Modeling How Diabetes Progresses

From Random Individual Paths to Population-Level Predictions

mV01.00

Gergely Pósfai[1] Mantana Chudtong[2] Gareth W. Peters[3] Andrea De Gaetano[1,2,4]

[1]Dept. of Biomatics and EKIK, Óbuda University, Budapest, Hungary
[2]Dept. of Mathematics, Mahidol University, Bangkok, Thailand
[3]Dept. of Statistics, University of California, Santa Barbara - USA
[4]Consiglio Nazionale delle Ricerche, Istituto per l'Analisi dei Sistemi e Informatica (CNR-IASI), Rome , Italy

September 8, 2026 Óbuda University Minisymposium

Outline

Why model chronic disease progression?

- Chronic diseases like Type 2 diabetes (T2DM) evolve through clinically meaningful stages, with **random** transition times between stages.
- Public health planning, cost-effectiveness analysis, and insurance/actuarial risk pricing all require a quantitative model of how a population moves between health states over time.
- Two complementary modeling languages exist:
  - **Stochastic**: individual-level random processes (Markov chains).
  - **Deterministic**: population-level differential equations (ODEs).
- This work builds a bridge between the two, and calibrates it using real epidemiological survey data (NHANES) — overcoming the difficulty of using cross-sectional data to infer a longitudinal process.

The core problem, and our approach

1. Represent T2DM progression as a **continuous-time Markov chain** (CTMC) on clinical states, with a deterministic **ODE** counterpart at the population level (a "fluid limit").
2. The catch: we **do not have longitudinal** (repeated, per-patient) data to fit transition rates directly.
3. We only have **cross-sectional** snapshots: what fraction of the population is in each state *right now*.
4. Solution: recover plausible transition rates from the equilibrium population proportions, using a principled, least-biased statistical criterion — **Maximum Caliber**.
5. Applied here to NHANES survey data (2006–2020) together with literature-derived transition information for T2DM.

What is a continuous-time Markov chain?

- A stochastic process $(X_t)_{t\geq 0}$ describes a system whose state changes randomly over time; here X_t is the clinical state of a patient at time t.
- **Markov**: no memory — the future depends on the present state only.
- **Continuous time**: transitions can happen at any real instant, not on a fixed clock tick — matching disease progression, which does not follow a calendar.

Definition (Generator)

$(X_t)_{t\geq 0}$ on a finite state space $S = \{1, \dots, m\}$ is a CTMC with **generator** $\mathbf{Q} = (q_{ij})$ if, for $i \neq j$,

$$q_{ij} \geq 0, \qquad q_{ii} = -\sum_{j \neq i} q_{ij}, \qquad \lim_{h \downarrow 0} \frac{\mathbb{P}(X_{t+h} = j \mid X_t = i)}{h} = q_{ij}.$$

- q_{ij} is the instantaneous *rate* ("propensity") of jumping $i \to j$; the matrix $\mathbf{Q}$ fully determines the process.

Rates, holding times, and the embedded chain

- Think of q_{ij} as "jumps per unit time" — like a hazard rate: over a short interval h, $\mathbb{P}(\text{jump } i \to j) \approx q_{ij} h$.

Proposition (Holding times)

If $X_t = i$, the waiting time τ_i to the next jump is $\mathrm{Exp}(\lambda_i)$, $\lambda_i = -q_{ii}$; mean holding time $\mathbb{E}[\tau_i] = 1/\lambda_i$.

- Useful because clinical literature reports *mean times* or *event probabilities*, not raw rates.
- Factorization: $\mathbf{Q} = \Lambda(P - I)$, $\Lambda = \mathrm{diag}(\lambda_i)$, $P_{ij} = q_{ij}/\lambda_i$ — separates *how fast* you leave a state (Λ) from *where* you go next (P, the **embedded jump chain**).
- Transition probabilities over a finite time: $P_t = e^{t\mathbf{Q}}$; state probabilities evolve by the **Kolmogorov forward equation** $\frac{d}{dt}\boldsymbol{p}(t) = \boldsymbol{p}(t)\mathbf{Q}$.

