# Evolution on graphs and the transition to cancer

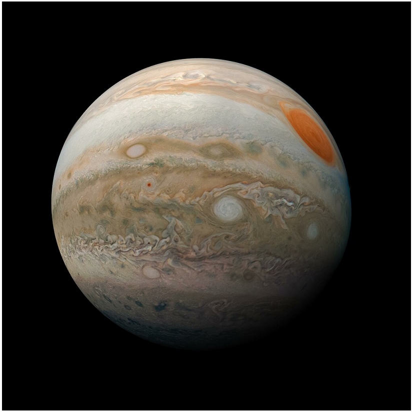
Chay Paterson

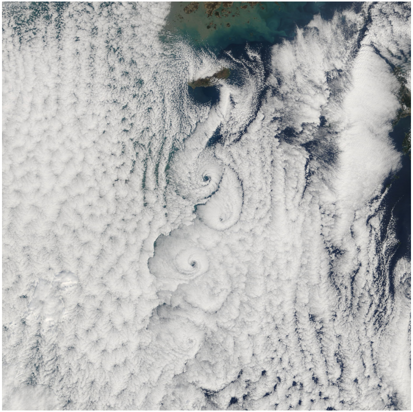
University of Manchester

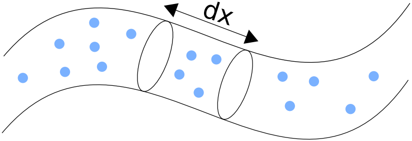
June 25, 2024

# Introduction

## Differential equations

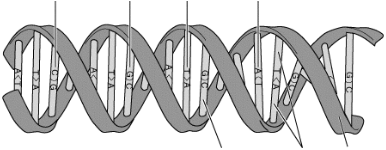






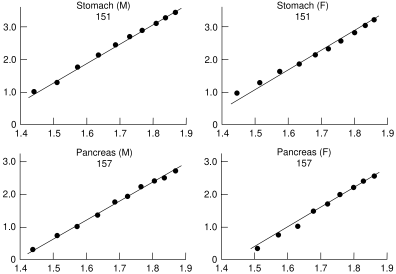
# Introduction

## Driver mutations



# Multi-stage models

## P. Armitage and R. Doll[1][2]

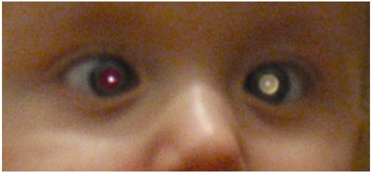


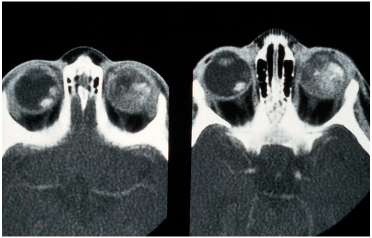
[1] P. Armitage and R. Doll, British Journal of Cancer 1954; 8: 1–12

[2] P. Armitage and R. Doll, British Journal of Cancer 1957; 11(2): 161-169

# Multi-stage models

A.G. Knudson[1][2]





Mutation and Retinoblastoma 823

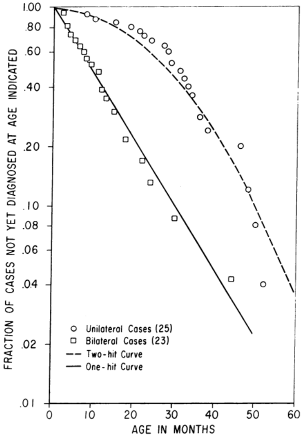


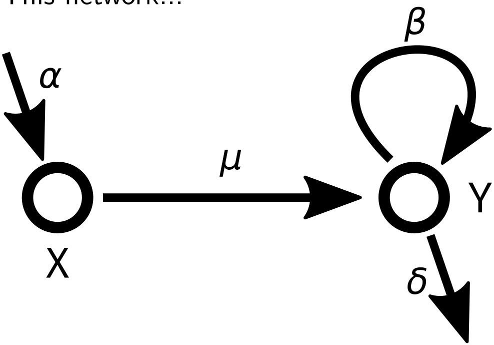
Fig. 1. Semilogarithmic plot of fraction of cases of retinoblastoma not yet diagnosed ($S$) vs. age in months ($t$). The one-hit curve was calculated from $\log S = -t/30$, the two-hit curve from $\log S = -4 \times 10^{-5} t^2$.

