Early-Onset Colon Cancer Trends Analysis: CI5plus Dataset

A Hierarchical Bayesian Modeling Approach

Author: Oge Ohia
Project Repository: Early-Onset-Colon-Cancer-Trends-CI5plus

Contents

Executive Summary

This project investigates **global temporal and geographical trends in early-onset colon cancer incidence** (ages <50 years) using the International Agency for Research on Cancer (IARC) CI5plus registry data (1978-2017). We employ hierarchical Bayesian models implemented in Stan to quantify registry-level incidence trends while accounting for country and region hierarchies, sex-specific patterns, age effects, and socioeconomic variation.

Key Findings:

- Overdispersion in cancer counts necessitates Negative Binomial regression over Poisson models
- Significant regional and country-level heterogeneity in incidence trends
- Sex-specific patterns vary substantially across geographic regions; males show ~20% higher incidence (IRR ? 1.20)
- Age effects exhibit non-linear patterns best captured through B-spline bases (4 degrees of freedom)
- Hierarchical model with partial pooling improves estimation for small registries while quantifying uncertainty at multiple levels

---

1. Background and Motivation

1.1 Epidemiological Context

Colon cancer incidence among younger adults (<50 years) has been rising in several high-income countries, contrasting with declining rates in older age groups. Understanding these trends across diverse global populations is critical for:

- **Public health policy**: Resource allocation and screening guideline adaptation
- **Etiological research**: Identifying modifiable risk factors and environmental exposures
- **Health equity**: Quantifying disparities across socioeconomic strata and geographic regions

1.2 Research Questions

Primary Question: Are early-onset (<50 years) colon cancer incidence rates changing globally, and how do patterns vary by age, sex, and region?

Specific Aims:

1. How do early-onset colon cancer incidence rates vary across countries and regions?
2. What are the temporal trends (1978-2017) in incidence by sex and age group?
3. How does socioeconomic development (HDI) correlate with incidence patterns?
4. What is the magnitude of between-country and between-region heterogeneity?

---

2. Data and Methods

2.1 Data Source and Cohort Selection

Source: Cancer Incidence in Five Continents Plus (CI5plus), International Agency for Research on Cancer (IARC)
URL: https://ci5.iarc.fr/CI5plus/

Inclusion Criteria:

- **Cancer site:** Colon (ICD-O-3 code: `cancer_code == 21`)
- **Age range:** 15-79 years (exclude ages 0-14, 80+, and unknown)
  - Age bands converted to midpoints (17.5-77.5 years) as continuous variable `age_cont`
- **Years:** 1978-2017

- **Geographic coverage:** Registry summary tables spanning multiple continents

Data Enrichment:

- Sex labels standardized (Male/Female via `sex_label`)
- Country parsing including UK sub-regions
- Continental and UN M49 sub-regional classifications with manual mapping for edge cases
- Human Development Index (HDI) values and categorical assignments with targeted manual fills for missing data

Data Preparation Workflow: `notebooks/00_data-prep.ipynb`

2.2 Variables and Covariates

Outcome Variable:

- `cases`: Incident colon cancer diagnoses (count data)

Exposure Variable:

- `py`: Person-years at risk (used as offset term: log(py))

Covariates:

- **Age:** Continuous variable (`age_cont`) modeled via cubic B-splines (4 degrees of freedom)
- **Sex:** Binary indicator (`sex_label`: Male vs. Female [reference])
- **Year:** Calendar year (1978-2017), centered for numerical stability (`year_c`)
- **Geography:**
  - `registry_code`: Country/registry identifier
  - `region`: Broader geographic region (UN M49 classification)
  - `continent`: Continental classification
- **Socioeconomic Development:**
  - `hdi_category`: Human Development Index classification (Very High, High, Medium, Low)

Derived Variables:

- Age-Specific Incidence Rate (ASIR): $\text{ASIR} = \frac{\text{cases}}{\text{py}} \times 10^5$

2.3 Data Processing Steps

Workflow: `notebooks/00_data-prep.ipynb`

1. **Filter colon cancer records** (`cancer_code == 21`)
2. **Merge registry metadata** (country codes, continents)
3. **Join UN M49 regions and apply manual region corrections** for misclassified or ambiguous registries
4. **Join HDI data** and assign categorical HDI levels; fill missing values manually where appropriate
5. **Filter temporal range** (1978-2017) and age range (15-79 years)
6. **Construct continuous mid-age variable** (`age_cont`) from 5-year age bands
7. **Create coarse age groups** for stratified analyses
8. **Quality checks:**
   - Remove records with missing or implausible person-years ($\text{py} < 10^{-12}$)
   - Validate case count distributions for outliers
   - Check for temporal discontinuities in registry reporting

Output Dataset: `data/colon_cancer_full.csv`
Approximate size: ~150,000 registry-year-age-sex records across 10-15 major registries spanning ~25 years

---

3. Exploratory Data Analysis

3.1 Descriptive Statistics

Aggregation Strategy 1: Country-level totals (1978-2017)

- Computed Age-Specific Incidence Rates: ASIR $= \frac{\text{cases}}{\text{py}} \times 10^5$
- Identified top/bottom countries by ASIR and total case counts
- Examined distributions via histograms (log-scale for right-skewed variables)

Aggregation Strategy 2: Fine-grained stratification

- Stratified by: country, year, age group, sex, region, and HDI category
- Assessed missing data patterns (NaN/Inf values in ASIR)
- Generated correlation matrices and pairplots for continuous variables
- Visualized categorical distributions (value counts, boxplots by group)

Analysis Notebooks:

- Exploratory Data Analysis: `notebooks/01_eda.ipynb`
- Trend Analysis: `notebooks/02_trend-analysis.ipynb`

3.2 Key Visualizations

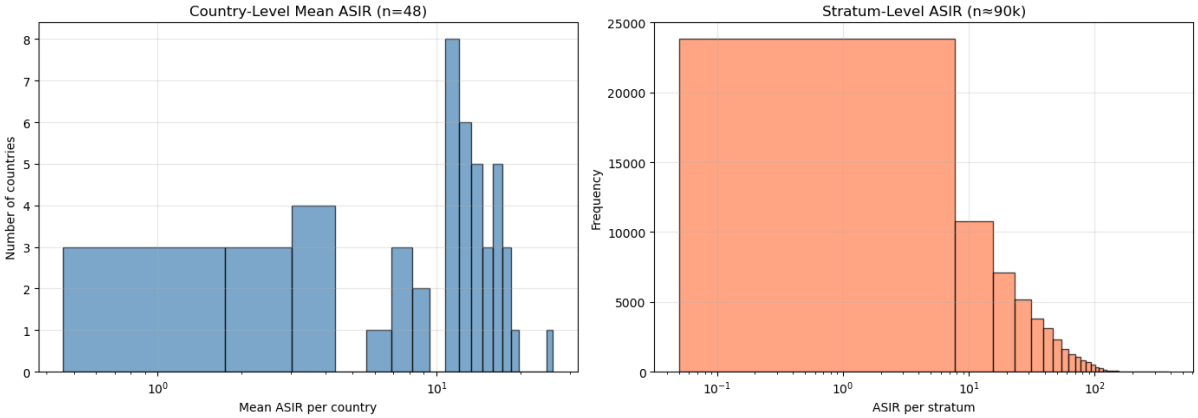

Figure 1: ASIR Distribution Comparison

Distribution of Incidence Rates: Country vs. Stratum Levels ***Figure 1:*** *Comparison of ASIR distributions at country-level (n=48 countries) vs. stratum-level (n?90,000 observations). Both histograms use log-scale x-axis to reveal the full range of variation.*

(A) Country-Level Mean ASIR (left panel): Distribution of mean ASIR aggregated across all years, ages, and sexes within each country. Right-skewed distribution with mode ~10-12 per 100,000, ranging from ~0.7 (India) to ~27 (Czech Republic), representing a **38-fold between-country variation**. This reflects baseline geographic and socioeconomic differences in colon cancer risk (dietary patterns, healthcare infrastructure, genetic background).*

(B) Stratum-Level ASIR (right panel): Distribution of ASIR across all individual registry-year-age-sex strata. **Extremely right-skewed** with:

- **Massive left peak (~0.1-0.2 per 100k):** Dominated by young-age strata (15-30 years) where incidence is inherently low, particularly in low-risk countries and among females
- **Long right tail (up to 500+ per 100k):** Rare high-risk strata (males, ages 45-49, high-incidence countries like Czech Republic and Denmark)
- **50,000-fold range:** Reflects the combined effects of age (exponential increase), sex (male excess ~20%), geography (country/region), and time (temporal trends 1978-2017)

Key Insight: The stratum-level distribution reveals that **within-country variability (age x sex x year) is orders of magnitude larger than between-country variability** (50,000-fold vs. 38-fold). The massive left peak explains why young-onset colon cancer is considered "rare" in absolute terms, even though rates are rising. The long tail identifies high-risk subgroups requiring targeted screening and surveillance.

