Argumentation Based Systems Journal

Journal homepage: www.absj.org
eISSN: 3120-1024

Bayesian Hierarchical Modeling and Clustering for Malignant Cancer Diagnosis

Sijia Zhu[1,†], Jonathan Ma[1,†], Zhe Liu[2,*]

[1] Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore 21218, USA
[2] School of Computer Sciences, Universiti Sains Malaysia, Pennag 11800, Malaysia

ARTICLE INFO

Article history:
Received 29 December 2025
Received in revised form 2 February 2026
Accepted 6 February 2026
Available online 7 February 2026

Keywords:
Late-stage cancer; Bayesian hierarchical logistic regression; No-U-Turn sampler; Patient heterogeneity; Cancer risk cluster

ABSTRACT

Late-stage cancer diagnosis remains a major barrier to improving survival rates, yet the relative contributions of tumor, patient, and region-level factors have not been well quantified. In this study, we develop a 3-layer Bayesian hierarchical logistic regression model to investigate late-stage cancer diagnosis. The model includes fixed effects for tumor characteristics and random effects for patients and regions. Model parameters are estimated using the No-U-Turn Sampler, and posterior samples are evaluated with effective sample sizes and convergence diagnostics. From intra-class correlation estimates, we find that patient-level variation has a substantially stronger influence on late-stage diagnosis than region-level variation. Lastly, we utilize a Gaussian Mixture Model to cluster posterior patient-level random effects, identifying nine distinct clusters characterized by age, sex, and tumor features. Our findings suggest that individualized, patient-focused strategies may be more effective than geographically targeted approaches for promoting earlier cancer detection.

1. Introduction

Cancer remains one of the leading causes of mortality in the United States, making early detection essential for improving survival outcomes [1–3]. However, timely diagnosis is not uniformly distributed across populations. For example, Oh et al. [4] reported substantial geographic disparities in cancer diagnosis between metropolitan and rural regions. This observation has motivated extensive research on regional differences in cancer outcomes [5]. Beyond regional factors, patient-level characteristics also play a critical role. Even among patients presenting with the same tumor type, demographic

* *Corresponding author.*
E-mail address: liuzhe921@gmail.com
† *These authors contributed equally to this work.*

https://doi.org/10.59543/w22sya71

characteristics such as sex and age can vary considerably. These factors may further interact with tumor features and influence the likelihood of late-stage diagnosis. [6].

While prior studies have examined either geographic or demographic influences on cancer diagnosis, few have explicitly quantified how multiple hierarchical levels contribute to late-stage cancer. We address this gap by decomposing the variation across tumor-level, patient-level, and region-levels in late-stage cancer diagnosis. Finally, we perform patient-level clustering based on key covariates to identify groups with the highest risk of late-stage diagnosis. This work can play an important role in earlier cancer detection and in reducing the incidence of late-stage diagnosis.

Although substantial work has examined disparities in cancer diagnosis [7–9], most existing studies focus on a single analytical layer and fail to account for how tumor-, patient-, and system-level factors jointly shape clinical outcomes when interactions are present [10, 11]. For example, Flanary et al. [12] developed a multivariate regression model and demonstrated that patients with Medicaid or no insurance have significantly higher odds of being diagnosed with late-stage cancer compared with those covered by the military health system. Similarly, Wassie et al. [13] employed multivariable logistic regression to identify sociodemographic and clinical factors associated with advanced-stage cancer at diagnosis among adult patients.

Beyond these single-layer analyses, an additional challenge lies in the high dimensionality and complexity of many cancer datasets, which makes it difficult to disentangle sources of variation using conventional regression approaches [14–16]. To address this issue, McGlothlin et al. [17] adopted Bayesian hierarchical logistic regression (BHLR) models to improve uncertainty quantification. Similarly, Lin et al. [18] developed a BHLR framework to integrate multiple family health histories for cancer risk prediction, while Allen et al. [19] applied a BHLR model to estimate the effects of risk factors on bladder cancer. Collectively, these studies underscore the importance of Bayesian hierarchical methods for modeling complex clinical data and motivate their application to the study of late-stage cancer diagnosis [20–24].

Clustering is a useful tool for identifying heterogeneity in cancer risk and disease stage [25–27]. For example, Vitelli et al. [28] applied a rank-based Bayesian clustering approach to RNA-seq data from 12 tumor types to identify molecular subgroups across cancers. Similarly, Nicholls et al. [29] proposed a Bayesian nonparametric clustering framework that accounts for individual-level observation uncertainty and demonstrated its utility on gene expression data. However, these clustering approaches have largely been developed independently of hierarchical logistic regression models and have not been applied to population-level epidemiologic outcomes, such as stage at cancer diagnosis.

In this study, we propose a three-level Bayesian hierarchical logistic regression framework to quantify the relative contributions of tumor-, patient-, and region-level factors to late-stage cancer diagnosis. Using SEER data, we explicitly decompose the variability of results across hierarchical levels and perform Bayesian inference using the No-U-Turn Sampler. Furthermore, we get intra-class correlations to characterize the heterogeneity and demonstrate that patient-level variation exceeds regional effects. Building on the hierarchical model, we conduct posterior patient-level random effect clustering using a Gaussian Mixture Model to identify latent risk subgroups with distinct demographic and tumor characteristics. Together, these contributions provide insights into cancer diagnosis disparities, highlighting the primacy of individual-level heterogeneity and strategies for earlier detection in public health.

