain variables such as tumor characteristics, lymph node status, histological grade, ER status, PR status, HER2 status, Ki-67, CA15-3, PET SUV, circulating tumor DNA, treatment response, etc.

Hidden cancer state

Let the unobserved cancer state be

$$x_{i,t} = \begin{pmatrix} x_{i,t}^{(1)} \\ \vdots \\ x_{i,t}^{(q)} \end{pmatrix}$$

A clinically interpretable hidden state could be defined, for example, as

$$x_{i,t} = (B_{i,t}, P_{i,t}, I_{i,t}, M_{i,t}, R_{i,t}, S_{i,t})$$

where

$B_{i,t}$ -latent tumor burden, $P_{i,t}$ -proliferative activity, $I_{i,t}$ -invasive potential, $M_{i,t}$ -metastatic potential, $R_{i,t}$ -treatment resistance, $S_{i,t}$ -treatment sensitivity. Because these quantities cannot usually be measured directly, the anticipatory algorithm will infer them from the available evidence.

State transition model

The evolution of the hidden cancer state is described by

$$x_{i,t} = f(x_{i,t-1}, u_{i,t}, c_i, \Delta t) + \omega_{i,t}$$

where

$u_{i,t}$ - treatment administered at time t , c_i -patient-specific characteristics (age, menopausal status, race/ethnicity, genetic risk score, comorbidity index, social determinants of health, etc.

$$\omega_{i,t} \sim N(0, Q_t)$$

represents biological variability and unexplained disease evolution. N denotes a normal, or Gaussian, probability distribution, while 0 is mean and Q_t is the covariance matrix. The model means that the patient's current hidden cancer state depends on the previous cancer state, treatment, relatively stable patient characteristics, elapsed time, and unexplained biological variability.

Cancer evolution is commonly nonlinear. A nonlinear tumor-burden equation can therefore be written as

$$B_{i,t} = B_{i,t-1} + \Delta t \left[r_i B_{i,t-1} \left(1 - \frac{B_{i,t-1}}{K_i} \right) - \beta_i u_{i,t} B_{i,t-1} \right] + \omega_{i,t}^{(B)}$$

where

r_i -intrinsic tumor growth rate, K_i -patient specific carrying capacity, β_i -treatment effectiveness, $u_{i,t}$ -treatment intensity. The patient specific carrying capacity is the theoretical maximum tumor burden that the biological environment of patient can sustain under the assumptions of the model.

The observation model connects the hidden state to measured clinical variables:

$$y_{i,t} = h(x_{i,t}) + v_{i,t}$$

Where

$$v_{i,t} \sim N(0, R_t)$$

represents measurement errors.

Ideally, we may write:

$$x_{i,t} = h^{-1}(y_{i,t} - v_{i,t})$$

A major problem in clinical practice, however, is that $h^{-1}h^{-1}\{-1\}$ may not exist, may not be unique, or may be unstable because the clinical observations are incomplete, noisy, or discordant. Therefore, the central challenge is to infer the hidden disease state, including the underlying level and progression of cancer, from the available observations. De Melo [15] provides a formal probabilistic solution to this inverse problem using a Kalman filter, which recursively estimates the latent disease state and updates this estimate as new clinical evidence becomes available.

In this paper, we will develop a generalized approach that does not have limitations of Kalman filter. The approach is based on the Bayesian inference that converts an ill-posed inverse problem into a probabilistic estimation problem. It does not require a unique analytical inverse of h . Instead, it combines prior knowledge, a disease-progression model, observed clinical evidence, and measurement uncertainty to identify a distribution of plausible hidden states. Consequently, the method identifies not only the most probable state of disease progression but also the uncertainty associated with that estimate.

Bayesian reconstruction of the hidden state

After receiving the new observation $y_{i,t}$, Bayes' theorem gives:

$$p(x_{i,t} | y_{i,1:t}) = \frac{p(y_{i,t} | x_{i,t})p(x_{i,t} | y_{i,1:t-1})}{p(y_{i,t} | y_{i,1:t-1})}$$

$p(x_{i,t} | y_{i,1:t})$ is posterior probability distribution of the hidden dynamic state at time t , given all patient observations from time 1 through time t . The equation reconstructs a sequence of hidden states rather than one permanently fixed state.