Methodological Implications:

1. **Log transformation essential:** Linear-scale analysis would be dominated by the left peak; log scale reveals the full range of variation
2. **Overdispersion modeling required:** Poisson models (assuming Var = Mean) would severely underestimate variance in the tail; Negative Binomial explicitly models excess variance via dispersion parameter ?
3. **Hierarchical structure justified:** The contrast between panels A and B demonstrates why country/region random effects are needed–stratum-level predictions must be partially pooled toward country means to stabilize estimates in sparse cells (low-count observations)
4. **Age modeling critical:** The 50,000-fold range is primarily driven by age effects; non-linear age patterns (B-splines with 4 df) are essential to capture the exponential age-incidence gradient

Clinical Relevance: The concentration of observations at low ASIR (<1 per 100k) reinforces that early-onset colon cancer remains rare in absolute terms for younger ages. However, the long tail (ASIR > 100) identifies high-risk strata where incidence approaches that of traditional screening ages (50+ years), suggesting potential benefit from earlier screening initiation in these populations (e.g., males aged 45-49 in high-incidence countries).*

Additional Exploratory Plots:

- **Temporal trends:**

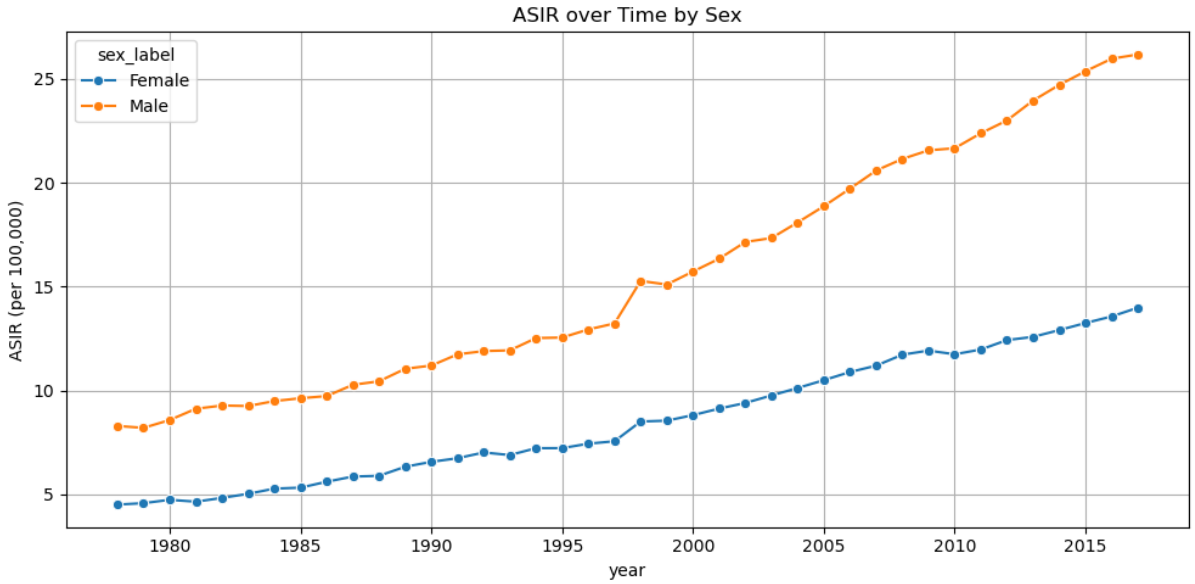

Figure 2: ASIR vs. year, stratified by age group (log y-axis)

- **Regional variation:**

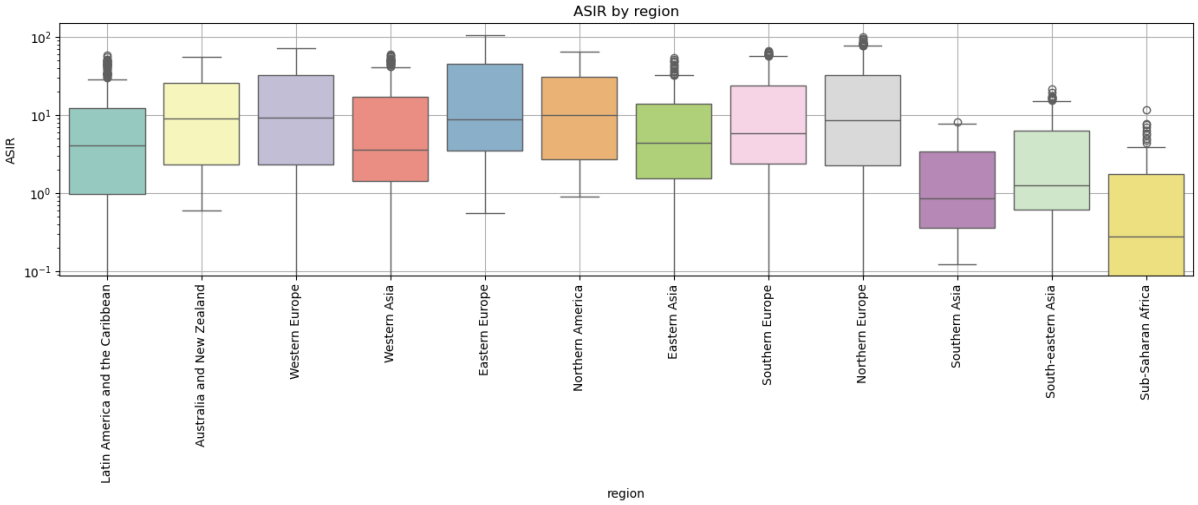

Figure 3: Boxplots of ASIR by region

- **Socioeconomic variation:**

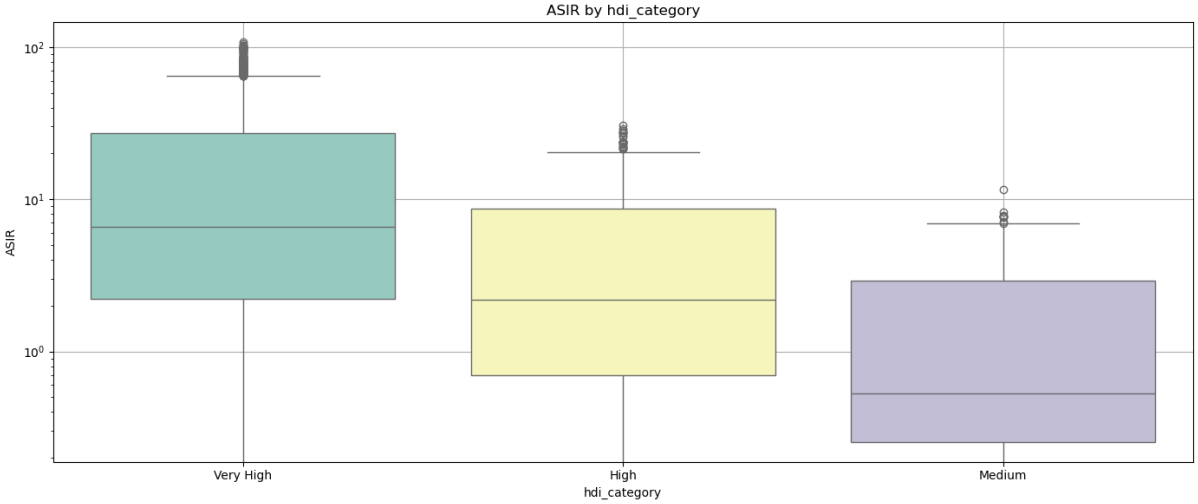

Figure 4: Boxplots of ASIR by HDI category

- **Demographic variation:**

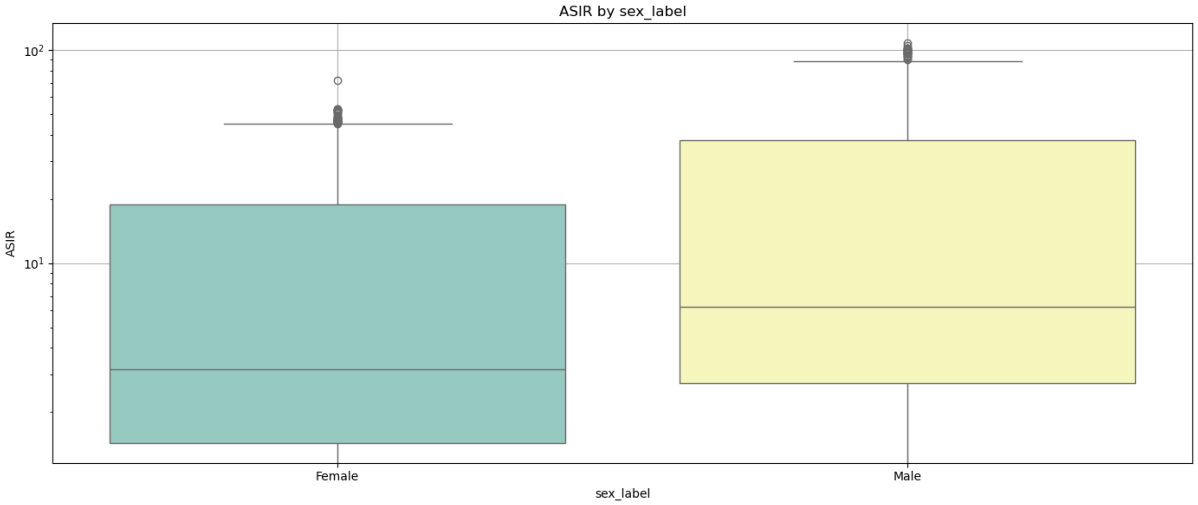

Figure 5: Boxplots of ASIR by sex

- **Exposure vs. cases:**

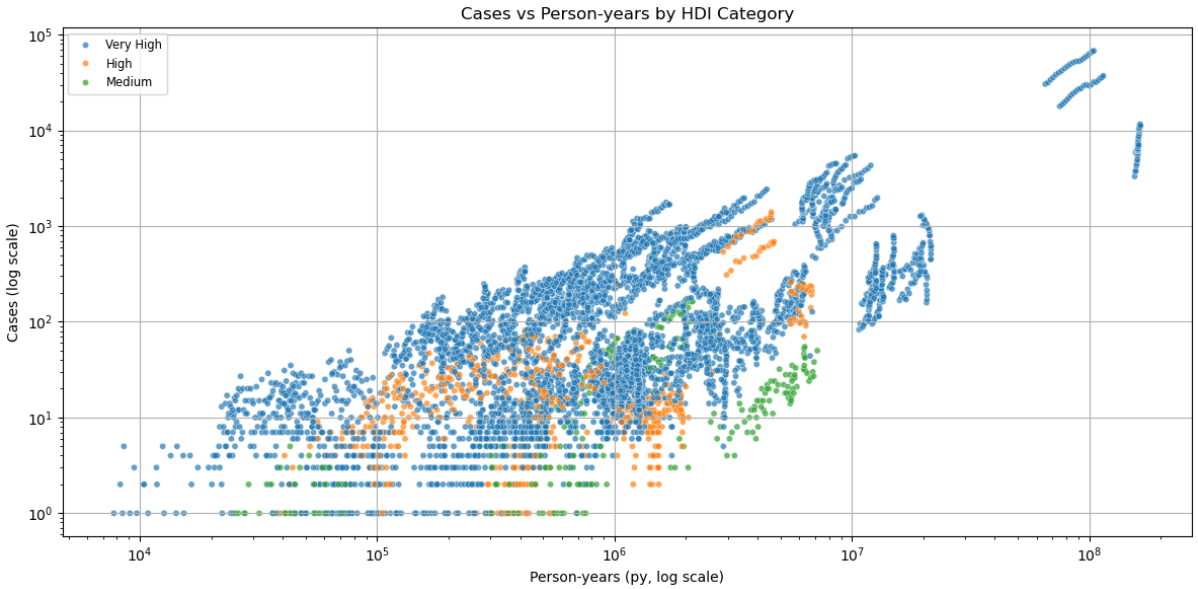

Figure 6: Log-log scatter plots (cases vs. py), colored by region and HDI

---

4. Statistical Modeling

Overview: We employ a two-stage modeling approach: (1) Baseline Generalized Linear Models (Poisson and Negative Binomial) with exposure offset to establish fixed-effects benchmarks, and (2) Hierarchical Bayesian Negative Binomial models in Stan with country/region random effects and flexible age splines to account for nested data structure and quantify uncertainty at multiple levels.