2. Data

In this section, we introduce the dataset, how we handle missingness, and all explore-transform-load (ETL) steps.

2.1 SEER Dataset

We use cancer data from the *Surveillance, Epidemiology, and End Results* (SEER)[*] provided by the National Cancer Institute (NCI). It contains standardized information on cancer cases across multiple U.S. regions.

2.2 Missingness Assessment

We evaluate whether missing data are related to observed covariates under the Missing Completely at Random (MCAR) assumption:

First, we define an indicator variable M_i for each observed covariate:

$$M_i = \begin{cases} 1, & \text{if the observation is missing,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

We then model the probability of missingness due to chance using logistic regression:

$$\Pr(M_i = 1) = \text{logit}^{-1}(\gamma_0 + \gamma_1 \text{Age}_i + \gamma_2 \text{Race}_i + \gamma_3 \text{Region}_i + \gamma_4 \text{Size}_i + \gamma_5 \text{Grade}_i)\,. \tag{2}$$

The regression results are shown in Table 1 and Table 2. They indicate that multiple covariates (including sex, race, region, and age) were significantly associated with missingness, so for this study, we will assume data is missing at random (MAR). As a result, we keep only complete cases for analysis, resulting in a sample size of 209,123.

2.3 Data Pre-processing

We take further steps to prepare our data. Firstly, the tumor grade variable is grouped into low (grades one and two) and high (grades three and four) which enables us to use a logistic model. Region identification is changed to show three levels: large metropolitan, small metropolitan, and nonmetropolitan with adjacent areas. Originally, nine levels were present, but we chose three to ensure a relatively stable balance We then remove observations that show missing size measurements along with those labeled as "Unknown" or "Blank(s)". We pool rare categories together to ensure a relative balance. Continuous measures, such as the size of the disease, are kept in their original form, and the distributions are examined to confirm that these measures are relatively unbiased.

Table 3 shows the organization for our final dataset, consisting of 37,530 observations.

3. Proposed Method

In this section, we will introduce the basic concept of our hierarchical model and clustering.

3.1 Bayesian Hierarchical Logistic Model

For each tumor t of patient p in region r, we define a dichotomous result:

$$Y_{tpr} = \begin{cases} 1, & \text{late-stage diagnosis,} \\ 0, & \text{early-stage diagnosis.} \end{cases} \tag{3}$$

Conditional on the success probability p_{tpr}, we assume

$$Y_{tpr} \mid p_{tpr} \sim \text{Bernoulli}(p_{tpr}). \tag{4}$$

[*] https://seer.cancer.gov/

Table 1: Significant Coefficients for the Missingness Model (Pt. 1).

Term	Est.	SE	z Stat	Prob
Age 15-19	-0.84	0.10	-8.60	0.0000
Age 20-24	-1.68	0.10	-17.14	0.0000
Age 25-29	-2.17	0.09	-22.88	0.0000
Age 30-34	-2.61	0.09	-28.16	0.0000
Age 35-39	-2.70	0.09	-29.90	0.0000
Age 40-44	-2.78	0.09	-31.34	0.0000
Age 45-49	-2.76	0.09	-31.43	0.0000
Age 50-54	-2.86	0.09	-32.75	0.0000
Age 55-59	-2.88	0.09	-33.10	0.0000
Age 60-64	-2.98	0.09	-34.27	0.0000
Age 65-69	-3.08	0.09	-35.40	0.0000
Age 70-74	-3.11	0.09	-35.72	0.0000
Age 75-79	-3.12	0.09	-35.82	0.0000
Age 80-84	-3.07	0.09	-35.18	0.0000
Age 85-89	-3.01	0.09	-34.24	0.0000
Age 90+	-2.91	0.09	-32.43	0.0000
Married	-0.11	0.01	-8.13	0.0000
Unmarried	0.06	0.02	3.31	0.0009
Grade II	3.03	0.05	66.58	0.0000
Grade Null	1.91	0.33	5.72	0.0000
Grade III	3.40	0.05	75.01	0.0000
Grade IV	3.88	0.05	83.75	0.0000
Grade 1	3.29	0.05	71.06	0.0000

Level 1: Tumor level model

We apply a logistic regression model to decompose variation of late-stage diagnosis into fixed covariate effects and multilevel random effects.

$$\operatorname{logit}(p_{tpr}) = \mathbf{z}_{tpr}^{\top}\beta_T + \mathbf{w}_p^{\top}\beta_P + u_p + v_r, \tag{5}$$

The variable $\mathbf{z}_{tpr}$ denotes the tumor-level covariate vector, $\mathbf{w}_p$ represents the patient-level covariate vector, β_T and β_P are the fixed effect coefficients for tumor-level and patient-level predictors. u_p is used for patient-level random effects, and v_r for region-level random effects.