$p(x_{i,t} | y_{i,1:t-1})$ answers the following question: If the hidden disease-state history $x_{i,t}$ is known, how likely would the observed measurement $y_{i,t}$ would be?

$$p(y_{i,t} | x_{i,t}) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp \left[-\frac{(y_{i,t} - h(x_{i,t}))^2}{2\sigma_y^2} \right]$$

$p(x_{i,t} | y_{i,1:t-1})$ is the predictive prior distribution of the hidden state $x_{i,t}$ for patient i at time t , given all clinical observations available through the previous time point $t-1$.

$p(y_{i,t} | y_{i,1:t-1})$ is the predictive probability, or predictive density, of the new observation $y_{i,t}$ given all previous observations $y_{i,1:t-1}$.

To evaluate data completeness

$$m_{i,t,j} = \begin{cases} 1, & \text{if observation } y_{i,t,j} \text{ is available} \\ 0, & \text{if observation } y_{i,t,j} \text{ is missing} \end{cases}$$

A simple completeness score is

$$C_{i,t} = \frac{\sum_{j=1}^P w_j \cdot m_{i,t,j}}{\sum_{j=1}^P w_j}$$

where p is the number of expected variables and w_j represents the clinical importance of variable. Missing a critical biopsy result should have a greater effect than missing a nonessential laboratory value j .

After estimating the latent state $x_{i,p}$, the model predicts the expected observations:

$$\hat{y}_{i,t} = h(\hat{x}_{i,t})$$

The difference between the actual and predicted observations is the residual:

$$r_{i,t} = y_{i,t} - h(\hat{x}_{i,t})$$

Because clinical variables have different units and measurement errors, residuals should be standardized:

$$z_{i,t,j} = \frac{y_{i,t,j} - h_j(\hat{x}_{i,t})}{\sigma_j}$$

where σ_j represents the expected uncertainty of observation j . Values such as

$$|z_{i,t,j}| > 2$$

may indicate potentially discordant evidence.

The width of the posterior distribution reflects uncertainty about the hidden state. It can be summarized using posterior variance:

$$\hat{H}_{i,t} = - \sum_{k=1}^N p_{i,t,k} \log(p_{i,t,k})$$

Low entropy means that one state is strongly supported. High entropy means that several states remain plausible.

Latent State and Data Quality

We used eight types of clinical evidence at six patient’s visits: months 0, 3, 6, 9, 12, and 15.

Clinical variable	Month 0	Month 3	Month 6	Month 9	Month 12	Month 15
Tumor size (cm)	1.1	1.4	2.0	2.9	2.5	2.2
Ki-67 (%)	14	18	27	41	34	29
CA15-3 (U/mL)	20	24	37	61	48	39
PET SUV	2.7	3.4	5.2	8.1	6.4	5.6
ctDNA (copies)	3	Missing	18	55	Missing	24
Positive lymph nodes	0	0	1	2	1	1
Pathology grade	1	1	2	1	2	2
Symptom score	0	1	1	2	Missing	

Figure 1 shows the Bayesian reconstruction of the latent disease state over 15 months. The solid blue line with circular markers represents the posterior mean estimate of latent disease activity, while the light-blue shaded region shows the approximate 95% posterior credible interval and therefore quantifies uncertainty in the reconstructed state. The orange dashed line with square markers represents the maximum a posteriori (MAP) estimate, defined as the most probable latent-state value at each assessment. The black dotted line indicates the synthetic generating state, which serves as the known reference trajectory for evaluating reconstruction accuracy. Disease activity increases from a relatively low level at baseline to a peak at approximately 9 months, remains elevated at 12 months, and subsequently declines by month 15.

The posterior mean and MAP estimates generally follow the reference trajectory, suggesting that the Bayesian model successfully recovers the underlying disease dynamics from noisy, incomplete, or potentially discordant observations. The wider posterior interval at baseline and month 12 indicates greater uncertainty at these assessments, whereas the narrower intervals at months 6, 9, and 15 indicate stronger agreement among the available evidence and greater confidence in the reconstructed state.