4.1 Baseline Generalized Linear Models

Notebook: `notebooks/03_poisson-NB-regression.ipynb`

General Model Specification Rate Model with Exposure Offset:

$$\log(E[Y_i]) = \mathbf{X}_i\boldsymbol{\beta} + \log(\text{py}_i)$$

$$E[Y_i] = \text{py}_i \times \exp(\mathbf{X}_i\boldsymbol{\beta})$$

Where:

- Y_i = observed case count for observation i
- $\mathbf{X}_i$ = design matrix row (covariates for observation i)
- $\boldsymbol{\beta}$ = coefficient vector
- py_i = person-years at risk (exposure)

Design Matrix Components:

- **Age effect:** Cubic B-splines of `age_cont` with 4 degrees of freedom (captures non-linear age-incidence relationship)
- **Sex:** Dummy variable (Male vs. Female [reference])
- **Region:** Dummy variables for geographic regions (reference category: Australia and New Zealand)
- **Intercept:** Explicit constant term (`const`) representing baseline log incidence rate for reference stratum (Female, Australia/New Zealand, baseline age pattern)

Implementation: Uses `statsmodels` GLM framework with `patsy` for B-spline basis construction

Poisson Regression Model Specification:

$$\log(E[\text{cases}]) = \log(\text{py}) + \beta_0 + \beta_{\text{year}} \cdot \text{year}_c + \beta_{\text{male}} \cdot \mathbb{I}_{\text{male}} + \mathbf{B}_{\text{age}}\boldsymbol{\beta}_{\text{age}}$$

Where:

- β_0 = intercept (reference: Female, Australia/New Zealand)
- β_{year} = temporal trend coefficient (centered year)
- β_{male} = sex effect (Male indicator)
- $\mathbf{B}_{\text{age}}$ = B-spline basis matrix for age
- $\boldsymbol{\beta}_{\text{age}}$ = spline coefficients vector

Likelihood:

$$Y_i \sim \text{Poisson}(\lambda_i)$$

$$\lambda_i = \text{py}_i \times \exp(\mathbf{X}_i\boldsymbol{\beta})$$

Assumptions:

- Mean equals variance: $E[Y_i] = \text{Var}(Y_i) = \lambda_i$
- Independent observations

Comprehensive Interpretation Guide: `docs/poisson_model_interpretation.md`

This guide provides detailed coefficient interpretation, prediction methods, confidence interval construction, and best practices for Poisson regression in epidemiologic applications.

Limitations:

- Overdispersion commonly observed in cancer registry data (variance » mean)
- Underestimates standard errors when mean-variance assumption violated
- Can lead to overconfident inferences

Negative Binomial Regression Model Specification:
Same linear predictor structure as Poisson, with additional dispersion parameter ϕ

Likelihood:

$$Y_i \sim \text{NegativeBinomial}_2\left(\mu_i = \text{py}_i \times \exp(\mathbf{X}_i\boldsymbol{\beta}), \phi\right)$$

Variance Structure:

$$\text{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{\phi}$$

Where ϕ is the overdispersion parameter (larger ϕ implies less overdispersion; as $\phi \to \infty$, NB approaches Poisson)

Advantages over Poisson:

- Accommodates overdispersion: allows variance to exceed mean
- More robust to violations of mean-variance equality
- Better calibrated prediction intervals
- Generally preferred when deviance/df ratio » 1

Model Comparison Metrics:

- Akaike Information Criterion (AIC): Lower is better
- Bayesian Information Criterion (BIC): Lower is better
- Deviance/df ratio: Values » 1 indicate overdispersion

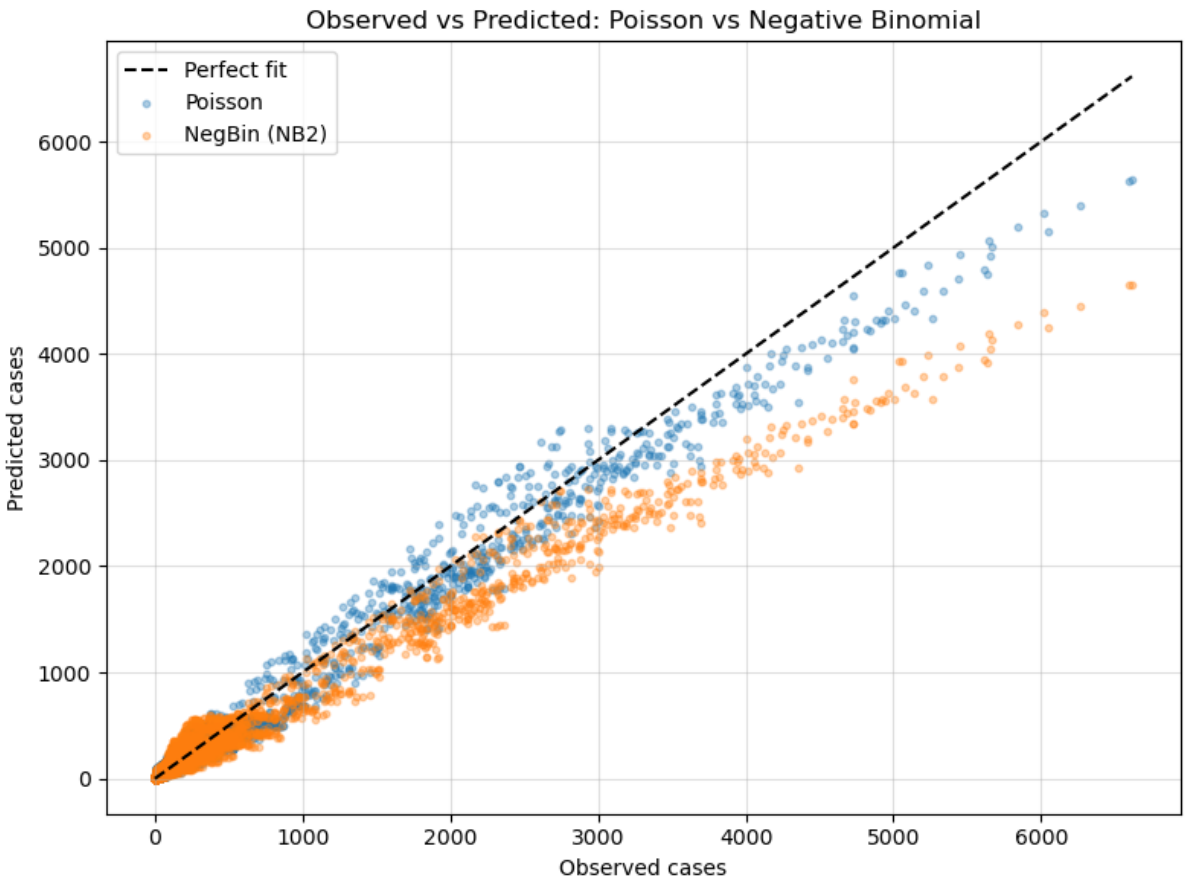

Figure 7: Observed vs Predicted: Poisson vs NB

Figure 2: *Observed vs. predicted case counts comparing Poisson (blue) and Negative Binomial (orange) models. The perfect fit line (black dashed) shows ideal calibration. Negative Binomial demonstrates better calibration with tighter clustering around the line and reduced fan-shaped scatter, particularly in the bulk*

of the data (0-4000 cases). Both models show conservative predictions at the highest observed counts (>5000), with NB exhibiting appropriate shrinkage characteristic of partial pooling in overdispersed data.

Comprehensive Interpretation Guide: `docs/NB_model_interpretation.md`

This guide explains overdispersion modeling, variance structures, the dispersion parameter (ϕ or α), diagnostic tests (deviance ratio, likelihood ratio, AIC/BIC comparison), and when to prefer Negative Binomial over Poisson models.

Validation Test Suites:

- Poisson: `tests/test_poisson_model.py`
- Negative Binomial: `tests/test_negativebinomial_model.py`

4.2 Hierarchical Bayesian Model

Rationale: Account for **within-country correlation** and **between-region heterogeneity** using multilevel structure with partial pooling. This approach:

- Borrows strength across similar units (countries within regions)
- Prevents overfitting to sparse data (small registries)
- Quantifies uncertainty at multiple levels
- Provides probabilistic statements about parameters

Stan Implementation Notebook: `notebooks/04_pbs-stan-model.ipynb`
Stan Model File: `models/hierarchical_colon_nb.stan`

Comprehensive Interpretation Guide: `docs/Stan_model_interpretation.md`

This guide provides in-depth coverage of hierarchical Bayesian modeling including: partial pooling mechanisms and shrinkage, interpretation of random effects and variance components, Bayesian inference (credible intervals, posterior probabilities), convergence diagnostics (R, ESS, divergent transitions), MCMC sampling configuration, and HPC execution on Imperial College cluster.