Level 2: Patient-level Model

$$u_p \mid \sigma_u^2 \sim N(0, \sigma_u^2), \qquad p = 1, \ldots, P, \tag{6}$$

This layer represents heterogeneity between patients and captures correlation among multiple tumors from the same patient.

Table 2: Significant Coefficients for the Missingness Model (Pt. 2)

Term	Est.	SE	z Stat	Prob
size10	1.10	0.13	8.27	0.0000
size13	-1.18	0.36	-3.29	0.0011
size14	-1.15	0.18	-6.44	0.0000
size15	-0.78	0.20	-3.93	0.0009
size21	-3.99	0.24	-16.41	0.0000
size22	-3.58	0.07	-49.79	0.0000
size23	-3.83	0.13	-28.81	0.0000
size24	-2.22	0.10	-21.95	0.0000
size25	0.43	0.04	10.33	0.0000
size28	-3.10	0.41	-7.58	0.0000
size30	0.16	0.01	12.20	0.0000
size31	3.67	0.03	135.55	0.0000
size32	-0.47	0.07	-6.89	0.0000
size33	-3.08	0.13	-22.98	0.0000
size34	1.11	0.12	9.36	0.0000
size35	-0.83	0.25	-3.36	0.0008
size36	-3.53	0.41	-8.61	0.0000
size40	0.45	0.01	32.01	0.0000
size41	-0.51	0.03	-17.11	0.0000
size42	1.25	0.03	47.89	0.0000
size43	2.36	0.04	59.77	0.0000
size45	0.55	0.04	12.23	0.0000
size47	-0.70	0.15	-4.86	0.0000
size50	-1.47	0.02	-70.97	0.0000
size51	-3.81	0.12	-30.84	0.0000
size52	-1.96	0.11	-17.63	0.0000
size53	-1.82	0.13	-14.06	0.0000
size54	-1.26	0.15	-8.55	0.0000
size60	2.16	0.02	132.51	0.0000
size62	-4.79	1.00	-4.79	0.0000
size70	-1.09	0.09	-11.78	0.0000
size80	-0.67	0.11	-6.04	0.0000
size90	1.36	0.05	28.68	0.0000
size98	6.78	0.05	149.72	0.0000
size99	1.72	0.07	24.95	0.0000

Table 3: Hierarchical Dataset Structure.

Level	Variable
Tumor-level	Tumor Stage
	Tumor Grade
	Tumor Size
	Tumor Site
Patient-level	Patient ID
	Age Group
	Sex
	Race/Ethnicity
	Year of Diagnosis
	Marital Status
Region-level	Region Size Classification

Level 3: Region-level Model

$$v_r \mid \mu_v, \sigma_v^2 \sim N(\mu_v, \sigma_v^2), \qquad r = 1, \ldots, R, \tag{7}$$

The term μ_v is used for average regional effects, and σ_v^2 represents between-region variability.

3.2 Model-based Clustering

We use risk stratification to identify subgroups from our results. The term $\bar{u}_p$ denotes the posterior mean of the patient-level random effect u_p from the hierarchical logistic model. We model the distribution of $\bar{u}_p$ by a Gaussian Mixture Model (GMM) through Maximum Likelihood Estimation:

$$\bar{u}_p \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\bar{u}_p \mid \mu_k, \sigma_k^2), \tag{8}$$

The term π_k is the mixing weight (with $\sum_k \pi_k = 1$), and the tuple (μ_k, σ_k^2) represents the mean and variance parameters of kth Gaussian component. We estimate the parameter vector $(\pi_k, \mu_k, \sigma_k^2)_{k=1}^K$ by the Expectation-Maximization (EM) algorithm.

Given the estimated parameters, we assign each patient a soft cluster label using the posterior probability of belonging to cluster k:

$$\gamma_{pk} = \frac{\pi_k \mathcal{N}(\bar{u}_p \mid \mu_k, \sigma_k^2)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\bar{u}_p \mid \mu_j, \sigma_j^2)}. \tag{9}$$

To obtain a hard decision boundary, we use these probabilities γ_{pk} to assign patients to the most probable cluster via $\arg\max_k \gamma_{pk}$.

3.3 No-U-Turn Sampler

The No-U-Turn Sampler (NUTS) [30] is an extension of Hamiltonian Monte Carlo (HMC). It automatically determines the trajectory length during sampling and uses the Metropolis-Hastings Accept-Reject criteria. This avoids the need for manual tuning of the number of leapfrog steps and improves mixing in BHLR models. The detailed algorithm is shown in Algorithm 1.

Algorithm 1 No-U-Turn Sampler

Input: Data y, covariates X, model structure, priors.
Initialize: Choose initial parameter vector $\Theta^{(0)}$ and momentum $r^{(0)} \sim \mathcal{N}(0, M)$.
for $t = 1, 2, \ldots, T$ **do**
    Resample momentum: $r^{(t)} \sim \mathcal{N}(0, M)$.
    Use leapfrog integration to explore candidate states and build a binary tree of proposals (Θ', r').
    Continue expanding the tree until the trajectory forms a U-turn, indicating that further expansion would retrace previous steps.
    Select one valid proposal $\Theta^{(t)}$ uniformly from the built tree.
    Apply the Metropolis-Hastings rule to accept or reject the proposal.
end for
Output: Posterior samples $\{\Theta^{(t)}\}_{t=1}^{T}$.