[1] AG. Knudson, PNAS 68.4 (1971): 820-823.

[2] F. Michor, Y. Iwasa and MA. Nowak, Nature Reviews Cancer 2004; 4: 197-205 doi:10.1038/nrc1295

# Graphs

This network...



corresponds to this stochastic process:...[1]

$$\begin{aligned} \emptyset &\to X, & &\text{rate } \alpha \\ X &\to X + Y, & &\text{rate } \mu X \\ Y &\to Y + Y, & &\text{rate } \beta Y \\ Y &\to \emptyset, & &\text{rate } \delta Y \end{aligned}$$

[1] C. Paterson, I. Bozic, H. Clevers, PNAS 2020; 117(34): 20681-20688

# How do these studies work?

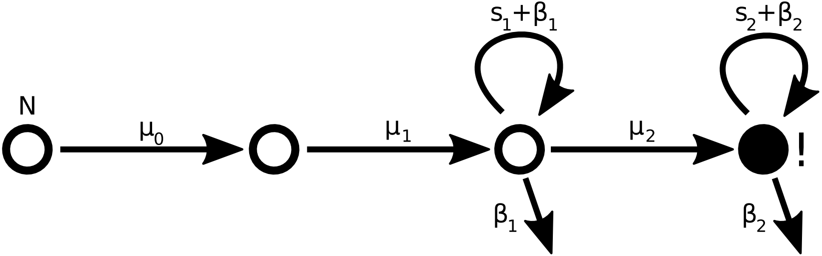
What are the relevant observables in this type of longitudinal study & data analysis?

Track a cohort that initially contains $N$ patients in the study:

- Age-specific incidence $I(a)$: rate at which new cases are recorded in the cohort with ages between $a$ and $a + da$
- Survival function $S(a)$: probability to survive to age $a$ without being diagnosed.

$$I(a) = -N\frac{dS}{da} = -N\;S(a)\frac{d\ln S}{da} \tag{1}$$

# So what?



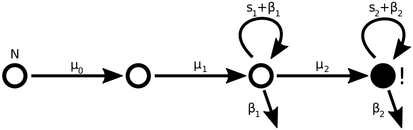
- The survival curve S(a) determines the incidence curve $I(a)$
- The model determines the survival curve: the probability not to end up at one of the end nodes of the graph. We want to know what model+parameters best agree with data from studies.
- To fit (or “train”) the model to longitudinal data, we need to compute $S(a)$!

This is the central mathematical problem in cancer epidemiology.
*How do we compute $S$?*

# Multi-stage clonal expansion models

## 2-3 rate limiting steps[1,2,3]

Problem: how to compute $S(t)$ for a given model?

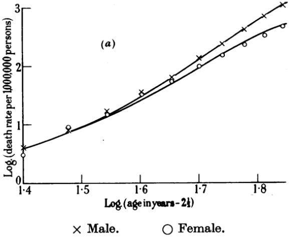


Different methods: Fast:

- Armitage + Doll's approximation[1]
- Moolgavkar + Venzon's quadrature[2]

Very slow:

- Gillespie algorithm + sampling [3]



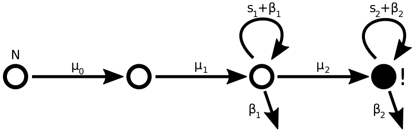
↓?

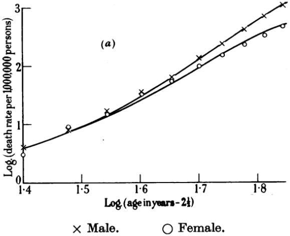
[1] P. Armitage and R. Doll, British Journal of Cancer 1957; 11(2): 161-169

[2] S. Moolgavkar and G. Luebeck, JNCI 1992; 84(8): 610-618

[3] C. Paterson, I. Bozic, H. Clevers, PNAS 2020; 117(34): 20681-20688 (supp. material)

# Armitage and Doll's approximation





- Assume all the probabilities are small: $1 - S \ll 1$
- Then the relevant probability $S(a)$ is expressible in terms of expected values/population means, which implies

$$S(a) \propto a^k(e^{sa} - 1) \qquad (2)$$

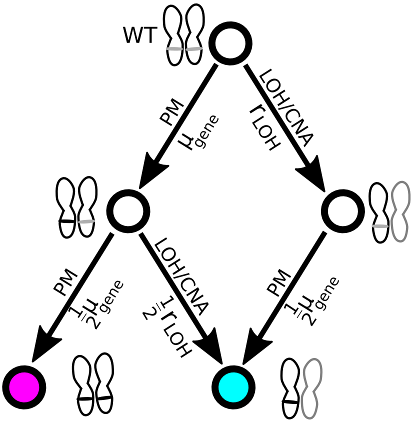
with constants $k$ and $s$.