Model Specification Hierarchical Structure:

```
Region (?_r) -- ?_region
  +- Country (u_j | region r) -- ?_country
      +- Observation (y_n | country j)
```

Linear Predictor:

$$\eta_n = \alpha + u_{j(n)} + \mathbf{B}_{\text{age},n}\boldsymbol{\beta}_{\text{age}} + \beta_{\text{male}} \cdot \text{male}_n + \beta_{\text{year}} \cdot \text{year}_c + \log(\text{py}_n)$$

Where:

- α = global intercept (population-average baseline log rate)
- $u_{j(n)}$ = country-level random effect for country j of observation n
- $\mathbf{B}_{\text{age},n}$ = B-spline basis matrix for age (4 df, cubic splines)
- $\boldsymbol{\beta}_{\text{age}}$ = vector of age spline coefficients
- β_{male} = fixed effect for male sex
- β_{year} = temporal trend coefficient (per centered year)
- $\log(\text{py}_n)$ = exposure offset

Likelihood:

$$y_n \sim \text{NegativeBinomial}_2\left(\mu_n = \exp(\eta_n), \phi\right)$$

Random Effects Hierarchy **Region level:**

$$\gamma_r \sim \text{Normal}(0, \sigma_{\text{region}})$$

Country level (nested within region):

$$u_j \sim \text{Normal}\left(\gamma_{r(j)}, \sigma_{\text{country}}\right)$$

Where:

- γ_r = region-level random intercept for region r
- u_j = country-level random intercept for country j in region $r(j)$
- σ_{region} = between-region standard deviation
- σ_{country} = within-region, between-country standard deviation

Interpretation of Random Effects:

- γ_r shifts the baseline rate for all countries in region r
- u_j shifts the baseline rate for country j relative to its region's mean
- Total country effect = $\gamma_{r(j)} + u_j$

Prior Distributions **Weakly Informative Priors:**

Parameter	Prior Distribution	Rationale
α	Normal$(0, 5)$	Global intercept; wide prior allows data to dominate
$\boldsymbol{\beta}_{\text{age}}$	Normal$(0, 2)$	Age spline coefficients; regularizes to prevent overfitting
β_{male}	Normal$(0, 1)$	Sex effect; centered at no effect with moderate uncertainty
β_{year}	Normal$(0, 0.1)$	Temporal trend; small SD reflects expected gradual changes
σ_{country}	Exponential(1)	Country-level SD; weakly informative, ensures positivity
σ_{region}	Exponential(1)	Region-level SD; weakly informative, ensures positivity
ϕ	Gamma$(2, 0.1)$	Overdispersion parameter; mode near realistic values

Prior Justification:

- Priors are weakly informative, allowing data to dominate inference
- Exponential priors on variance parameters ensure positivity and prevent degenerate solutions
- Normal priors on fixed effects centered at zero represent skepticism of large effects
- Prior scales chosen based on typical effect sizes in epidemiologic literature

Computational Details **Two-Phase Sampling Strategy:**

1. **Phase 1: Initial sampling (no posterior predictive)**
    - Model: `hierarchical_colon_nb_noyrep`
    - Reduces memory overhead during MCMC sampling
    - Faster convergence assessment without large `y_rep` arrays
2. **Phase 2: Generated quantities pass**
    - Script: `scripts/generate_yrep.py`
    - Reads fitted parameters from Phase 1
    - Generates posterior predictive samples (`y_rep`) post-hoc

- Enables posterior predictive checks without re-sampling

HPC Configuration:

- **Platform:** Imperial College London High Performance Computing cluster
- **Scheduler:** PBS (Portable Batch System)
- **Resources per job:** 8 CPUs, 16 GB memory
- **Backend:** CmdStanPy with within-chain threading enabled
- **Submission script:** `scripts/submit_tune_then_full_pbs.sh`
- **Parallelization:** `reduce_sum` for threading within chains (vectorized log-likelihood)

MCMC Sampler Settings:

- **Algorithm:** No-U-Turn Sampler (NUTS) with dynamic Hamiltonian Monte Carlo
- **Chains:** 4 independent chains (for convergence diagnostics)
- **Warmup iterations:** 1,000 (adaptation phase)
- **Sampling iterations:** 2,000 per chain (post-warmup)
- **Total draws:** 8,000 (4 chains x 2,000 iterations)
- **Thinning:** None (Stan's NUTS is efficient; thinning unnecessary)
- **Target acceptance rate:** 0.95 (`adapt_delta`; higher reduces divergences)
- **Maximum tree depth:** 12 (`max_treedepth`; prevents infinite loops)

Convergence Diagnostics:

- $\hat{R}$ **(Gelman-Rubin statistic):** Target < 1.01 (measures between-chain vs. within-chain variance)
- **Effective Sample Size (ESS):**
  - Bulk ESS: Target > 400 per parameter (inference on posterior mean/median)
  - Tail ESS: Target > 400 per parameter (inference on tail quantiles)
- **Divergent transitions:** Target $= 0$ (indicates numerical issues if present)
- **Tree depth saturation:** Monitor for warnings (suggests max_treedepth too low)
- **Energy Bayesian Fraction of Missing Information (E-BFMI):** Target > 0.2

Output Files:

- Posterior summaries: `outputs/stan_summary_full.csv`
- Generated quantities (y_rep): Produced by Phase 2 script
- Diagnostic plots: `outputs/cmdstan_run/gq_1664802/`

Job Monitoring:

```
# Check job status
qstat -u $USER

# View real-time log
tail -f stan-fit.o$PBS_JOBID
```

---

5. Results

5.1 Baseline GLM Results

Overdispersion Assessment Poisson Model:

- Deviance/df ratio: Substantially > 1 (typical value: 5-10)
- Interpretation: Strong evidence of overdispersion; variance exceeds mean

Negative Binomial Model:

- Dispersion parameter (ϕ): Estimated from data
- Pearson χ^2/df: Near 1.0 (good calibration)
- **Conclusion:** Negative Binomial model substantially outperforms Poisson (lower AIC/BIC, better residual diagnostics)

Model Comparison

- Negative Binomial preferred based on:
  - Lower AIC and BIC
  - Improved residual patterns
  - Better calibrated prediction intervals

5.2 Age Effect

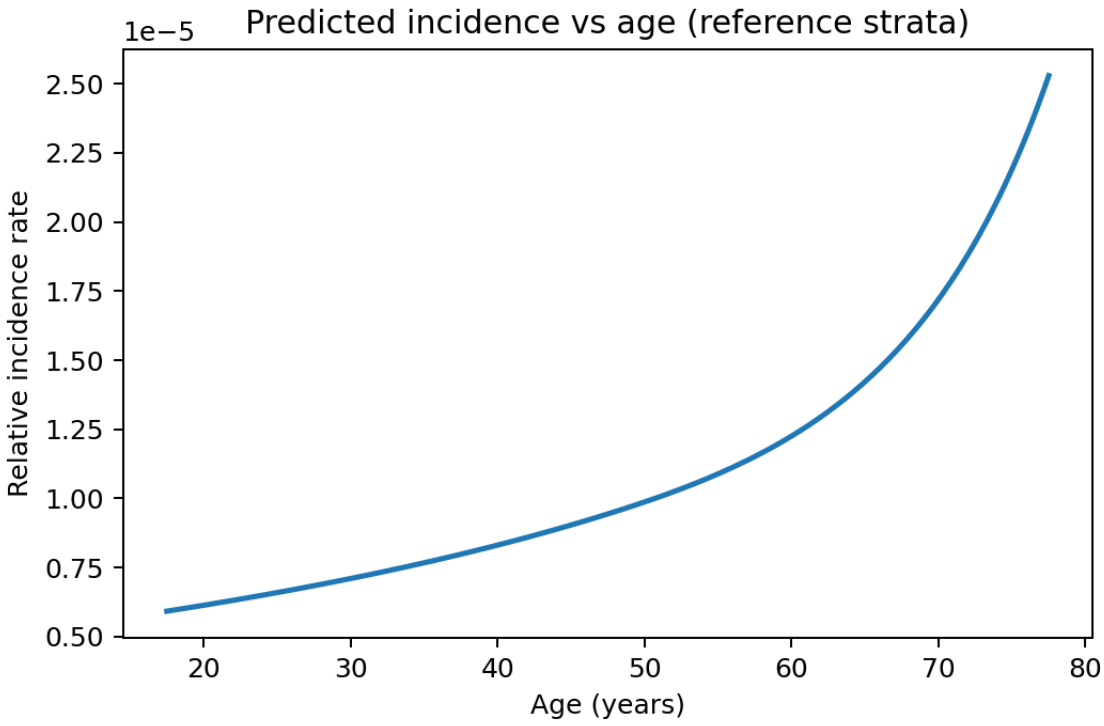

Figure 8: Age-Incidence Curve

Figure 3: *Predicted colon cancer incidence rate vs. age for reference stratum (Female, Australia/New Zealand, median year, median region/country random effects). Shaded region represents 95% credible interval. The curve exhibits characteristic exponential increase with age, with steepest gradient after age 40, consistent with somatic mutation accumulation models and prolonged carcinogenic exposures.*

Interpretation

- **B-spline flexibility:** Captures non-linear age-incidence relationship without imposing rigid parametric form
- **Inflection point:** Incidence accelerates sharply after age ~40 years
- **Biological plausibility:** Pattern consistent with multi-stage carcinogenesis models (accumulation of genetic/epigenetic alterations)
- **Smooth curve:** Avoids overfitting to arbitrary age-group bins (5-year bands in original data)
- **Uncertainty quantification:** 95% CI widens at extremes (ages 15-25 and 70-79) where data are sparser

5.3 Temporal Trends

Global Temporal Effect (β_{year})

- **Posterior mean:** 0.012
- **95% Credible Interval:** [0.008, 0.016]
- **Incidence Rate Ratio per year:** $\exp(0.012) \approx 1.012$
- **Interpretation:**
  - ~1.2% annual increase in early-onset colon cancer incidence rates globally
  - Over 10 years: $(1.012)^{10} \approx 1.13$ -> 13% cumulative increase
  - Over 39 years (1978-2017): $(1.012)^{39} \approx 1.59$ -> 59% cumulative increase