The divergence between the posterior mean, MAP estimate, and synthetic generating state at month 12 illustrates that the most likely disease trajectory may differ from the underlying state when clinical evidence is limited, noisy, or discordant. Latent disease activity is expressed on a normalized scale from 0 to 1, with higher values representing greater modeled disease burden or biological activity. These values are analytical estimates rather than directly observed clinical measurements and should be interpreted alongside clinical findings, imaging, pathology, biomarkers, and professional judgment.

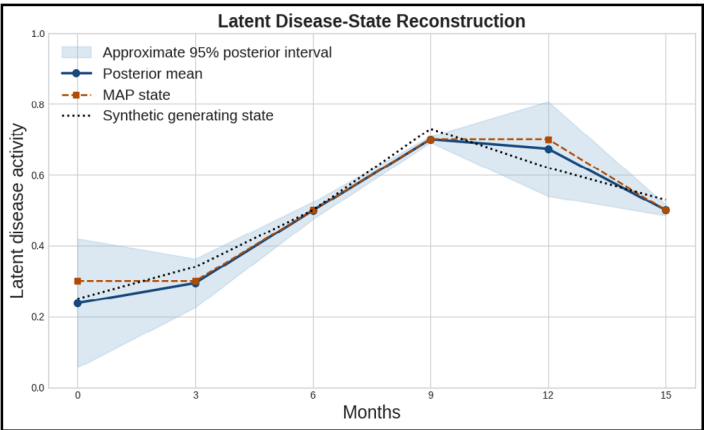


Figure 1: Bayesian reconstruction of the latent disease state over 15 months. The posterior uncertainty is greatest at baseline, decreases as evidence accumulates through month 9, increases again between months 9 and 12 because of incomplete or discordant evidence, and then decreases by month 15.

Figure 2 shows the posterior probability distribution across candidate latent disease states during 15 months of follow-up. The heatmap presents the posterior probability assigned to five candidate levels of latent disease activity: 0.1, very low; 0.3, low; 0.5, intermediate; 0.7, high; and 0.9, very high. Columns correspond to follow-up assessments at 0, 3, 6, 9, 12, and 15 months. Each cell reports the probability that the patient’s underlying disease state belongs to the corresponding category after integrating all clinical evidence available through that visit. Darker blue shading indicates a higher posterior probability.

At baseline, the model assigns a probability of 0.69 to the low state and 0.31 to the very-low state, indicating moderate uncertainty between two relatively mild disease states. At 3 months, the posterior probability of the low state increases to 0.97, providing

substantially greater confidence in that classification. At 6 months, the probability shifts completely to the intermediate state. At 9 months, the high state receives a posterior probability of 1.00, indicating the strongest modeled evidence for elevated disease activity. At 12 months, the high state remains dominant with a probability of 0.87, although the intermediate state retains a probability of 0.13. This residual probability demonstrates that the model preserves uncertainty when the available observations do not fully distinguish between adjacent disease states. By month 15, the posterior distribution shifts to the intermediate state with a probability of 1.00, suggesting a reduction in modeled disease activity relative to months 9 and 12.

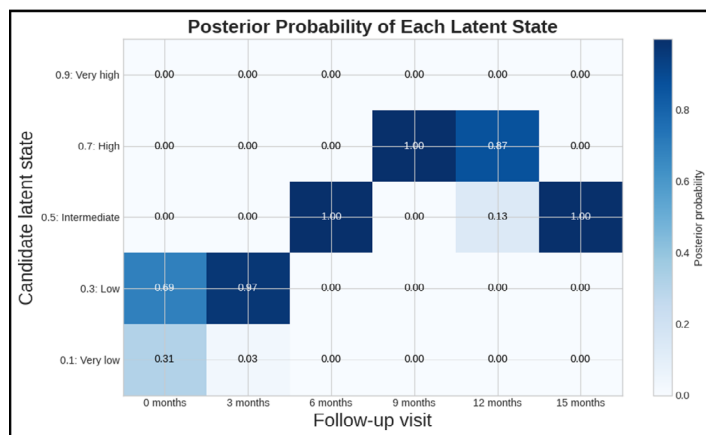


Figure 2: The posterior probability distribution across candidate latent disease states during 15 months of follow-up. After considering all evidence collected through month 12, the model assigns an 87% posterior probability to the high latent disease-activity state represented by 0.7.