Stationary distributions, and a subtlety

- $\boldsymbol{\pi}$ is **stationary** (an equilibrium) if $\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$, $\sum_i \pi_i = 1$: inflow balances outflow everywhere; if the population starts at $\boldsymbol{\pi}$, it stays there.
- Stationary distributions are what a single cross-sectional survey can plausibly estimate — a snapshot, not a trajectory.
- **Subtlety**: the embedded chain's stationary law is *not* $\boldsymbol{\pi}$ itself. Stationarity $\boldsymbol{\pi}^\top Q = 0$ is equivalent to

$$(\boldsymbol{\pi} \odot \boldsymbol{\lambda})^\top P = (\boldsymbol{\pi} \odot \boldsymbol{\lambda})^\top,$$

  so the embedded chain sees $\pi_i \lambda_i$ — the **stationary throughput** of state i — not π_i alone.
- This distinction matters directly for how the optimization problem ahead is weighted.

Why is finding Q from π hard?

- Suppose we only observe the equilibrium $\boldsymbol{\pi}$ (e.g., from survey data) and want to recover the generator $\mathbf{Q}$ that produced it.
- This **inverse problem** is severely **ill-posed**:
  1. **Dimensionality**: $m(m-1)$ unknown off-diagonal rates, but only $m-1$ independent balance equations.
  2. **Scale invariance**: if $\mathbf{Q}$ satisfies $\boldsymbol{\pi}\mathbf{Q} = 0$, so does $k\mathbf{Q}$ for any $k > 0$ — uniformly speeding everything up doesn't change the equilibrium.
- The solution set for $\mathbf{Q}$ is an entire **cone**, not a single point.

- **Sparsity**: clinical knowledge tells us which transitions are biologically plausible (e.g., no direct jump from prediabetes to death). A sparsity mask shrinks the unknowns from $m(m-1)$ down to just the permitted edges.
- **Known holding times**: mean sojourn times $\tau_i = 1/\lambda_i$ from follow-up studies fix the scale for those states, resolving the $k\mathbf{Q}$ ambiguity.
- **Flow picture**: define edge flows $f_{ij} = \pi_i q_{ij}$. Feasibility requires, at every state i,

$$\sum_j f_{ij} = \pi_i \lambda_i = \sum_j f_{ji}.$$

- This makes transparent *why* some sparse graphs simply cannot support a prescribed, strictly positive equilibrium — a fact that returns, with real numbers, later for diabetes.

The Maximum Entropy idea

- Jaynes' Principle of Maximum Entropy: among all distributions/models consistent with known constraints, choose the one that adds the *least* extra, unjustified information.
- Goal: avoid implicitly assuming things about the unknowns that the data does not actually say.
- **Static** case: maximize Shannon entropy subject to constraints.
- **Dynamic** case (processes evolving in continuous time): the analogous principle on *path measures* (probabilities over entire trajectories) is called **Maximum Caliber** (MaxCal).

Maximum Caliber: the optimization program

- Compare a candidate generator Q to a reference/prior generator $Q^{(0)}$ (e.g., literature-based rates), both with stationary law $\boldsymbol{\pi}$. The **relative entropy rate** between path measures is

$$\mathcal{R}(Q\|Q^{(0)};\pi) = \sum_i \pi_i \Big[\sum_{j\neq i} q_{ij} \log \frac{q_{ij}}{q_{ij}^{(0)}} - (\lambda_i - \lambda_i^{(0)}) \Big] \geq 0,$$

  with equality iff $Q = Q^{(0)}$.

- **MaxCal program**: minimize $\mathcal{R}(Q\|Q^{(0)};\pi)$ subject to
  - *Stationarity*: $\boldsymbol{\pi}^\top Q = 0$;
  - *Sparsity*: $q_{ij} = 0$ off the admissible edges;
  - *Holding times*: $\lambda_i = 1/\tau_i$ where known.
- Convex problem: existence and uniqueness of the minimizer are guaranteed whenever the feasible set is non-empty.