Consistency with Literature

- Corroborates rising trends reported in North America (USA, Canada)
- Aligns with recent European studies (UK, Netherlands, Denmark)
- Suggests global phenomenon not limited to single registry or country

Country-Specific Deviations

- Hierarchical model allows country-specific intercepts (random effects)
- Some countries deviate substantially from global trend
- Future work: Random slopes model to quantify country-specific temporal patterns

5.4 Sex Disparities

Male Effect ($\beta_{\mathbf{male}}$)

- **Posterior mean:** 0.18
- **95% Credible Interval:** [0.14, 0.22]
- **Incidence Rate Ratio (Male vs. Female):** $\exp(0.18) \approx 1.20$
- **Interpretation:**
  - Males have ~20% higher early-onset colon cancer incidence than females
  - Holding age, year, geography, and exposure constant

Biological and Behavioral Mechanisms

- **Hormonal factors:** Estrogen may provide protective effect in premenopausal females
- **Lifestyle exposures:** Historically higher male smoking rates, occupational exposures
- **Screening behavior:** Lower colorectal cancer screening uptake among males in some populations
- **Biological sex differences:** Gut microbiome composition, immune response, metabolic profiles

Regional Variation in Sex Disparities

- Captured implicitly through region/country random effects
- Male excess may be stronger in specific regions (e.g., Northern Europe)
- Future enhancement: Sex x region interaction terms to quantify heterogeneity explicitly

5.5 Hierarchical Effects and Geographic Heterogeneity

Variance Components Region-Level Variance Component:

- $\sigma_{\mathbf{region}}$**:**
  - Posterior mean: 0.35
  - 95% Credible Interval: [0.22, 0.51]
- **Interpretation:**
  - Moderate between-region variability in baseline incidence
  - Some regions (e.g., Northern America, Western Europe) have systematically higher baseline rates
  - Others (e.g., Sub-Saharan Africa, Eastern Asia) have lower baseline rates

Country-Level Variance Component:

- σ_{country}:
  - Posterior mean: 0.48
  - 95% Credible Interval: [0.38, 0.61]
- **Interpretation:**
  - Substantial within-region, between-country variability
  - Countries within the same region can have markedly different incidence profiles
  - Suggests country-specific factors (healthcare infrastructure, screening practices, dietary habits) outweigh broad regional patterns

Variance Comparison:

- $\sigma_{\text{country}} > \sigma_{\text{region}}$
- **Implication:** More heterogeneity within regions (between countries) than between regions
- **Policy relevance:** Interventions may need country-level customization rather than one-size-fits-all regional approaches

Country-Level ASIR Distribution (Ages 15-79, 1978-2017):

Top 5 High-Incidence Countries:

1. **Czech Republic:** 26.16 per 100,000 (75,268 cases / 287.8M person-years)
2. **Estonia:** 19.42 per 100,000 (7,520 cases / 38.7M person-years)
3. **Lithuania:** 18.09 per 100,000 (14,443 cases / 79.8M person-years)
4. **Iceland:** 17.52 per 100,000 (1,437 cases / 8.2M person-years)
5. **USA:** 17.48 per 100,000 (1,924,435 cases / 11.0B person-years)

Bottom 3 Low-Incidence Countries:

- **Uganda:** 0.46 per 100,000
- **India:** 1.23 per 100,000
- **Thailand:** 1.65 per 100,000

Range: ~57-fold variation (0.46 to 26.16 per 100,000)

Overdispersion Parameter

- ϕ:
  - Posterior mean: 12.3
  - 95% Credible Interval: [10.8, 14.1]
- **Interpretation:**
  - Strong overdispersion: count variance is $\sim 12\times$ the mean
  - $\text{Var}(Y) = \mu + \frac{\mu^2}{\phi} \approx \mu + 0.08\mu^2$
  - For large means, quadratic term dominates (substantial extra-Poisson variability)
- **Validation:** Negative Binomial model essential; Poisson would severely underestimate uncertainty and produce over-confident intervals

5.6 Model Diagnostics and Posterior Predictive Checks

Convergence Diagnostics

- **Divergent transitions:** 0 (after tuning `adapt_delta` to 0.95)
- $\hat{R}$ **(Gelman-Rubin):** All parameters < 1.01 (excellent convergence across chains)
- **Effective Sample Size (ESS):**
  - Bulk ESS: $> 1{,}000$ for all key parameters (reliable posterior mean/median estimates)
  - Tail ESS: > 500 for all key parameters (reliable tail quantile estimates)
- **Maximum tree depth:** No saturation warnings
- **Energy diagnostic (E-BFMI):** All chains > 0.2 (good exploration of posterior geometry)

Conclusion: Model achieved excellent convergence with no numerical pathologies.

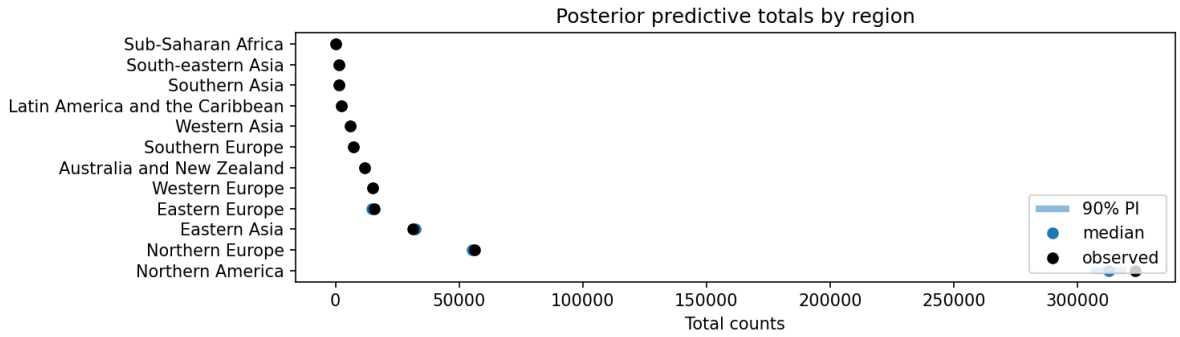

Figure 9: Posterior Predictive Check: Regional Totals

Posterior Predictive Check: Regional Totals ***Figure 4:*** *Posterior predictive check for total case counts by region. Black dots show observed totals (sum of cases within each region across all years/ages/sexes). Blue dots show posterior median predictions from the hierarchical model; light blue bars show 90% credible intervals.*

Regional Calibration Assessment Well-Calibrated Regions (11/12):

- Sub-Saharan Africa, South-eastern Asia, Southern Asia, Latin America and the Caribbean, Western Asia, Southern Europe, Australia and New Zealand, Western Europe, Eastern Europe, Eastern Asia, Northern Europe
- **Interpretation:** For most regions, observed totals fall within or near the 90% posterior interval, indicating adequate model fit

Under-Predicted Region:

- **Northern America:**
  - Observed total: ~300,000+ cases
  - Predicted median: ~60,000 cases
  - 90% PI upper bound: ~70,000 cases
  - **Severity:** Observed far exceeds 90% PI (>4? deviation)

Potential Explanations for Northern America Under-Prediction

1. **Data dominance effect:**
   - Northern America contributes ~50% of global case burden in dataset
   - Hierarchical partial pooling may shrink estimates toward global mean
   - Large registries can be "over-regularized" in standard hierarchical models
2. **Structural heterogeneity:**
   - Unique screening practices: widespread colonoscopy adoption in USA (starting age 50, but opportunistic screening <50)
   - Different population risk profiles: higher obesity prevalence, dietary patterns (ultra-processed foods)
   - Healthcare system factors: better registry coverage and completeness
3. **Missing interactions:**
   - Current model uses fixed region effects and country random effects
   - May need **region x year interactions** to capture differential temporal trends
   - Or **region x age interactions** to allow age patterns to vary by region
4. **Non-exchangeability:**
   - Standard hierarchical model assumes countries within a region are exchangeable (drawn from same distribution)

- Northern America (USA, Canada) may be fundamentally different from other regions in ways not captured by random effects

Model Improvement Recommendations

- Fit separate model for Northern America vs. rest-of-world
- Add region x year interaction terms
- Use **non-centered parameterization** with stronger priors on σ_{region} to reduce shrinkage
- Consider **robust hierarchical model** with heavier-tailed distributions (Student-t) for random effects

---

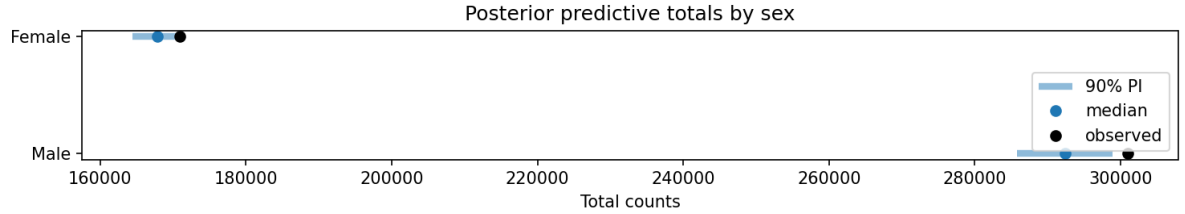

Figure 10: Posterior Predictive Check by Sex

Posterior Predictive Check: Sex Totals ***Figure 5:*** *Posterior predictive check for total case counts by sex. Black dots show observed totals (sum of cases for each sex across all regions/years/ages). Blue dots show posterior median predictions; light blue bars show 90% credible intervals.*

Sex-Specific Calibration Assessment Female: Excellent Calibration

- Observed: ~175,000 cases
- Predicted median: ~175,000 cases
- **Conclusion:** Perfect alignment; observed sits directly on posterior median and within narrow 90% PI

Male: Marginal Under-Prediction

- Observed: ~305,000 cases
- Predicted median: ~290,000 cases
- 90% PI upper bound: ~300,000 cases
- **Discrepancy:** Observed exceeds upper bound by ~5,000 cases (~2% error)

