Characterizing Oncology Patient Journeys and Health State Transitions using a Data-Driven Markov Transition Matrix in Large-scale Electronic Health Records

#MSR68

Youngwon Kim [1], Wilson Lau [1], Ehsan Alipour [1], Sihang Zeng [2], Anand Oka [1]
[1]Truveta Inc, Bellevue, WA, [2]Univeristy of Washington, Seattle, WA

Background

Existing knowledge

- Oncology care trajectories are highly heterogeneous and non-linear, with patients transitioning across multiple comorbid and symptom states over time, making longitudinal characterization challenging in real-world data
- Most EHR-based oncology studies rely on predefined feature sets or disease-specific cohorts, which may miss latent, cross-domain comorbidity patterns and temporal dependencies present in large-scale longitudinal data
- Markov models are widely used in health economics and outcomes research, but are typically parameterized using simplified or assumption-driven transition structures rather than empirically derived real-world trajectories
- Data-driven Markov Transition Matrices (MTMs) enable unbiased estimation of state transitions directly from EHR data, offering a scalable framework to capture condition persistence and comorbidity without prior feature selection

Objective

- To evaluate the descriptive utility, face validity, and scalability of MTMs for characterizing condition persistence and comorbidity dynamics in a large cancer population and inform parameterization of Markov models

Methods

Data

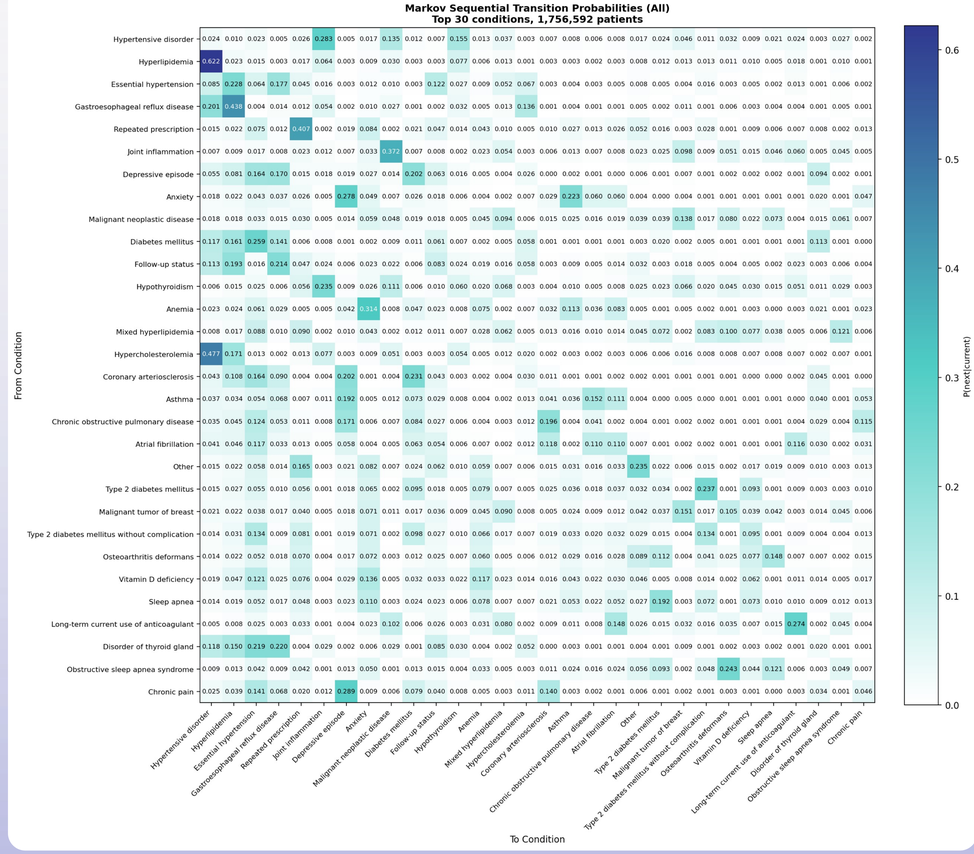
- Real-world US electronic health record (EHR) data (Truveta Research database): aggregated, normalized, de-identified)
- ~1.76M cancer patients in the U.S.
- Data included conditions and demographics.
- Index event: first cancer diagnosis; 24-month lookback window