Solving it: row-wise projection and exponential tilt

- Fixing exit rates λ_i and writing $Q = \Lambda(P - I)$, $Q^{(0)} = \Lambda(P^{(0)} - I)$, the objective reduces to

$$\mathcal{R}(Q \| Q^{(0)}; \pi) = \sum_i \underbrace{\pi_i \lambda_i}_{w_i} \sum_{j \neq i} P_{ij} \log \frac{P_{ij}}{P_{ij}^{(0)}}$$

  — a weighted sum of row-wise KL divergences, weighted by the stationary throughputs w_i, not by π_i alone.
- The feasible set is a convex product of simplices; the KL objective is strictly convex, so the minimizer $P^\star$ exists and is unique whenever the prior is strictly positive on the allowed edges.
- In Lagrangian form, $P^\star$ is an **exponential tilt** of the prior — adjusted just enough to satisfy the global flow constraints, exactly as exponential families arise from ordinary Maximum Entropy.

The fuid limit theorem

- Consider N individuals, each independently following the same CTMC generator Q. Let $x_i^{(N)}(t)$ be the fraction of individuals in state i at time t.
- As N grows, random fluctuations should average out — a law-of-large-numbers argument.

Theorem (Fluid Limit; Kurtz)

As $N \to \infty$, $x_i^{(N)}(t)$ converges in probability to a deterministic $x_i(t)$ satisfying

$$\frac{dx_i}{dt} = \sum_{j \neq i} x_j \, q_{ji} - x_i \sum_{j \neq i} q_{ij}.$$

- Remarkably, *the same functional form* as the Kolmogorov forward equation for $\boldsymbol{p}(t)$ — but $x(t)$ is now a deterministic population proportion, not one individual's probability law.
- This is the rigorous justification (Kurtz, Ethier) for using ODE models at the population scale.

From theory to practice: calibration and caveats

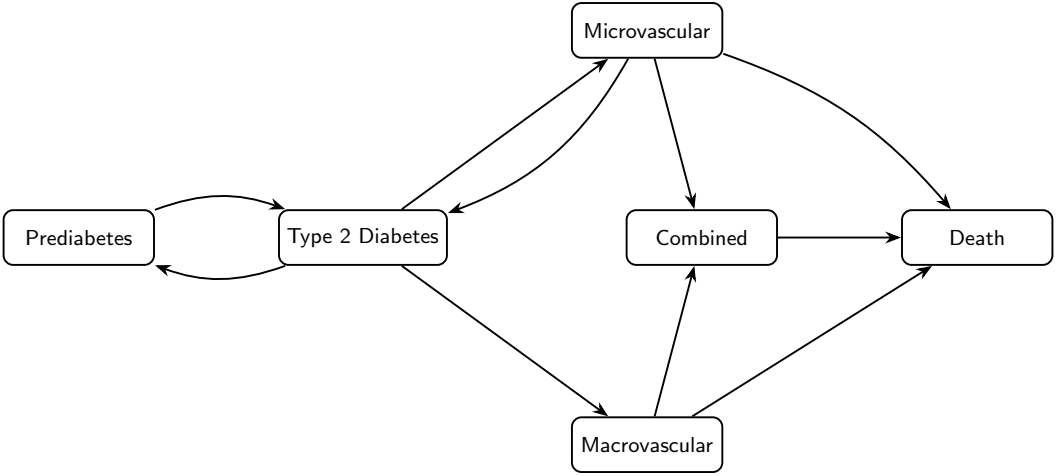
- **Sanity check** (simplex invariance): if $x(0)$ is a probability vector and Q is conservative, $x(t) = x(0)e^{tQ}$ stays in the probability simplex for all $t \geq 0$ — proportions never leave $[0, 1]$ or fail to sum to 1.
- The ODE approximates *expected* behavior; the approximation degrades for small populations, rare transitions, or possible extinction of a subpopulation.
- **Calibrating ODE rates from residence times**: with mean residence $\bar{\tau}_i$ in state i, set $\lambda_i = 1/\bar{\tau}_i$, factor $q_{ij} = \lambda_i P_{ij}$, and solve the same row-wise KL projection as before. The result, $k_{ij} = q^\star_{ij} = \lambda_i P^\star_{ij}$, gives the calibrated ODE rate coefficients directly.

Clinical background, and seven clinical states

- Diabetes mellitus: a metabolic disorder of carbohydrate metabolism causing hyperglycemia. T2DM drives micro- and macrovascular complications — major sources of morbidity, mortality, and cost.
- T2DM strongly accelerates atherosclerotic disease (myocardial infarction, stroke, peripheral arterial disease), soon after diagnosis; coexisting vascular complications sharply raise long-term mortality.