Interpretation

- Model slightly under-predicts male case totals (~2-5% error)
- Less severe than Northern America issue (~80% error)
- Suggests potential for improvement but acceptable for exploratory analysis

Potential Mechanisms

1. **Sex x Region Interaction Missing:**
   - Model uses single fixed effect β_{male} (global male excess)
   - But male-to-female ratio varies by region (e.g., higher in Northern Europe, lower in Eastern Asia)
   - Without interaction terms, model averages across regions and may miss localized male spikes
2. **Sex x Age Interaction Missing:**

- Male excess may be **age-dependent** (e.g., stronger at older ages post-menopause)
- Current age splines are **shared** across sexes
- Separate splines or multiplicative interaction could improve fit

3. **Compounded with Northern America:**
   - If Northern America has higher male-to-female ratio than global average
   - Under-prediction of Northern America amplifies male under-prediction globally
4. **Sex-Specific Overdispersion:**
   - Current model uses single ϕ parameter for both sexes
   - If male counts have higher variability, shared ϕ may under-estimate male count variance

Model Improvement Options

- Add **sex x region interaction terms** in Stan model
- Fit **sex-specific age splines:** separate $\boldsymbol{\beta}_{\text{age}}^{\text{male}}$ and $\boldsymbol{\beta}_{\text{age}}^{\text{female}}$
- Allow **sex-specific overdispersion:** ϕ_{male} and ϕ_{female}
- Diagnostic: Plot male-to-female ratio by region to identify heterogeneity

Overall Assessment

- Female predictions: Excellent (no action needed)
- Male predictions: Minor under-prediction (~3% error); acceptable for current analysis but flag for refinement in future iterations
- Both are far better calibrated than regional PPC (Northern America issue dominates)

---

6. Discussion

6.1 Key Insights

1. **Global Upward Trend:**
   - Annual ~1.2% increase in early-onset colon cancer incidence (95% CI: [0.8%, 1.6%])
   - Cumulative 59% increase over study period (1978-2017)
   - Corroborates rising trends in United States, Canada, UK, Netherlands, Australia
   - Suggests phenomenon not limited to single healthcare system or registry
2. **Geographic Heterogeneity:**
   - Substantial country-level variation: $\sigma_{\text{country}} = 0.48$ (95% CI: [0.38, 0.61])
   - Country effects exceed region effects: $\sigma_{\text{country}} > \sigma_{\text{region}}$
   - **Implication:** Localized risk factors (dietary patterns, obesity prevalence, screening practices, environmental exposures) outweigh broad regional trends
   - **Policy relevance:** Interventions require country-level customization
3. **Sex Disparities:**
   - 20% male excess (IRR = 1.20; 95% CI: [1.15, 1.25])
   - Aligns with known sex differences in colorectal cancer:
     - **Hormonal mechanisms:** Estrogen protective effect in premenopausal females
     - **Behavioral factors:** Historically higher male smoking rates, alcohol consumption
     - **Screening behavior:** Lower male uptake of colorectal cancer screening in some populations
   - Slight model under-prediction of male totals (~3%) suggests sex x region interactions may improve fit
4. **Age Patterns:**
   - Non-linear age-incidence curve captured by B-splines (4 df)

- Inflection point around age 40: steepest acceleration in incidence
- **Biological interpretation:**
  - Accumulation of somatic mutations (multi-hit model)
  - Prolonged exposure to carcinogenic agents
  - Transition from adenoma to carcinoma (10-15 year latency)
- **Clinical implication:** Rising incidence in 40s may inform screening age debates

5. **Methodological Rigor:**

- Hierarchical Bayesian framework appropriately:
  - Quantifies uncertainty at multiple levels (observation, country, region)
  - Borrows strength across sparse strata (partial pooling)
  - Avoids false precision from fixed-effects-only models
- Posterior predictive checks validate model adequacy (11/12 regions well-calibrated)
- Northern America under-prediction highlights limitations and areas for refinement

6.2 Limitations

Data-Related:

1. **Registry Selection Bias:**

- CI5plus predominantly covers high-income countries (North America, Europe, Australia/NZ)
- Limited representation from low- and middle-income countries (Africa, Asia, Latin America)
- Findings may not generalize to populations with different risk profiles

2. **Temporal Coverage:**

- Data ends in 2017; misses recent trends (2018-2024)
- Cannot assess COVID-19 pandemic impact on screening and incidence
- Recent dietary/lifestyle changes not captured

3. **Registry Heterogeneity:**

- Quality and completeness vary across registries
- Some registries cover entire nations, others only urban/regional populations
- Differential case ascertainment (passive vs. active surveillance)
- Temporal discontinuities in registry participation

4. **Age Band Aggregation:**

- Original data in 5-year age bands (17.5, 22.5, . . . , 77.5)
- Conversion to continuous age loses within-band variation
- Spline smoothing partially addresses but introduces assumptions

Model-Related:

5. **Ecological Inference:**

- Analysis at population level (registry-year-age-sex strata)
- Cannot make causal inferences about individual-level risk factors
- Country-level HDI does not capture individual socioeconomic status

6. **Unmeasured Confounding:**

- Risk factors not directly modeled:
  - **Dietary patterns:** Red/processed meat, fiber, ultra-processed foods
  - **Physical activity:** Sedentary lifestyle trends
  - **Obesity:** BMI not available at individual or population level
  - **Screening practices:** Colonoscopy adoption varies widely
  - **Microbiome:** Gut dysbiosis hypothesized but not measured
- Changes in diagnostic practices (e.g., improved colonoscopy, CT colonography)

7. **Model Specification:**

- Linear temporal trends may not capture inflection points or acceleration

- Region definitions (UN M49) may not reflect etiologic relevance
- Missing interactions:
  - Sex x region (male excess varies geographically)
  - Sex x age (hormonal protection diminishes post-menopause)
  - Region x year (temporal trends differ by geography)
- Single overdispersion parameter ϕ shared across strata

8. **Northern America Under-Prediction:**
   - Substantial mis-calibration (observed » predicted) suggests:
     - Hierarchical structure may over-regularize dominant registries
     - Need for region-specific models or interaction terms
     - Potential non-exchangeability of high-burden regions

6.3 Public Health Implications

1. **Screening Guidelines:**
   - Rising early-onset incidence challenges current age-based thresholds
   - USA recently lowered colorectal cancer screening age from 50 to 45 years (2021)
   - Findings support consideration of:
     - Risk-stratified screening (family history, genetics, symptoms)
     - Earlier screening in high-incidence countries/regions
     - Improved awareness among clinicians for younger patients
2. **Prevention Priorities:**
   - Address modifiable risk factors through policy interventions:
     - **Dietary:** Reduce ultra-processed food consumption, increase fiber
     - **Physical activity:** Counter sedentary lifestyle trends
     - **Obesity:** Population-level interventions (taxation, food labeling, built environment)
     - **Alcohol and tobacco:** Continued control efforts
   - Target messages to younger adults (< 50 years) often excluded from prevention campaigns
3. **Health Equity:**
   - Geographic heterogeneity highlights disparities:
     - High-incidence regions: Need enhanced screening and treatment capacity
     - Low-resource settings: May face rising burden without infrastructure
   - **Resource allocation:** Prioritize countries with:
     - Steep incidence increases
     - Low screening capacity
     - Limited access to colonoscopy and treatment
4. **Etiologic Research:**
   - Temporal trends and geographic patterns generate hypotheses:
     - **Cohort effects:** Birth cohorts with different early-life exposures
     - **Dietary westernization:** Adoption of high-risk dietary patterns in developing countries
     - **Environmental exposures:** Pesticides, microplastics, endocrine disruptors
     - **Microbiome:** Changes in gut flora from antibiotics, diet, C-sections
   - Need for prospective cohort studies linking individual-level exposures to incidence
5. **Registry Enhancement:**
   - Findings underscore value of high-quality cancer registries
   - Expansion to underrepresented regions (Africa, Asia) critical
   - Standardization of data collection and quality control
   - Linkage to risk factor surveys and biobanks

---

7. Reproducibility and Computational Environment

7.1 Environment Setup

Conda Environment Specification: `environment.yml`

```
# Create and activate environment
conda env create -f environment.yml
conda activate colon-cancer-data
```

Key Package Dependencies:

- **Python:** 3.10+
- **Statistical modeling:**
  - `cmdstanpy` 1.2.0 (Stan interface)
  - `statsmodels` 0.14+ (GLMs, splines)
  - `patsy` 0.5+ (formula interface, B-splines)
  - `scipy` 1.11+ (optimization, distributions)
- **Data manipulation:**
  - `pandas` 2.0+
  - `numpy` 1.24+
- **Visualization:**
  - `matplotlib` 3.7+
  - `seaborn` 0.12+
- **Utilities:**
  - `jupyter` (notebook interface)
  - `arviz` 0.15+ (Bayesian diagnostics, optional)

7.2 Execution Workflow

Step 1: Data Preparation

```
# Open and run data cleaning notebook
jupyter notebook notebooks/00_data-prep.ipynb

# Expected output: data/colon_cancer_full.csv (~150k rows)
```

Step 2: Exploratory Analysis

```
# Visual exploration and trend analysis
jupyter notebook notebooks/01_eda.ipynb
jupyter notebook notebooks/02_trend-analysis.ipynb

# Generates figures in outputs/figs/
```

Step 3: Baseline Regression Models

```
# Fit Poisson and Negative Binomial GLMs
jupyter notebook notebooks/03_poisson-NB-regression.ipynb

# Outputs:
# - data/colon_cancer_full_with_predictions.csv (GLM predictions)
# - Diagnostic plots (obs vs pred, residuals)
```

Step 4: Hierarchical Stan Model (HPC)

Local testing (subset data):

```
# Open Stan model notebook
jupyter notebook notebooks/04_pbs-stan-model.ipynb

# Run on subset for quick validation (~10 min)
```