4. Bayesian Inference

In this section, we provide more details and justification for our model and clustering specification.

4.1 Model Specification

We define the likelihood, prior distributions, and joint posterior for the Bayesian hierarchical logistic regression model.

Likelihood

The full likelihood of the binary tumor level result y is given as

$$p(y \mid \beta, u, v) = \prod_{p=1}^{P} \prod_{r=1}^{R} \prod_{t=1}^{n_{pr}} p_{tpr}^{y_{tpr}} (1 - p_{tpr})^{1 - y_{tpr}}. \tag{10}$$

Prior

We assume patient-level and region-level random effects follow normal priors with bigger variances to make sure no overfitting occurs:

$$u_p \sim \mathcal{N}(0, 4), \qquad v_r \sim \mathcal{N}(0, 25). \tag{11}$$

Based on [31], we apply $\sigma \sim t(3, 2.5)$ with density $p(\sigma) \propto \left(1 + \frac{\sigma^2}{3 \cdot 2.5^2}\right)^{-2}$, $\sigma > 0$.

Joint Posterior Distribution

Combining likelihood and priors gives the posterior:

$$p(\beta, u, v, \sigma_u, \sigma_v \mid y) \propto p(y \mid \beta, u, v) \left(\prod_{p=1}^{P} \mathcal{N}(u_p \mid 0, \sigma_u^2) \right) \left(\prod_{r=1}^{R} \mathcal{N}(v_r \mid 0, \sigma_v^2) \right) p(\beta)\, p(\sigma_u)\, p(\sigma_v). \tag{12}$$

4.2 Posterior Inference

We employ NUTS to sample from the joint posterior distribution:

$$p(\Theta \mid y, X), \quad \Theta = (\boldsymbol{\beta}, \{u_p\}, \{v_r\}, \sigma_u^2, \sigma_v^2). \tag{13}$$

We ran 4 Markov chains with 3,000 iterations each, including 1,500 warm-up iterations. After burn-in, a total of 6,000 post–warm-up draws were retained for inference. To improve exploration of the posterior geometry, particularly for hierarchical variance parameters, we set step size `adapt_delta = 0.9995` to reduce divergent transitions and `max_treedepth = 15` to allow sufficiently long trajectories without causing a burden on our CPU constraints.

4.2.1 Model Diagnostic

We verified convergence via E-BFMI (Expected Bayes Fraction of Missing Information), effective sample sizes (ESS), and $\widehat{R}$ diagnostics (all $\widehat{R} \approx 1.00$), indicating satisfactory posterior exploration. The model diagnostics are shown in Table 4.

Table 4: Posterior Estimates and Convergence.

Term	Estimate	SE	2.5%	97.5%	Rhat	Bulk ESS	Tail ESS
Hyperparameters							
sd(Patient)	9.11	1.85	5.69	13.01	1.01	494	593
sd(Region)	0.79	0.89	0.02	3.23	1.00	2931	3694
Intercept	-4.32	1.36	-7.17	-1.75	1.00	1677	2623
age01-04	-0.34	1.89	-4.05	3.32	1.00	9027	5229
age05-09	-1.17	1.92	-4.82	2.58	1.00	9688	5284
age10-14	0.09	1.88	-3.47	3.91	1.00	8516	4916
age15-19	0.07	1.90	-3.65	3.87	1.00	10007	4664
age20-24	-1.28	1.83	-4.95	2.31	1.00	7836	4698
age25-29	-0.24	1.67	-3.52	3.03	1.00	6268	4781
age30-34	1.14	1.46	-1.81	4.08	1.00	4090	4215
age35-39	0.49	1.26	-2.05	2.92	1.00	3444	4116
age40-44	0.45	1.09	-1.69	2.63	1.00	3134	4143
age45-49	0.38	0.93	-1.44	2.19	1.00	2876	3588
age50-54	0.98	0.89	-0.73	2.76	1.00	2551	3412
age55-59	0.12	0.84	-1.78	1.48	1.00	2555	3562
age60-64	-0.23	0.82	-1.82	1.38	1.00	2305	3031
age65-69	0.36	0.80	-1.22	1.93	1.00	2668	3376
age70-74	-0.23	0.82	-1.82	1.38	1.00	2305	3031
age75-79	1.42	0.86	-0.24	3.12	1.00	2719	3650
age80-84	-0.58	0.94	-2.49	1.24	1.00	2326	3739
age85-89	-0.05	1.04	-2.10	2.00	1.00	2787	3505
age90+	-0.62	1.18	-2.99	1.66	1.00	3452	4375
sexMale	0.58	0.45	-0.28	1.51	1.00	2666	3249
racethOther	1.79	0.81	0.29	3.43	1.00	1990	3574
racethWhite	-0.03	0.67	-1.37	1.25	1.00	2804	3382
gradeStart	6.06	1.13	3.96	8.41	1.01	635	881
size_std	1.22	0.30	0.69	1.87	1.01	793	1682
year_std	-0.02	0.21	-0.44	0.39	1.00	2784	3506
marryUnmarried	0.47	0.46	-0.40	1.41	1.00	2780	2985

Figure 1 shows that the observed mean late-stage diagnosis rate aligns well with the distribution of replicated values under the posterior, indicating that the model captures key data features and shows no systematic misfit.