1. **Healthy**: normoglycemic, no diagnosis.
2. **Prediabetes (P)**: impaired fasting glucose/glucose tolerance.
3. **Type 2 Diabetes (T)**: diagnosed, uncomplicated.
4. **Microvascular (Mi)**: retinopathy, nephropathy, or neuropathy.
5. **Macrovascular (Ma)**: coronary, cerebrovascular, or peripheral arterial disease.
6. **Combined (B)**: both micro- and macrovascular.
7. **Death (D)**: absorbing state.

The disease transition graph



- No direct P $\rightarrow$ complication jump, no remission from macrovascular states, death is absorbing — structural assumptions motivated by clinical irreversibility.

The sparse generator matrix

State order (P, T, Mi, Ma, B, D):

$$Q = \begin{pmatrix} -q_{PT} & q_{PT} & 0 & 0 & 0 & 0 \\ q_{TP} & -(q_{TP} + q_{TMi} + q_{TMa}) & q_{TMi} & q_{TMa} & 0 & 0 \\ 0 & q_{MiT} & -(q_{MiT} + q_{MiB} + q_{MiD}) & 0 & q_{MiB} & q_{MiD} \\ 0 & 0 & 0 & -(q_{MaB} + q_{MaD}) & q_{MaB} & q_{MaD} \\ 0 & 0 & 0 & 0 & -q_{BD} & q_{BD} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

- Ten free rate parameters govern the entire disease dynamics.

From literature reports to rates

- Clinical literature rarely reports raw rates q_{ij} directly; it reports either:
  - a **mean time** T to transition $\Rightarrow$ set $q = 1/T$;
  - an **event probability** p over a horizon T $\Rightarrow$ assume constant hazard: $q = -\ln(1-p)/T$.
- Rare-event approximation bound:

$$0 \leq -\frac{\log(1-p)}{T} - \frac{p}{T} \leq \frac{p^2}{2T(1-p)},$$

so the naive shortcut $q \approx p/T$ is only valid when p is small — and is *not* used in this paper's numerics.

Literature-based edge rates

Edge	Reported quantity	Rate value (yr^{-1})
$P \to T$	9 years mean	0.1110
$T \to P$	9.2% in 3 yrs	0.0321
$T \to Mi$	12% in 3 yrs	0.0426
$T \to Ma$	5.8% in 3 yrs	0.0198
$Mi \to T$	4.6% in 2 yrs	0.0235
$Mi \to B$	7.6% in 3 yrs	0.0265
$Mi \to D$	38.2% in 13 yrs	0.0370
$Ma \to B$	17.8% in 3 yrs	0.0649
$Ma \to D$	5 years mean	0.2000
$B \to D$	5 years mean	0.2000

NHANES: the cross-sectional target

- NHANES: a large US National Health and Nutrition Examination Survey, cross-sectional by design; six living categories by biomarker thresholds ("healthy" requires *explicit* evidence of neither diabetes nor prediabetes).
- Cycle-specific proportions (2006, 2008, . . . , 2020) averaged to one empirical equilibrium target.
- Key limitation: death is not part of the living cross-sectional composition, so mortality outflow is invisible to a naive closed six-state model.

Living state	Average proportion
Healthy	0.8364
Prediabetes	0.0529
T2DM, uncomplicated	0.0642
Microvascular only	0.0145
Macrovascular only	0.0236
Micro- and macrovascular	0.0084