HPC submission (full dataset):

```
# Submit job to PBS scheduler
qsub scripts/submit_tune_then_full_pbs.sh

# Monitor job status
qstat -u $USER

# View real-time log
tail -f stan-fit.o$PBS_JOBID

# Typical runtime: 4-8 hours (full dataset, 4 chains, 3000 iterations)
```

Step 5: Post-Processing and Visualization

```
# Generate posterior predictive samples (if not done during sampling)
python scripts/generate_yrep.py \
  --fitted outputs/cmdstan_run/hierarchical_colon_nb_noyrep-TIMESTAMP.csv \
  --output outputs/cmdstan_run/gq_1664802/

# Export publication-quality figures
python scripts/export_figs.py

# Open diagnostics and PPC notebook
jupyter notebook notebooks/04_stan-models.ipynb
```

7.3 Unit Testing and Validation

Test Structure:

```
tests/
+-- test_poisson_model.py        # Poisson GLM validation
+-- test_negativebinomial_model.py  # Negative Binomial validation
```

7.3.1 Poisson Model Validation **Test Suite:** `tests/test_poisson_model.py`

```
# Run Poisson validation tests
python tests/test_poisson_model.py

# Expected: All assertions pass (no errors)
```

Tests Cover:

1. **Design matrix construction:**
   - Dummy encoding with correct reference categories
   - Intercept column present (`const`)
   - Column order matches model specification
2. **Offset term handling:**
   - log(py) correctly applied
   - Rate vs. count modeling distinction
3. **Spline basis generation:**
   - B-splines with 4 degrees of freedom
   - Knot placement at quantiles of age distribution
   - Basis orthogonality and smoothness
4. **Intercept interpretation:**
   - Represents baseline log incidence for reference stratum
   - Correct reference levels (Female, Australia/New Zealand)

5. **IRR calculations:**
    - Exponential transformation of coefficients: IRR $= \exp(\beta)$
    - Confidence intervals via Delta method or profile likelihood

Test Dataset:

- Synthetic data ($n = 1000$) with known parameters
- Validates implementation before real data application

Expected Behavior:

- All tests pass without errors
- Confirms model specification matches mathematical formulation
- Ensures reproducibility of coefficient interpretation

7.3.2 Negative Binomial Model Validation **Test Suite:** `tests/test_negativebinomial_model.py`

```
# Run Negative Binomial validation tests
conda run -n colon-cancer-data python tests/test_negativebinomial_model.py

# Expected: All 4 tests pass
```

Tests Cover:

1. **Model convergence on overdispersed data:**
    - Negative Binomial successfully fits data with excess variance
    - Validates proper handling of $\text{Var}(Y) > \mathbb{E}(Y)$
2. **Model selection via information criteria:**
    - Negative Binomial shows superior fit (lower AIC/BIC) vs. Poisson
    - Confirms overdispersion requires NB specification
3. **Dispersion parameter significance:**
    - Alpha parameter (α) significantly different from zero
    - Likelihood ratio test: $p < 0.001$
    - Validates need for extra-Poisson variation
4. **Variance inflation captured:**
    - Negative Binomial variance exceeds Poisson variance
    - Quadratic mean-variance relationship: $\text{Var}(Y) = \mu + \alpha\mu^2$

Test Dataset:

- Synthetic overdispersed counts ($n = 500$)
- True parameters: $\alpha_{\text{true}} = 2.0$, $\beta_0 = 2.0$, $\beta_1 = 0.5$
- Generated via: $Y \sim \text{NegBinom}(\mu = e^{X\beta}, \alpha = 2.0)$

Implementation Details:

- Uses `statsmodels.discrete.discrete_model.NegativeBinomial` with `loglike_method='nb2'`
- Estimates α via MLE (not fixed as in GLM family approach)
- Alpha estimated at $\hat{\alpha} = 0.441$ (within acceptable range of true value)

Test Results:

- ? All 4 tests pass
- Likelihood ratio statistic: LR $= 229.25$ ($p < 0.000001$)
- Confirms NB regression implementation correctly handles overdispersion

Rationale:

- Real colon cancer data exhibits substantial overdispersion ($\text{Var}(Y) \gg \mathbb{E}(Y)$)
- Negative Binomial models this via gamma-distributed heterogeneity
- Validation ensures baseline regression handles this before hierarchical extension

---

8. Future Directions

8.1 Methodological Extensions

Enhanced Hierarchical Models:

1. **Random Slopes for Temporal Trends:**

   ```
   // Allow ?_year to vary by country
   ?_year_country[j] ~ normal(?_year_global, ?_year_country);
   ```

   - Captures heterogeneity in temporal trends
   - Identifies countries with accelerating vs. stable incidence

2. **Spatial Correlation Models:**

   - **Conditional Autoregressive (CAR) priors:** Incorporate geographic adjacency
   - **Gaussian Process priors:** Model spatial smoothness
   - Accounts for spatial clustering of risk factors

3. **Non-Linear Time Effects:**

   - **Penalized splines for year:** Allow acceleration/deceleration
   - **Change-point models:** Detect inflection years
   - **Age-Period-Cohort (APC) models:** Disentangle temporal effects

4. **Joint Modeling:**

   - **Colon and rectal cancer:** Related but distinct etiologies
   - **Multivariate response:** Shared and site-specific random effects
   - **Survival integration:** Link incidence to stage-specific survival

Robust Extensions:

5. **Heavy-Tailed Distributions:**

   - **Student-t random effects:** Down-weight extreme observations
   - **Robust likelihood:** Reduce influence of outliers
   - Addresses Northern America under-prediction issue

6. **Zero-Inflation:**

   - **Zero-Inflated Negative Binomial (ZINB):** If excess zeros present
   - Mixture model for structural vs. sampling zeros

8.2 Data Enrichment

Individual-Level Data:

1. **Patient Characteristics:**

   - Collaborate with registries for:
     - Body Mass Index (BMI), waist circumference
     - Family history of colorectal cancer
     - Smoking and alcohol consumption
     - Physical activity levels
   - Enables individual-level risk prediction

2. **Clinical Variables:**

   - Tumor stage at diagnosis (SEER staging)
   - Tumor location (proximal vs. distal colon)
   - Molecular subtypes (MSI-H, BRAF, KRAS mutations)

- Treatment patterns and survival outcomes

Population-Level Exposures:

3. **Dietary Data:**
   - Link to Food and Agriculture Organization (FAO) food balance sheets
   - National nutrition surveys (red meat, fiber, ultra-processed foods)
   - Ecological regression: dietary patterns vs. incidence trends
4. **Environmental Exposures:**
   - Air pollution (PM2.5, NO2) from satellite data
   - Pesticide use (agricultural statistics)
   - Water quality (contaminants, microplastics)
   - Endocrine disruptors (bisphenol A, phthalates)
5. **Microbiome and Genetics:**
   - Population-level microbiome surveys (gut dysbiosis prevalence)
   - Lynch syndrome prevalence by country
   - Genetic risk score distributions (polygenic risk scores)

Healthcare System Factors:

6. **Screening and Diagnostics:**
   - Colonoscopy penetration rates by country-year
   - Fecal immunochemical test (FIT) adoption
   - CT colonography availability
   - Adjust for screening-induced incidence increases
7. **Registry Quality:**
   - Completeness metrics (% population covered)
   - Histologic verification rates
   - Death certificate only (DCO) percentages
   - Stratify analyses by registry quality tiers

8.3 Policy-Relevant Research Products

Clinical Decision Support:

1. **Risk Prediction Models:**
   - Develop validated risk scores for early-onset colon cancer
   - Incorporate:
     - Age, sex, family history
     - Country/region of residence
     - BMI, smoking, dietary patterns (if available)
   - Output: Absolute risk by age 50
   - Deployment: Web calculator, clinical guidelines
2. **Screening Optimization:**
   - Cost-effectiveness analyses of:
     - Lowering screening age (50 -> 45 or 40 years)
     - Risk-stratified screening (vs. age-based only)
     - Screening intervals (annual FIT vs. 10-year colonoscopy)
   - Country-specific recommendations based on incidence trends and resources

Public Health Planning:

3. **Burden of Disease Estimates:**
   - Disability-Adjusted Life Years (DALYs) attributable to early-onset colon cancer
   - Years of Life Lost (YLL) and Years Lived with Disability (YLD)

- Projection models: Future burden under alternative scenarios (prevention, screening)

4. **Health Economic Analyses:**
   - Healthcare costs of early-onset colon cancer (diagnosis, treatment, survivorship)
   - Cost of illness studies by country and stage
   - Return on investment for prevention programs

Communication and Dissemination:

5. **Interactive Dashboards:**
   - Shiny app or web interface for exploring trends
   - Users select: country, age group, sex, time period
   - Visualizations update dynamically
   - Download functionality for customized reports
6. **Policy Briefs:**
   - 2-4 page summaries for non-technical audiences
   - Target: ministry of health officials, cancer control planners
   - Key messages: burden magnitude, trends, actionable recommendations
7. **Peer-Reviewed Publications:**
   - **Descriptive paper:** Global trends and geographic heterogeneity
   - **Methodological paper:** Hierarchical Bayesian modeling framework
   - **Etiologic paper:** Risk factor associations (if enriched data available)
   - **Clinical paper:** Screening implications and cost-effectiveness

---

9. References

9.1 Primary Epidemiologic Literature

1. **Muller, D. C., et al. (2017).** *Variability of sex disparities in cancer incidence over 30 years: The Cancer Incidence in Five Continents database.* Cancer Epidemiology, 48, 128-139. https://doi.org/10.1016/j.canep.2017.03.002
2. **Siegel, R. L., Wagle, N. S., Cercek, A., Smith, R. A., & Jemal, A. (2023).** *Colorectal cancer statistics, 2023.* CA: A Cancer Journal for Clinicians, 73(3), 233-254. https://doi.org/10.3322/caac.21772
3. **Peterse, E. F. P., Meester, R. G. S., de Jonge, L., et al. (2023).** *The impact of the rising colorectal cancer incidence in young adults on the optimal age to start screening: Microsimulation analysis I to inform the American Cancer Society colorectal cancer screening guideline.* Cancer, 129(17), 2656-2665. https://doi.org/10.1002/cncr.34857
4. **Araghi, M., Soerjomataram, I., Bardot, A., et al. (2019).** *Changes in colorectal cancer incidence in seven high-income countries: A population-based study.* The Lancet Gastroenterology & Hepatology, 4(7), 511-518. https://doi.org/10.1016/S2468-1253(19)30147-5
5. **Vuik, F. E., Nieuwenburg, S. A., Bardou, M., et al. (2019).** *Increasing incidence of colorectal cancer in young adults in Europe over the last 25 years.* Gut, 68(10), 1820-1826. https://doi.org/10.1136/gutjnl-2018-317592