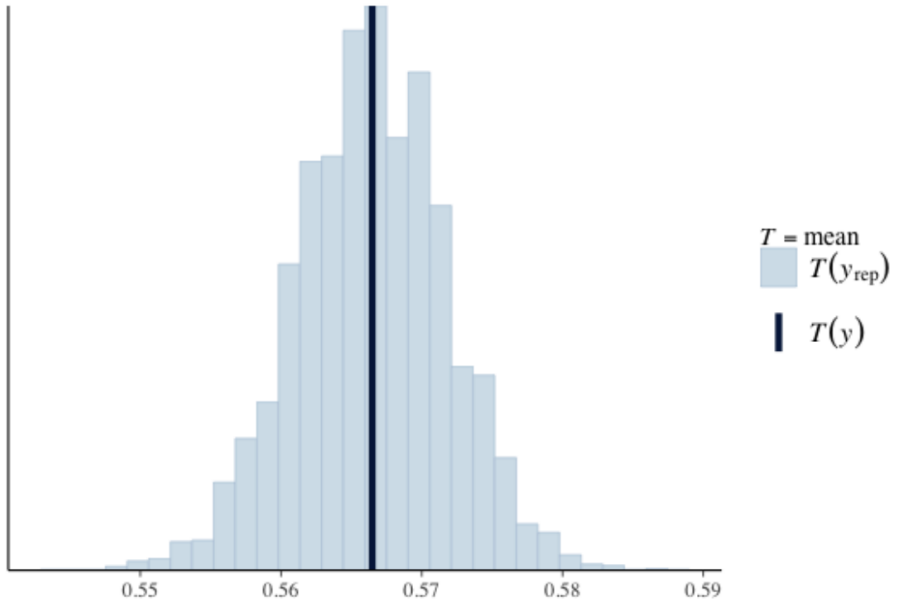

Figure 1: Posterior predictive distribution of the mean late-stage diagnosis rate.

4.3 Intra-class Correlation

For logistic models, the residual variance is $\frac{\pi^2}{3}$. Patient-level and region-level ICCs are calculated as follows:

$$\mathrm{ICC}_p = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2 + \pi^2/3}, \qquad \mathrm{ICC}_r = \frac{\sigma_v^2}{\sigma_u^2 + \sigma_v^2 + \pi^2/3}. \tag{14}$$

Using posterior estimates $\sigma_u \approx 9.11$ and $\sigma_v \approx 0.79$:

$$\mathrm{ICC}_p \approx 0.94, \qquad \mathrm{ICC}_r \approx 0.007,$$

This result confirms that nearly all unexplained variation arises at the patient level.

4.4 Shrinkage Effects

Patient level Under the prior $u_p \sim \mathcal{N}(0, \sigma_u^2)$, the posterior mean is

$$\mathbb{E}(u_p \mid \boldsymbol{\beta}, \{v_r\}, \sigma_u^2, \sigma_v^2) = w_p \, y_{\mu_p}, \qquad w_p = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2/n_p}. \tag{15}$$

The term μ_p is the sample mean. When n_p or σ_u^2 are small, the weight w_p decreases, and the estimate of u_p "shrinks" toward 0.

Region level Given the prior $v_r \sim \mathcal{N}(0, \sigma_v^2)$, the posterior mean is

$$\mathbb{E}(v_r \mid \boldsymbol{\beta}, \{u_p\}, \sigma_u^2, \sigma_v^2) = w_r \, y_{v_p}, \qquad w_r = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\varepsilon^2/n_r}. \tag{16}$$

The term y_{v_p} is the sample mean. Because n_r is typically small and σ_v^2 is much smaller than σ_u^2, the weight w_r becomes close to zero, resulting in strong shrinkage of all region effects.

At the patient level, the model reveals substantial variability: some patients have strongly negative intercepts (indicating a lower risk of late-stage diagnosis), while others have strongly positive values (suggesting a higher risk). The shrinkage effect tempers this variability by adjusting extreme values inward, leading to more conservative yet robust estimates. This is critical in cases where individual data are sparse, as it prevents the model from overreacting to noisy observations.

At the regional level, the shrinkage effect is even more pronounced. All three regions show posterior means close to zero, and the wide credible intervals indicate substantial uncertainty. This means that after adjusting for patient-level covariates, regional differences contribute very little to the variability in late-stage diagnosis. This shrinkage reinforces the dominance of individual-level heterogeneity, aligning with the high patient-level intra-class correlation.

Figure 2 visualizes the strong shrinkage of region-level raw rates toward the posterior mean. This result reflects both the low ICC_r and high posterior uncertainty, emphasizing that observed regional variation may largely be driven by noise rather than structural effects.