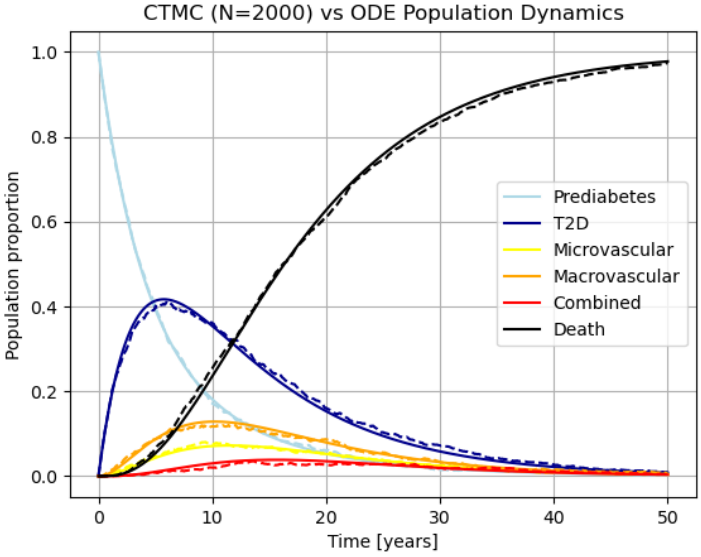
An infeasibility, and its fix

- **Key finding**: the closed six-state disease-only equilibrium equations **do not admit a positive exact solution** at the NHANES composition when all ten literature rates are fixed.
- Why: literature rates, compiled from heterogeneous cohorts with different inclusion criteria, are not jointly consistent with a single population-level stationary distribution.
- **Fix**: augment the model with an auxiliary **open-population healthy reservoir** H, feeding into prediabetes and absorbing mortality outflow — used *only* during calibration, not in the final disease simulation.
- Extended system (H, P, T, Mi, Ma, B): 11 unknown rates, equilibrium matrix has rank 5 $\Rightarrow$ fix 6 rates, solve for the remaining 5. Enumerate minimal fixed-rate combinations, insert literature values where possible, filter for exact *positive* solutions, rank by closeness to literature.

Calibrated versus literature rates

Rate	Calibrated	Status	Literature
q_{HP}	0.0103	inferred	n/a
q_{PT}	0.2020	inferred	0.1110
q_{TP}	0.0321	fixed	0.0321
q_{TMi}	0.0426	fixed	0.0426
q_{TMa}	0.0972	inferred	0.0198
q_{MiT}	0.0235	fixed	0.0235
q_{MiB}	0.0109	inferred	0.0265
q_{MiD}	0.1544	inferred	0.0370
q_{MaB}	0.0649	fixed	0.0649
q_{MaD}	0.2000	fixed	0.2000
q_{BD}	0.2000	fixed	0.2000

What we estimate



What changed, and why it matters

- Equilibrium-constrained calibration implies:
  - **Faster** progression from prediabetes to uncomplicated diabetes.
  - **Much faster** progression from uncomplicated diabetes to macrovascular disease (nearly 5$\times$ the literature rate).
  - **Higher** mortality hazard directly from the microvascular state.
  - **Lower** rate of transition from microvascular to combined-complication state.
- These are not arbitrary: they are the minimal, entropy-optimal adjustments needed to reconcile heterogeneous literature hazards with the observed U.S. population equilibrium.

Simulating the calibrated model: ODE vs. Gillespie SSA

- **ODE integration**: forward Euler, $\Delta t = 0.1$ yr, horizon $T = 50$ yr.
- **Gillespie SSA**: exact stochastic simulation — draw the waiting time from $\mathrm{Exp}(\lambda_i)$, then jump to j with probability q_{ij}/λ_i; $N = 2000$ trajectories, fixed seed. Death is absorbing.
- Two starting cohorts, same generator:
  - All-prediabetes: $x(0) = (1, 0, 0, 0, 0, 0)$.
  - Baseline co-morbidity (from clinical cohort literature): $x(0) = (0, 0.65, 0.12, 0.17, 0.06, 0)$.
- **Result**: the ODE trajectory and the mean of 2000 SSA paths agree closely, confirming internal consistency of the two methods.
- Under the calibrated (faster) hazards, uncomplicated T2DM peaks just above 0.4 within the first decade, then complications and mortality dominate; by year 50 death contains nearly the whole cohort. Starting from the co-morbid baseline produces an even steeper, front-loaded mortality curve.