- Don't use correlations, variances, or higher moments in stem cell populations – just ignore these.

# Models on graphs

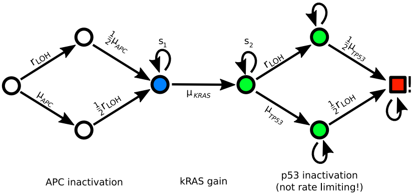
1. Most methods for computing $S(a)$ do not consider graphs with multiple end nodes
2. To study **specific genes** and mechanisms of interest (SNVs, LOH, CNA, etc.), we need to evaluate $S(a)$ for a model defined on a graph (right)
3. What methods do we actually have?

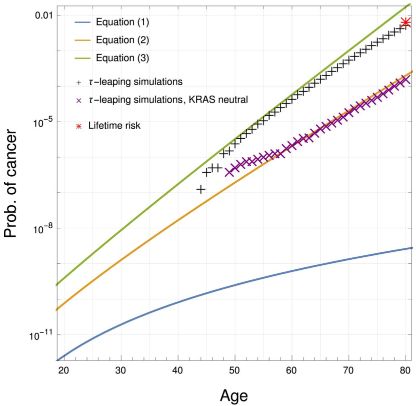
example model



This gets us the incidence of *specific karyotypes*

# Colorectal adenocarcinoma model





- Can use A-D type approximation, or stochastic simulations

but:

- Mean-field breaks down at old ages / large probabilities
- Stochastic simulations are *extremely slow*

[1] C. Paterson, I. Bozic, H. Clevers, PNAS 2020; 117(34): 20681-20688

# Alternative approach: Kolmogorov forward equations

Define a more general generating function $\Psi$:

$$\Psi(t, \vec{q}) = \mathbb{E}[\left[\prod_j q_j^{N_j}\right] \tag{3}$$

and derive Kolmogorov forward equations instead. Then we can numerically integrate these, and get survival curves $S_i(a)$ for different types of cancer $i$. E.G. tumours with clonal LOH, or no clonal LOH.

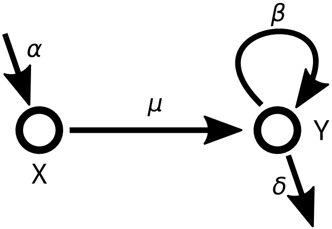
$$S_i = \Psi(t, q_j = 1, \ldots, q_i = 0) \tag{4}$$

People knew about this approach for a long time (since 1970s) but it was never considered as useful as backward equations.

[1] SH Moolgavkar and DJ Venzon, Math. biosc. 1979; 47(1): 55-77

[2] DW Quinn, Risk Analysis 1989; 9(3): 407-13

# Kolmogorov forward equations as wave equations



Briefly: the Kolmogorov forward equations in $P(N_0, \dots)$

$$\frac{dP(N_0, N_1, \dots)}{dt} = \sum_{vertices} \alpha(N_j - 1)P(\dots, N_j - 1, \dots)$$
$$- \alpha N_j P(\dots, N_j, \dots)$$
$$+ \beta \cdots + \mu \cdots + \delta \cdots$$

get transformed...

# Kolmogorov forward equations as wave equations

They transform with $P \rightarrow \Psi$:

$$\frac{\partial \Psi}{\partial t} = \sum_{vertices} \alpha(q_j - 1) q_j \frac{\partial \Psi}{\partial q_j} + \cdots = \mathcal{H}\Psi \tag{5}$$

where $\mathcal{H}$ is a hyperbolic differential operator.
Because this is a hyperbolic wave equation, we can solve for future values of $\Psi$ if we have initial values, by evolving them along *characteristics*.