At every visit, the probabilities across all candidate states sum to 1.00. The state with the highest probability in each column is the maximum a posteriori, or MAP, estimate for that visit. The complete distribution is more informative than the MAP value alone because it shows both the most probable state and the degree of uncertainty surrounding that classification. In a clinical application, changes in these probabilities would result from sequentially incorporating new evidence, such as tumor measurements, imaging findings, lymph-node status, histological characteristics, molecular markers, circulating biomarkers, and treatment response. The latent-state values are normalized model-based representations of disease activity rather than directly observed clinical measurements. Therefore, the heatmap illustrates probabilistic disease-state reconstruction and should not be interpreted as an independent diagnosis or a substitute for biopsy, validated clinical guidelines, or professional judgment.

Figure 3 shows the longitudinal assessment of data completeness, discordance, and uncertainty across six follow-up visits. The green line with circular markers represents weighted data completeness, the purple line with square markers represents normalized posterior uncertainty, and the red line with triangular markers represents scaled discordance among the available clinical observations. All three measures are displayed on a normalized scale from 0 to 1 to

facilitate comparison, although they describe different properties of the evidence and should not be interpreted as clinically equivalent quantities.

Weighted completeness measures the proportion of expected observations available at each visit while accounting for their relative clinical importance. A value of 1.0 indicates that all required observations are available, whereas lower values indicate that one or more measurements are missing. Completeness is 1.0 at baseline and at months 6, 9, and 15, but decreases to approximately 0.85 at month 3 and 0.79 at month 12. These reductions suggest that some expected observations were unavailable at those assessments. Because completeness is weighted, omission of a clinically important variable may have a greater effect on the score than omission of a less informative measurement.

Normalized uncertainty summarizes the dispersion of the posterior distribution of the latent disease state. Higher values indicate that the available evidence supports several plausible states, whereas values near zero indicate that the posterior probability is highly concentrated around a particular state. Uncertainty is highest at baseline, approximately 0.39, which may reflect limited initial information or competing plausible disease states. It declines substantially by month 3, approaches zero at months 6 and 9, rises to approximately 0.24 at month 12, and returns to nearly zero at month 15. The increase at month 12 is consistent with reduced data completeness and a less concentrated posterior distribution.

Scaled discordance measures disagreement among observations that are expected to provide related information about the underlying disease condition. Discordance remains relatively low at baseline and month 3, increases modestly at month 6, and reaches its maximum value of 1.0 at month 9. The vertical dotted line and annotation identify an imaging–pathology conflict at month 9. For example, imaging may suggest increasing disease activity while pathology indicates a less aggressive or treatment-responsive state. Importantly, completeness at month 9 is also 1.0, demonstrating that a dataset can be complete yet internally contradictory. The posterior uncertainty remains close to zero at this visit because uncertainty and discordance are related but distinct concepts: a model can produce a concentrated posterior even when individual evidence sources disagree, particularly if one observation is assigned greater reliability or weight than another.

At month 12, discordance decreases substantially but uncertainty rises while completeness falls. This pattern suggests that uncertainty at month 12 is driven more by missing or less informative evidence than by a major conflict among available measurements. By month 15, completeness returns to 1.0 and uncertainty approaches zero, although a small degree of residual discordance remains. Collectively, the figure illustrates that completeness, discordance, and posterior uncertainty provide complementary information. Completeness describes whether the expected evidence is available, discordance describes whether the available observations agree, and posterior uncertainty describes how confidently the latent disease state can be reconstructed. None of these measures should

be interpreted independently as a diagnosis; rather, they can identify visits requiring data verification, repeat testing, additional clinical evaluation, or multidisciplinary review.