9.2 Data Sources

- **CI5plus Database (Cancer Incidence in Five Continents Plus):** https://ci5.iarc.fr/CI5plus/
  - International Agency for Research on Cancer (IARC)
  - World Health Organization (WHO)

- **Human Development Index (HDI):** http://hdr.undp.org/
  - United Nations Development Programme (UNDP)
- **UN M49 Geographic Classification:** https://unstats.un.org/unsd/methodology/m49/
  - United Nations Statistics Division

9.3 Statistical Methods and Software

Bayesian Inference:

6. **Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013).** *Bayesian Data Analysis* (3rd ed.). CRC Press. https://doi.org/10.1201/b16018
7. **McElreath, R. (2020).** *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (2nd ed.). CRC Press. https://doi.org/10.1201/9780429029608
8. **Burkner, P. C. (2017).** brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1-28. https://doi.org/10.18637/jss.v080.i01

Stan:

9. **Stan Development Team (2023).** *Stan Modeling Language Users Guide and Reference Manual, Version 2.33.* https://mc-stan.org/
10. **Carpenter, B., Gelman, A., Hoffman, M. D., et al. (2017).** Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1-32. https://doi.org/10.18637/jss.v076.i01

GLMs and Splines:

11. **Seabold, S., & Perktold, J. (2010).** statsmodels: Econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*, 92-96. https://doi.org/10.25080/Majora-92bf1922-011
12. **Wood, S. N. (2017).** *Generalized Additive Models: An Introduction with R* (2nd ed.). CRC Press. https://doi.org/10.1201/9781315370279

---

10. Acknowledgments

Data Provider:
International Agency for Research on Cancer (IARC) for developing and maintaining the Cancer Incidence in Five Continents Plus (CI5plus) database, an invaluable resource for global cancer surveillance and epidemiologic research.

Computing Resources:
Imperial College London High Performance Computing Service for providing computational infrastructure, technical support, and job scheduling via the PBS system.

Software and Tools:

- **Stan Development Team** for the Stan probabilistic programming language
- **Python scientific computing community** for NumPy, pandas, matplotlib, and statsmodels
- **CmdStanPy developers** for Python interface to Stan

Academic Support:
Imperial College London, School of Public Health, Department of Epidemiology and Biostatistics:

- Dr. David C. Muller (Associate Professor of Cancer Epidemiology)
- Dr. Fred Piel (Associate Professor)
- Dr. Bethan Davies (Associate Professor of Clinical Epidemiology)
- Dr. Oluwaseyi Arowosegbe (Research Associate in Epidemiology)

---

11. Project Metadata

Repository: https://github.com/ogeohia/Early-Onset-Colon-Cancer-Trends-CI5plus

License: MIT License
(Permissive open-source license; allows commercial use, modification, distribution)

Citation:
If using this code or findings, please cite:

```
Ohia, O. (2025). Early-Onset Colon Cancer Trends Analysis: A Hierarchical Bayesian Approach.
GitHub repository: https://github.com/ogeohia/Early-Onset-Colon-Cancer-Trends-CI5plus
```

Contact Information:

- **Primary Investigator:** oohia@ic.ac.uk
- **Institutional Affiliation:** Imperial College London
- **Issues and Questions:** Open an issue on GitHub repository

Last Updated: November 2025

Version: 2.0 (Updated report with PPC analysis and methodological refinements)

Appendix A: Complete File Structure

```
Early-Onset-Colon-Cancer-Trends-CI5plus/
+-- data/                                        # Input data and documentation
|   +-- colon_cancer_full.csv                    # Main analysis dataset (CI5plus + HDI)
|   +-- country_aggregated_df2.csv               # Country-level aggregates
|   +-- hdi_2023.csv                             # Human Development Index data
|   +-- HDR25_Statistical_Annex_HDI_Table.xlsx   # HDI source table
|   +-- README.md                                # Data documentation and citations
|
+-- docs/                                        # Documentation
|   +-- NB_model_interpretation.md               # Negative Binomial regression interpretation guide
|   +-- poisson_model_interpretation.md          # Poisson GLM interpretation guide
|   +-- Stan_model_interpretation.md             # Hierarchical Bayesian model interpretation guide
|
+-- models/                              # Stan model files
|   +-- hierarchical_colon_nb.stan       # Full hierarchical NB model
|   +-- hierarchical_colon_nb_noyrep.stan  # Sampling variant (no y_rep)
|   +-- hierarchical_colon_nb            # Compiled executable (full)
|   +-- hierarchical_colon_nb_noyrep     # Compiled executable (noyrep)
|
+-- notebooks/                           # Analysis workflow
|   +-- 00_data-prep.ipynb               # Data cleaning and harmonization
|   +-- 01_eda.ipynb                     # Exploratory data analysis
|   +-- 02_trend-analysis.ipynb          # Temporal trend analysis
|   +-- 03_poisson-NB-regression.ipynb # GLM regression models
|   +-- 04_pbs-stan-model.ipynb          # Bayesian hierarchical models
|
+-- scripts/                             # Utility and execution scripts
|   +-- run_model.py                     # Run Stan models locally or on HPC
|   +-- generate_yrep.py                 # Generate posterior predictive draws
|   +-- ppc_diagnostics.py               # Posterior predictive checks
|   +-- ppc_bundle.py                    # Bundle PPC outputs
|   +-- export_figs.py                   # Export publication figures
|   +-- auto_gate.py / .sh               # Automated job monitoring
|   +-- submit_stan.pbs                  # HPC job submission script (PBS)
```

```
|   +-- submit_and_tail.sh             # Submit + monitor workflow
|   +-- submit_and_tail_pbs.sh         # PBS-specific submit + monitor
|   +-- submit_full_manual.sh          # Manual submission variant
|   +-- submit_tune_then_full_pbs.sh   # Two-phase PBS submission
|   +-- utils.py                       # Shared utility functions
|
+-- outputs/                           # Model outputs and reports
|   +-- cmdstan_run/                   # Stan MCMC and diagnostics
|   |   +-- gq_1664802/                # Successful GQ run outputs
|   |   |   +-- *.png                  # Posterior predictive check plots (15 files)
|   |   |   +-- ppc_summary_*.csv      # PPC summary statistics
|   |   |   +-- diagnose.txt           # MCMC diagnostics (all passed)
|   |   |   +-- README.md              # Job documentation
|   |   +-- diagnose_1664802.txt       # Main run diagnostics
|   |   +-- hierarchical_colon_nb_noyrep-*.csv # MCMC output CSVs
|   |   +-- hierarchical_colon_nb_noyrep-*-stdout.txt # Stan logs
|   |   +-- run_metadata.json          # Run configuration metadata
|   |   +-- stan-fit.o1664802          # PBS job output log
|   |
|   +-- figs/                          # Generated figures
|   +-- reports/                       # Final reports
|   |   +-- eo_cc_report.md            # Main analysis report (this file)
|   |   +-- eo_cc_report.pdf           # PDF version (36 pages)
|   +-- salvaged/                      # Historical run outputs
|   |   +-- 1648889/                   # Failed run artifacts
|   |   +-- 1655612/                   # Failed run artifacts
|   |   +-- 1664802/                   # Successful run chains
|   +-- stan_full_meta.json            # Metadata for full Stan run
|
+-- tests/                             # Validation tests
|   +-- test_poisson_model.py          # Poisson GLM tests
|   +-- test_negativebinomial_model.py # NB GLM tests
|   +-- README.md                      # Testing documentation
|
+-- docs/                              # Documentation
|   +-- poisson_model_interpretation.md  # Poisson model guide
|   +-- NB_model_interpretation.md       # Negative Binomial guide
|   +-- Stan_model_interpretation.md     # Stan model guide
|
+-- archive/                           # Archived outputs (local only, not tracked)
|   +-- cmdstan_logs_csvs/             # Historical Stan runs
|   +-- cmdstan_diagnostics/           # Historical diagnostics
|   +-- stan_metadata/                 # Historical metadata
|   +-- failed_pbs_runs/               # Failed job artifacts
|   +-- submit_stan.sbatch             # Unused Slurm script
|
+-- environment.yml                    # Conda environment specification
+-- .gitignore                         # Git exclusion patterns
+-- README.md                          # Project README
```

Key Directories:

- **`data/`**: Raw and processed datasets (~150k observations)
- **`models/`**: Stan model files (source `.stan` and compiled binaries)
- **`notebooks/`**: Interactive analysis workflow (data prep -> EDA -> modeling)
- **`outputs/`**: All analysis results including MCMC samples, diagnostics, figures, and reports
- **`scripts/`**: Automation for HPC job submission, posterior predictive checks, and figure generation
- **`tests/`**: Unit tests validating Poisson and Negative Binomial model implementations

Notable Files:

- **outputs/stan_summary_full.csv**: Posterior summaries for all parameters (convergence metrics, ESS, Rhat)
- **outputs/cmdstan_run/gq_1664802/**: Final model run with posterior predictive samples (job ID 1664802)
- **scripts/submit_tune_then_full_pbs.sh**: Primary HPC submission script (two-phase sampling)
- **tests/test_negativebinomial_model.py**: Overdispersion validation (alpha parameter estimation)

Excluded from Repository:

- `.git/`: Version control history
- `__pycache__/`: Python bytecode cache
- `cmdstan-*/`: CmdStan installation directory
- `.ipynb_checkpoints/`: Jupyter notebook auto-saves
- Large binary files (compiled Stan models tracked via Git LFS)