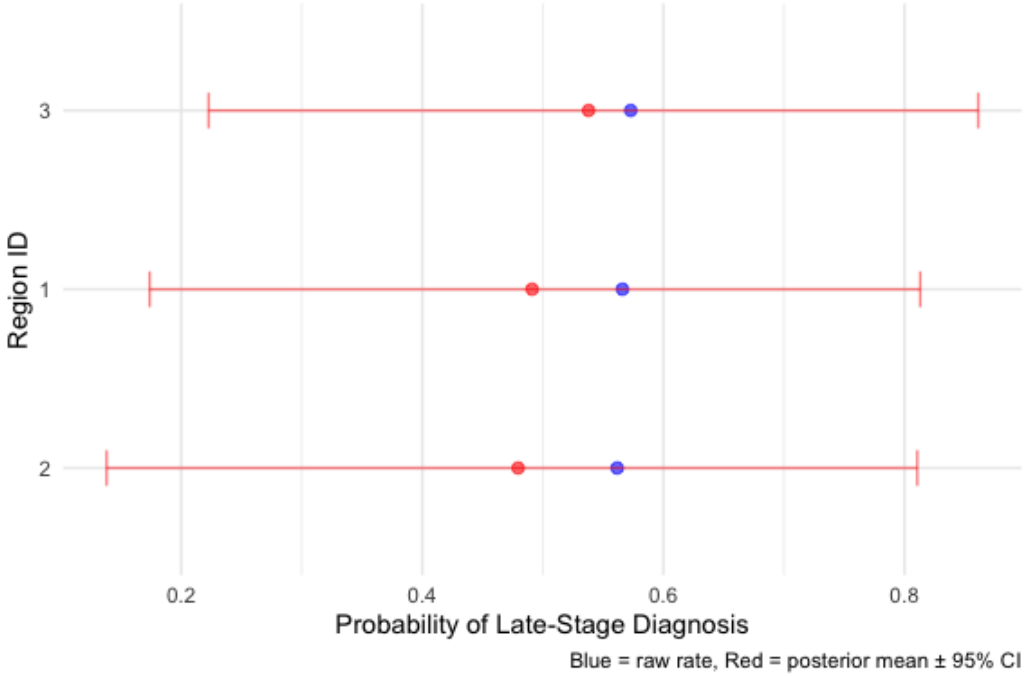

Figure 2: Shrinkage of Raw vs Posterior Region Level Estimates.

4.5 Clustering Analysis

To identify the distinct cancer risk characteristics in the risk subgroups, we fitted a Gaussian Mixture Model (GMM) for the cluster analysis. Table 5 summarizes the most common demographic characteristics within each cluster.

The table shows that clusters 2 and 6 are the largest groups, which mainly consist of White females aged 60–64 living in large metropolitan areas (Region 1). This demographic profile represents the most dominant risk characteristics in our sample. Most of the other clusters primarily consist of White female patients residing in large metropolitan areas (Region 1) with various age ranges, like clusters 1 and 8, which are an older group (75–79 years), signaling that increased risk of late-stage diagnosis is associated with age.

In contrast, Cluster 4 is the only group that primarily consisted of males, confirming the gender differences known empirically. Cluster 5 is the only cluster where individuals in the "Other" racial

category are most common, representing a different latent risk profile for minority metropolitan elder women. Clusters 3 likely represent rare or outlier patient profiles. Across all clusters, Region 1 (large metropolitan areas) is the most common geography, indicating minimal regional differentiation. Our clustering analysis divided the large homogeneous old White female metropolitan population into various latent risk groups and also isolated the smaller minority and male subpopulation whose patterns may need more investigation.

Table 5: Clustering results.

Cluster	Count	Age	Race	Sex	Region*
1	697	75-79 years	White	Female	1
2	777	60-64 years	White	Female	1
3	4	60-64 years	White	Female	1
4	314	70-74 years	White	Male	1
5	56	75-79 years	Other	Female	1
6	817	60-64 years	White	Female	1
7	101	60-64 years	White	Female	1
8	120	75-79 years	White	Female	1
9	145	70-74 years	White	Female	1

* Region=1 corresponds to large metropolitan areas (≥ 1 million population).

5. Conclusion

This study reinforces the dominant role of individual-level factors, particularly tumor size, tumor grade, age, and sex, in predicting late-stage cancer diagnosis. Despite prior work identifying spatial disparities, our analysis found negligible region-level variation (ICCregion ≈ 0.007) after accounting for covariates, with nearly all residual heterogeneity arising at the patient level (ICCpatient ≈ 0.94). The coarse regional granularity in SEER and the inclusion of strong individual-level predictors likely contribute to this result.

We employed Bayesian hierarchical logistic regression to model patient- and region-level effects, leveraging partial pooling and shrinkage to stabilize estimates, especially in sparsely populated strata. Model diagnostics confirmed good posterior convergence and predictive calibration. Additionally, posterior clustering of patient-level random effects uncovered nine latent subgroups, each characterized by distinct demographic and tumor profiles, emphasizing the potential of Bayesian modeling to uncover unobserved heterogeneity not captured by covariates alone.

From a clinical and policy perspective, these findings support a shift toward patient-focused interventions such as individualized risk profiling and targeted screening over region-based strategies. This aligns with emerging paradigms in precision public health.