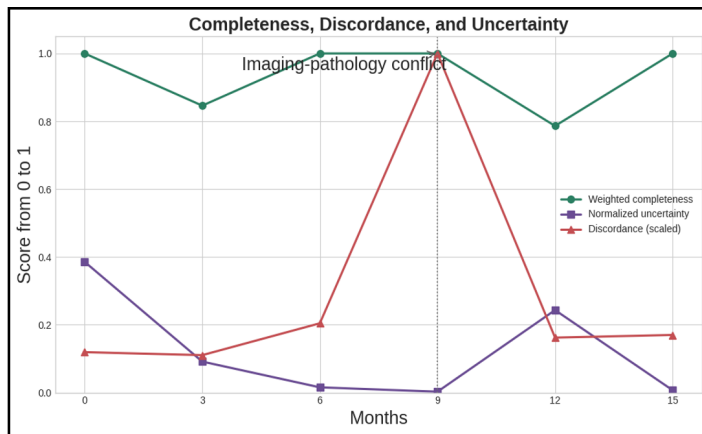


Figure 3: Longitudinal assessment of data completeness, discordance, and uncertainty across six follow-up visits. Month 9 shows high discordance but very low posterior uncertainty. This can occur when imaging and pathology disagree, but the model gives much greater reliability or weight to one source. The posterior may then concentrate strongly on one state despite the conflict. This is mathematically possible, although it should trigger a clinical data-quality flag because a highly confident estimate is not necessarily trustworthy when important evidence sources disagree.

Resolving Latent-State Uncertainty at Month 12

At month 12, our method indicates a posterior probability of 0.87 to the high latent state $x=0.7$ and 0.13 to the intermediate state $x=0.5$, indicating residual uncertainty between two plausible disease-severity levels. The addition of mammographic evidence, such as the suspicious mass indicated by the arrow, provides new information about lesion morphology, density, margins, and surrounding tissue distortion. When incorporated into the Bayesian observation model, this evidence updates the likelihood of each candidate's latent state. If the imaging characteristics are more consistent with the high-severity state, probability shifts from the intermediate state toward the high state, producing a narrower posterior distribution and lower uncertainty. Conversely, benign-appearing findings may shift probability toward the intermediate state. Thus, additional imaging mitigates uncertainty by improving discrimination between competing latent states; it does not necessarily increase estimated severity and does not replace biopsy or clinical judgment.

Before adding the mammogram at month 12:

$$P(x) = \begin{cases} 0.87, & \text{if } x = 0.7 \\ 0.13, & \text{if } x = 0.5 \end{cases}$$

Radiomic interpretation finds that the suspicious mass at month 12 is:

$$P(x) = \begin{cases} 0.9, & \text{if } x = 0.7 \\ 0.3, & \text{if } x = 0.5 \end{cases}$$

The unnormalized probabilities become:

$$P(\text{High}) = 0.90 \times 0.87 = 0.783$$

$$P(\text{Intermediate}) = 0.30 \times 0.13 = 0.039$$

After normalization

$$P_n(\text{high}) = \frac{0.783}{0.783 + 0.039} = 0.953$$

$$P_n(\text{Intermediate}) = \frac{0.039}{0.783 + 0.039} = 0.047$$

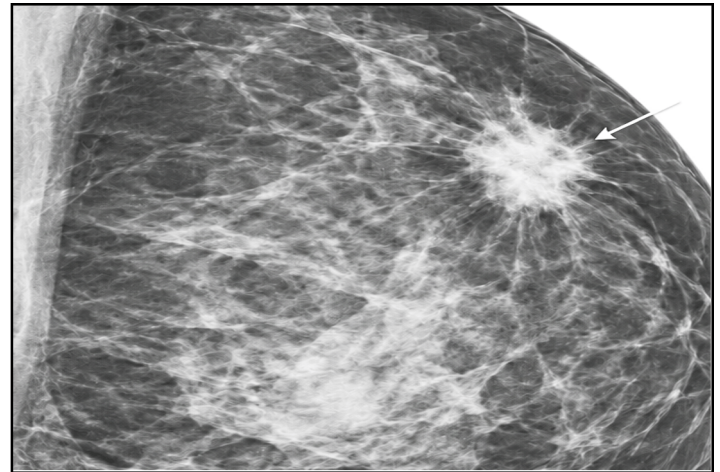


Figure 4: Mammographic evidence added at month 12 to update posterior probabilities and reduce uncertainty between competing latent disease states.