Markov Transition Matrix

- States: Top 30 conditions by patient count
- Conditions outside top 30 → chain-breaking events (preserves temporal semantics)
- Counted adjacent pairs ($C_i \rightarrow C_{i+1}$), including self-transitions
- $P(j \mid i) = \text{count}(i \rightarrow j) / \Sigma_k \text{count}(i \rightarrow k)$ (row-normalization)

Analysis

- Persistence (self-transitions), Comorbidity (cross-transitions)
- diagonal = persistence; off-diagonal = comorbidity progression



Patient characteristics by group

Characteristic	n (%)
Age (years), mean ± SD (median)	68.1 ± 17.0 (70)
Female sex, n (%)	1,020,313 (58.1%)
Race, n (%)	
White	1,399,011 (79.6%)
Black or African American	153,398 (8.7%)
Asian	47,459 (2.7%)
Other or Unknown Race	156,737 (9.0%)
Ethnicity, n (%)	
Hispanic or Latino	110,845 (6.3%)
Not Hispanic or Latino	1,645,761 (93.7%)
Marital status, n (%)	
Married	854,805 (48.7%)
Unmarried	241,689 (13.8%)
Widowed	155,012 (8.8%)
Other / Unknown	504,100 (28.7%)

Results

- Low overall persistence (mean ~0.07) with markedly higher stability in chronic/treatment states (repeated prescription 0.41, anticoagulant use 0.27, breast cancer 0.15, diabetes 0.13, AF 0.11)
  → symptom-level conditions remain highly transient (~0.01–0.02)
- Cardiometabolic conditions form the dominant subnetwork: Dense, high-probability transitions among hypertension, diabetes, and lipid disorders (e.g., diabetes → hypertension 0.26; CAD → diabetes 0.23)
- Largest transition magnitudes reflect care processes, not biology (e.g., hyperlipidemia → hypertensive disorder 0.62) → likely driven by shared screening, co-documentation, and coding practices
- Depression emerges as a global downstream "sink": Strong inflow from multiple conditions: chronic pain (0.29), anxiety (0.28), asthma (0.19), COPD (0.17) → reverse transitions minimal, indicating consequence rather than driver
- Marked directional asymmetry reveals disease trajectories: Transitions are not reciprocal (e.g., anxiety → depression ≫ reverse) → distinguishes progression from simple co-occurrence
- Respiratory conditions bridge cardiometabolic and mental health systems: Asthma → COPD (0.15), COPD → coronary disease (0.20), asthma → AF (0.11)
  → consistent with shared risk factors and systemic burden
- Atrial fibrillation acts as a multimorbidity hub in later disease stages: Distributes broadly to anticoagulation, coronary disease, hypertension, and COPD (~0.11–0.12)
- High-connectivity "hub" nodes reflect healthcare utilization (GERD, hypertensive disorder, joint inflammation)
  → broad outgoing transitions suggest systemic comorbidity and encounter-driven coding
- Cancer diagnoses remain relatively self-contained Higher persistence and limited cross-cluster spread
  → large transitions (e.g., joint inflammation → cancer 0.37) likely reflect diagnostic pathways

Conclusions

- Data-driven MTMs provide an interpretable, scalable framework for modeling condition persistence and comorbidity dynamics
- Recover clinically coherent multimorbidity patterns (cardiometabolic, respiratory, psychosomatic, oncologic)
- Directional asymmetry distinguishes progression from co-occurrence, with depressive episodes emerging as a downstream convergence state of chronic disease burden
- MTMs capture both biological relationships and healthcare utilization signals, providing a reproducible foundation for parameterizing Markov models in cost-effectiveness and outcomes research

Large-scale EHR-derived Markov transition matrices reveal stable oncologic trajectories and hidden comorbidity networks without prior feature selection..

Disclosures: All authors are employees of Truveta Inc.