However, several limitations remain. SEER's coarse regional identifiers may obscure finer-scale disparities, and residual patient-level variance likely reflects unmeasured confounding factors like screening access or insurance status. While clustering was applied to posterior means of random effects, this two-stage frequentist approach does not fully propagate uncertainty, potentially understating variation in subgroup assignments. Due to time and computational constraints, a fully Bayesian clustering method (e.g., DP-GMM) was not implemented.

Acknowledgement

This research was not funded by any grant.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Siegel, R. L., Kratzer, T. B., Giaquinto, A. N., Sung, H., & Jemal, A. (2025). Cancer statistics, 2025. *CA: A Cancer Journal for Clinicians*, *75*(1), 10–34. https://doi.org/https://doi.org/10.3322/caac.21863

[2] Arafeh, R., Shibue, T., Dempster, J. M., Hahn, W. C., & Vazquez, F. (2025). The present and future of the cancer dependency map. *Nature Reviews Cancer*, *25*(1), 59–73. https://doi.org/https://doi.org/10.1038/s41568-024-00763-x

[3] Xiong, X., Zheng, L.-W., Ding, Y., Chen, Y.-F., Cai, Y.-W., Wang, L.-P., Huang, L., Liu, C.-C., Shao, Z.-M., & Yu, K.-D. (2025). Breast cancer: Pathogenesis and treatments. *Signal Transduction and Targeted Therapy*, *10*(1), 49. https://doi.org/https://doi.org/10.1038/s41392-024-02108-4

[4] Oh, D. L., Wang, K., Goldberg, D., Schumacher, K., Yang, J., Lin, K., Gomez, S. L., & Shariff-Marco, S. (2024). Disparities in cancer stage of diagnosis by rurality in california, 2015 to 2019. *Cancer Epidemiology, Biomarkers & Prevention*, *33*(11), 1523–1531. https://doi.org/https://doi.org/10.1158/1055-9965.EPI-24-0373

[5] Laguna, J. C., García-Pardo, M., Alessi, J., Barrios, C., Singh, N., Al-Shamsi, H. O., Loong, H., Ferriol, M., Recondo, G., & Mezquita, L. (2024). Geographic differences in lung cancer: Focus on carcinogens, genetic predisposition, and molecular epidemiology. *Therapeutic Advances in Medical Oncology*, *16*, 17588359241231260. https://doi.org/https://doi.org/10.1177/17588359241231260

[6] Ullah, A., Kenol, G. S., & Lee, K. T. (2024). Chordoma: Demographics and survival analysis with a focus on racial disparities and the role of surgery, a us population-based study. *Clinical and Translational Oncology*, *26*(1), 109–118. https://doi.org/https://doi.org/10.1007/s12094-023-03229-8

[7] Vyas, A., Kumar, K., Sharma, A., Verma, D., Bhatia, D., Wahi, N., & Yadav, A. K. (2025). Advancing the frontier of artificial intelligence on emerging technologies to redefine cancer diagnosis and care. *Computers in Biology and Medicine*, *191*, 110178. https://doi.org/https://doi.org/10.1016/j.compbiomed.2025.110178

[8] Saeidnia, H. R., Firuzpour, F., Kozak, M., & Majd, H. S. (2025). Advancing cancer diagnosis and treatment: Integrating image analysis and ai algorithms for enhanced clinical practice. *Artificial Intelligence Review*, *58*(4), 105. https://doi.org/https://doi.org/10.1007/s10462-025-11117-w

[9] Ng, A. B. C. D., Asif, A., Agarwal, R., Panebianco, V., Girometti, R., Ghai, S., Gómez-Gómez, E., Budäus, L., Barrett, T., & Radtke, J. P. (2025). Biparametric vs multiparametric mri for prostate cancer diagnosis: The prime diagnostic clinical trial. *JAMA*, *334*(13), 1170–1179. https://doi.org/https://doi.org/10.1001/jama.2025.13722

[10] Kurzeder, C., Nguyen-Sträuli, B. D., Krol, I., Ring, A., Castro-Giner, F., Nüesch, M., Asawa, S., Zhang, Y. W., Budinjas, S., & Gvozdenovic, A. (2025). Digoxin for reduction of circulating tumor cell cluster size in metastatic breast cancer: A proof-of-concept trial. *Nature Medicine*, *31*(4), 1120–1124. https://doi.org/https://doi.org/10.1038/s41591-024-03486-6

[11] Zhu, S., Liu, Z., Letchmunan, S., Ulutagay, G., & Ullah, K. (2025). Novel distance measures on complex picture fuzzy environment: Applications in pattern recognition, medical diagnosis and clustering. *Journal of Applied Mathematics and Computing*, *71*(2), 1743–1775. https://doi.org/https://doi.org/10.1007/s12190-024-02293-z

[12] Flanary, J. T., Lin, J., Shriver, C. D., & Zhu, K. (2023). Cancer stage at diagnosis: Comparison of insurance status in seer to the department of defense cancer registry. *Cancer Medicine*, *12*(22), 20989–21000. https://doi.org/https://doi.org/10.1002/cam4.6655