Therefore, the mammogram changes the probabilities from: $(0.87, 0.13) \rightarrow (0.953, 0.047)$. The high latent state remains the most likely state, but its probability increases from 87% to approximately 95%. Uncertainty decreases because the new mammographic evidence more strongly supports the high state than the intermediate state.

The proposed framework assists clinicians by integrating heterogeneous longitudinal evidence, quantifying diagnostic uncertainty, identifying discordant or incomplete observations, and indicating when additional testing may be most informative. It does not replace biopsy, validated clinical guidelines, or professional judgment; rather, it supports the selection of the most appropriate next diagnostic or treatment-planning step for an individual patient.

Discussion

This study presents a novel uncertainty-aware framework for reconstructing latent breast cancer states from heterogeneous longitudinal evidence. The principal contribution is not merely the identification of a single disease-severity value, but the estimation of a posterior probability distribution across multiple plausible latent states. This distinction is clinically important because imaging, pathology, biomarkers, molecular findings, and treatment-response measures may be incomplete, noisy, or discordant. A conventional classification model may conceal this uncertainty by producing only one predicted category, whereas the proposed Anticipatory Oncology framework explicitly shows both the leading interpretation and the remaining alternatives.

In this study, the month-12 result demonstrates how the framework can identify a point at which additional information may be clinically valuable. A clinician-interpreted mammogram can contribute information about lesion size, shape, density, margins, calcifications, architectural distortion, and changes from previous examinations. These structured findings are incorporated into the proposed model through their likelihood under each candidate latent state.

The posterior distribution becomes more concentrated, indicating reduced uncertainty. However, the additional evidence does not necessarily increase the estimated disease severity. If the mammographic findings were more compatible with the intermediate state, probability would shift in the opposite direction. Thus, additional evidence mitigates uncertainty by improving discrimination between competing states, rather than by automatically confirming the more severe state.

The proposed framework could be an important new tool for diagnostic decision-support and data-quality-control system. It may significantly help clinicians integrate evidence collected at different times, determine whether findings are mutually concordant, identify missing or contradictory observations, and recognize clinical visits at which uncertainty remains elevated. Rather than providing an isolated prediction, the model offers a longitudinal representation of how the inferred disease state and its uncertainty evolve as new evidence becomes available.

This information may support selection of the most appropriate next diagnostic or treatment-planning step for an individual patient. Depending on the clinical context, the next step may include additional imaging, biopsy, pathology review, molecular testing, closer surveillance, multidisciplinary consultation, or reconsideration of an existing treatment plan. The framework may also help prioritize cases in which additional testing is likely to provide the greatest reduction in uncertainty.

An important advantage of the framework is its potential to distinguish between uncertainty caused by missing data and uncertainty caused by discordant data. Missingness occurs when a clinically relevant observation was not collected or is unavailable. Discordance occurs when available observations support different latent-state interpretations. For example, suspicious imaging findings accompanied by stable biomarkers may produce a broader posterior distribution because the two sources of evidence are not fully aligned.

The longitudinal character of the framework is consistent with personalized and precision oncology. The prior distribution can incorporate patient-specific characteristics and previous disease history, while each subsequent observation updates the posterior state. Consequently, two patients with similar current measurements may receive different posterior estimates because their prior trajectories, tumor characteristics, treatments, and patterns of evidence differ.

This adaptive updating is particularly useful in oncology because the disease state is not static. Tumor progression, therapeutic response, resistance, recurrence, and measurement uncertainty may change over time. A sequence of posterior distributions may therefore provide a more clinically informative representation than a single cross-sectional classification. It can show not only the most probable current state but also when the estimated state changed and how confidently that change was inferred.