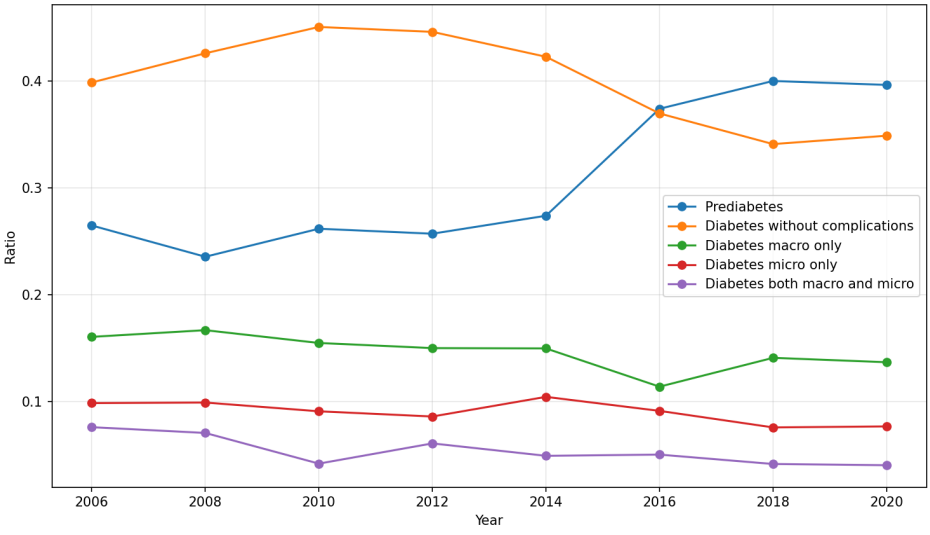
MaxCal versus alternative estimators

- **Constrained least squares**: fit a perturbation ΔQ minimizing a quadratic norm subject to approximate stationarity. Simple, but no information-theoretic interpretation, and can distort rare transitions disproportionately.
- **Spectral methods**: infer equilibrium/mixing behavior from eigenvalues/eigenvectors of the transition operator. Powerful with detailed transition data, but do not naturally encode holding-time or prevalence priors.
- **EM / likelihood methods**: need repeated per-patient longitudinal measurements — scarce and censored here; do not automatically enforce equilibrium.
- **MaxCal**: preserves prior clinical structure, enforces equilibrium exactly, and weights corrections by their contribution to *stationary flow* — the right tool exactly when data are cross-sectional prevalence plus summary hazards, as here.

What the calibration shift means

- Heterogeneous literature sources (different cohorts, follow-up durations, covariate adjustments) are not automatically consistent with one global population equilibrium.
- Equilibrium constraints introduce **global coupling**: adjusting one rate can shift the entire stationary distribution, so MaxCal coordinates adjustments across many transitions at once.
- The open-population extension reveals **unobserved mortality turnover** needed to reconcile cross-sectional living-state prevalence with the fact that death removes people from the living population entirely.
- Dynamical-systems view: calibration replaces a literature-implied drift field $F^{(0)}$ with a calibrated drift $\tilde{F}$ whose stationary point matches NHANES — this can change stability, transient dynamics, and long-run mortality, directly relevant to QALY/cost-effectiveness and actuarial applications, which depend on the whole trajectory, not just its endpoint.

Trust NHANES?



Applications, limitations, and future directions

- **Applications**: health-economic evaluation (QALYs, cost-effectiveness); actuarial risk pricing (premiums, reserves); population forecasting under current or counterfactual rates; the same MaxCal + fluid-limit pipeline transfers to any staged chronic disease with cross-sectional prevalence data.
- **Limitations**: structural choices (irreversible macrovascular disease, two aggregated complication classes); literature rates drawn from heterogeneous populations and study designs; the healthy reservoir is a calibration device, not a modeled disease state.
- **Future directions**: age-structured/time-varying models; richer data integration (partial longitudinal sub-cohorts alongside cross-sectional targets); Bayesian/hierarchical MaxCal for uncertainty quantification; application to other chronic diseases.

Key takeaways

- CTMCs and their mean-field ODEs are two views of the same underlying disease process, connected by a fluid-limit theorem.
- When only cross-sectional prevalence is available, Maximum Caliber offers a convex, uniquely-solvable method for recovering transition rates, which minimizes assumptions.
- Applying this to T2DM revealed that literature-only rates are inconsistent with the observed U.S. population equilibrium, requiring a mortality-aware open-population correction.
- The calibrated model implies substantially faster progression to macrovascular disease and death than a naive literature-based model, with real implications for health-economic and actuarial practice.

Thank you!

Dept. of Biomatics, Óbuda University, Budapest Hungary
Dept. of Mathematics, Mahidol University, Bangkok Thailand
CNR Institute for Systems Analysis and Informatics, Rome, Italy

andrea.degaetano@uni-obuda.hu