[13] Wassie, L. A., Mekonnen, C. K., Tiruneh, Y. M., Melkam, M., Belachew, E. A., & Zegeye, A. F. (2024). Advanced-stage presentation of cancer at the time of diagnosis and its associated factors among adult cancer patients at northwest amhara comprehensive specialized hospitals, northwest ethiopia, 2022. *BMC Cancer*, *24*, 68. https://doi.org/https://doi.org/10.1186/s12885-024-11835-4

[14] Cui, J., Li, Y., Shen, D., & Wang, Y. (2026). Mgcm: Multi-modal graph convolutional mamba for cancer survival prediction. *Pattern Recognition*, *169*, 111991. https://doi.org/https://doi.org/10.1016/j.patcog.2025.111991

[15] Wang, J., Zhang, Z., & Wang, Y. (2025). Utilizing feature selection techniques for ai-driven tumor subtype classification: Enhancing precision in cancer diagnostics. *Biomolecules*, *15*(1), 81. https://doi.org/https://doi.org/10.3390/biom15010081

[16] Zhang, J., Che, Y., Liu, R., Wang, Z., & Liu, W. (2025). Deep learning–driven multi-omics analysis: Enhancing cancer diagnostics and therapeutics. *Briefings in Bioinformatics*, *26*(4), bbaf440. https://doi.org/https://doi.org/10.1093/bib/bbaf440

[17] McGlothlin, A. E., & Viele, K. (2018). Bayesian hierarchical models. *JAMA*, *320*(22), 2365–2366. https://doi.org/https://doi.org/10.1001/jama.2018.17977

[18] Lin, J., Myers, M. F., Koehly, L. M., & Marcum, C. S. (2019). A bayesian hierarchical logistic regression model of multiple informant family health histories. *BMC Medical Research Methodology*, *19*, 107. https://doi.org/https://doi.org/10.1186/s12874-019-0700-5

[19] Allen, B. (2020). Bayesian hierarchical dose–response meta-analysis of bladder cancer and inorganic arsenic exposure. *Environment International*, *140*, 105709. https://doi.org/https://doi.org/10.1016/j.envint.2020.105709

[20] Mimnagh, N., & Prado, E. (2026). An introduction to bayesian hierarchical modelling applied to wildlife monitoring. https://doi.org/https://doi.org/10.1007/978-3-032-05821-8_8

[21] Vloeberghs, R., Urai, A. E., Desender, K., & Linderman, S. W. (2025). A bayesian hierarchical model of trial-to-trial fluctuations in decision criterion. *PLOS Computational Biology*, *21*(7), e1013291. https://doi.org/https://doi.org/10.1371/journal.pcbi.1013291

[22] Uddandarao, D. P. (2024). Improving employment survey estimates in data-scarce regions using dynamic bayesian hierarchical models: Addressing measurement challenges in developing countries. *Panamerican Mathematical Journal*, *34*(4), 2024. https://doi.org/https://doi.org/10.1007/s40819-024-01567-3

[23] Ouyang, L., Zhu, S., Ye, K., Park, C., & Wang, M. (2022). Robust bayesian hierarchical modeling and inference using scale mixtures of normal distributions. *IISE Transactions*, *54*(7), 659–671. https://doi.org/https://doi.org/10.1080/24725854.2021.2013550

[24] Ievoli, R., Gardini, A., & Palazzo, L. (2023). The role of passing network indicators in modeling football outcomes: An application using bayesian hierarchical models. *AStA Advances in Statistical Analysis*, *107*(1), 153–175. https://doi.org/https://doi.org/10.1007/s10182-023-00479-x

[25] Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika*, *65*(1), 31–38. https://doi.org/https://doi.org/10.1093/biomet/65.1.31

[26] Basch, E., Schrag, D., Jansen, J., Henson, S., Ginos, B., Stover, A. M., Carr, P., Spears, P. A., Jonsson, M., & Deal, A. M. (2025). Symptom monitoring with electronic patient-reported outcomes during cancer treatment: Final results of the pro-tect cluster-randomized trial. *Nature Medicine*, *31*(4), 1225–1232. https://doi.org/https://doi.org/10.1038/s41591-025-03507-y

[27] Eruslanov, E., Nefedova, Y., & Gabrilovich, D. I. (2025). The heterogeneity of neutrophils in cancer and its implication for therapeutic targeting. *Nature Immunology*, *26*(1), 17–28. https://doi.org/https://doi.org/10.1038/s41590-024-02029-y

[28] Vitelli, V. (2023). Transcriptomic pan-cancer analysis using rank-based bayesian clustering. *BMC Medical Genomics*, *16*, 112. https://doi.org/https://doi.org/10.1186/s12920-023-01541-2

[29] Nicholls, K., Kirk, P. D. W., & Wallace, C. (2024). Bayesian clustering with uncertain data. *PLoS Computational Biology*, *20*(9), e1012301. https://doi.org/https://doi.org/10.1371/journal.pcbi.1012301

[30] Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623. https://doi.org/https://doi.org/10.5555/2627435.2638586

[31] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533. https://doi.org/https://doi.org/10.1214/06-BA117A