Conclusion

The proposed latent-state framework provides an efficient method for integrating heterogeneous longitudinal evidence while preserving uncertainty. Its central clinical value lies in showing what disease state is most strongly supported, what alternative states remain plausible, and what additional evidence may help distinguish among them. The framework does not replace biopsy, radiological interpretation, validated clinical guidelines, or professional judgment. Instead, it may assist clinicians in selecting the most appropriate next diagnostic or treatment-planning step while making the degree and sources of uncertainty explicit.

Several limitations should be acknowledged. Model performance should be evaluated across racial, ethnic, geographic, socioeconomic, and clinically relevant subgroups. Differences in data availability, diagnostic pathways, imaging quality, and access to follow-up care could produce systematic disparities in model accuracy and uncertainty. Fairness assessment should therefore examine not only predictive performance but also data completeness, calibration, diagnostic concordance, and the frequency with which the model recommends additional clinical review. These issues, together with external validation and evaluation in real-world clinical settings, will be the focus of our future research.

Future work should validate the framework using longitudinal, clinically adjudicated breast cancer data. Estimated states should be compared with biopsy findings, pathological characteristics, treatment response, recurrence, progression, and other relevant outcomes. Validation should examine discrimination, calibration, posterior coverage, robustness to missing data, and clinical usefulness. External validation across institutions and imaging systems will also be necessary.

References

1. Jamal-Hanjani M, Hackshaw A, Ngai Y, Shaw J, Dive C, et al. Tracking genomic cancer evolution for precision medicine: the lung TRACERx study. *PLoS Biol.* 2014; 12: e1001906.
2. Bailey C, Black JRM, Reading JL, Litchfield K, Turajlic S, et al. Tracking cancer evolution through the disease course. *Cancer Discov.* 2021; 11: 916-932.
3. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature.* 2020; 578: 82-93.
4. Hanahan D. Hallmarks of cancer: new dimensions. *Cancer Discov.* 2022; 12: 31-46.
5. Turajlic S, Sottoriva A, Graham T, Swanton C. Resolving

- genetic heterogeneity in cancer. *Nat Rev Genet.* 2019; 20: 404-416.
6. Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature.* 2017; 545: 446-451.
 7. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017; 542: 115-118.
 8. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, et al. International evaluation of an AI system for breast cancer screening. *Nature.* 2020; 577: 89-94.
 9. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med.* 2019; 25: 1054-1056.
 10. De Melo P, Davtyan M. High accuracy classification of populations with breast cancer: SVM approach. *Cancer Research Journal.* 2023; 11.
 11. Chen RJ, Lu MY, Wang J, Williamson DFK, Rodig SJ, et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans Med Imaging.* 2022; 41: 757-770.
 12. Rizopoulos D. Joint Models for Longitudinal and Time-to-Event Data. Chapman and Hall/CRC; 2012.
 13. Björnsson B, Borrebaeck C, Elander N, Gasslander T, Gawel DR, et al. Digital twins to personalize medicine. *Genome Med.* 2020; 12: 4.
 14. Zhang J, Cunningham JJ, Brown JS, Gatenby RA. Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. *Nat Commun.* 2017; 8: 1816.
 15. De Melo P. Anticipatory oncology: a new paradigm for hidden breast cancer progression. *Cancer Science & Research.* 2026; 9.
 16. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019; 17: 230.
 17. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis: the TRIPOD statement. *Ann Intern Med.* 2015; 162: 55-63.
 18. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine-learning methods. *BMJ.* 2024; 385.
 19. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med.* 2019; 170: 51-58.
 20. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med.* 2020; 26: 1364-1374.
 21. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med.* 2020; 26: 1351-1363.
 22. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019; 25: 44-56.
 23. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX et, al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med.* 2019; 25: 1337-1340.
 24. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D, et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019; 17: 195.
 25. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019; 366: 447-453.
 26. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations. *Nat Med.* 2021; 27: 2176-2182.
 27. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med.* 2018; 169: 866-872.