[1] SH Moolgavkar and DJ Venzon, Math. biosc. 1979; 47(1): 55-77

# Kolmogorov forward equations

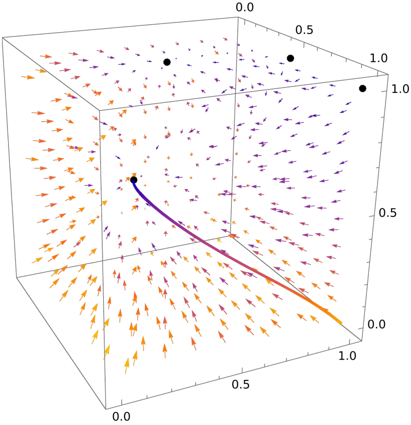
Using the big generating function $\Psi$, find the corresponding wave equation:

$$\frac{d\Psi}{dt} = \mathcal{H}\Psi \qquad (6)$$

This can be solved using the method of characteristics, so we can instead integrate

$$\frac{d\vec{\gamma}}{dt} = \vec{X} \qquad (7)$$

numerically, using an appropriate time stepper.



The vector field $\vec{X}$ and a characteristic $\vec{\gamma}$

[1] SH Moolgavkar and DJ Venzon, Math. biosc. 1979; 47(1): 55-77

# Random sampling

## Error analysis

To compare methods, ask under what conditions the errors are similar. Stochastic algorithms have

$$\epsilon \sim N^{-1/2} \tag{8}$$

and will thus need

$$N \sim \mathcal{O}(\epsilon^{-2}) \tag{9}$$

runs, so overall runtime $T \propto N$.

# Method of characteristics

## Error analysis

Why wasn't this method ever used? Wave equation+characteristic methods were studied before, but used Euler integration, which has error

$$\epsilon \sim \Delta t \tag{10}$$

and required two passes, so it ran in

$$T \sim \Delta t^{-2} \sim \mathcal{O}(\epsilon^{-2}) \tag{11}$$

this is asymptotically just as bad as random sampling!

[1] SH Moolgavkar and DJ Venzon, Math. biosc. 1979; 47(1): 55-77

[2] DW Quinn, Risk Analysis 1989; 9(3): 407-13

# Fast forward method

## Error analysis

But what happens if we only need one pass, and replace Euler integration with a Runge-Kutta scheme?

$$\epsilon \sim \Delta t^2 \tag{12}$$

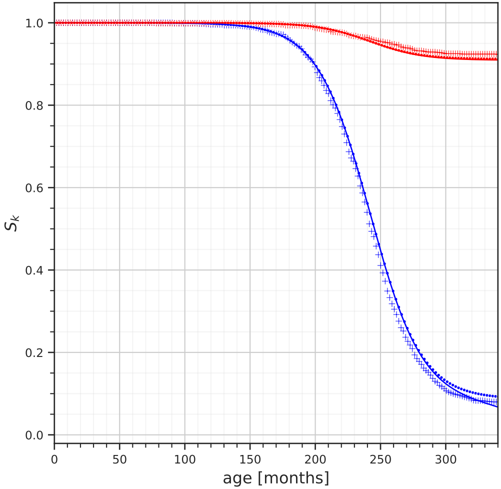
and runs in

$$T \sim \Delta t^{-1} \sim \mathcal{O}(\epsilon^{-1/2}) \tag{13}$$

so new runtime $\sim \mathcal{O}\left( \text{ old runtime }^{1/4}\right)$.

Amazing!

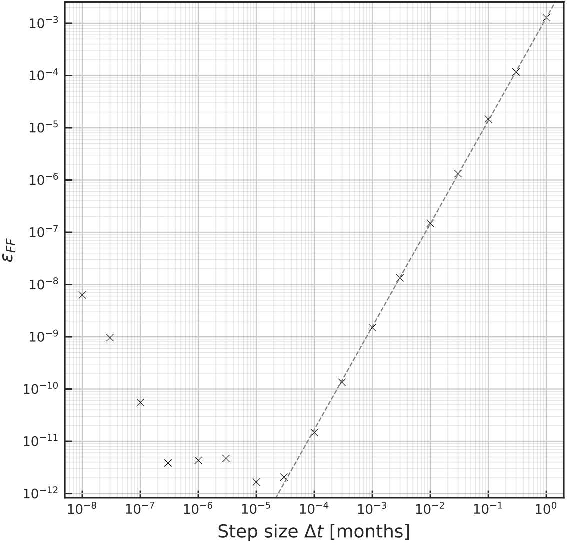
# Fast forward method



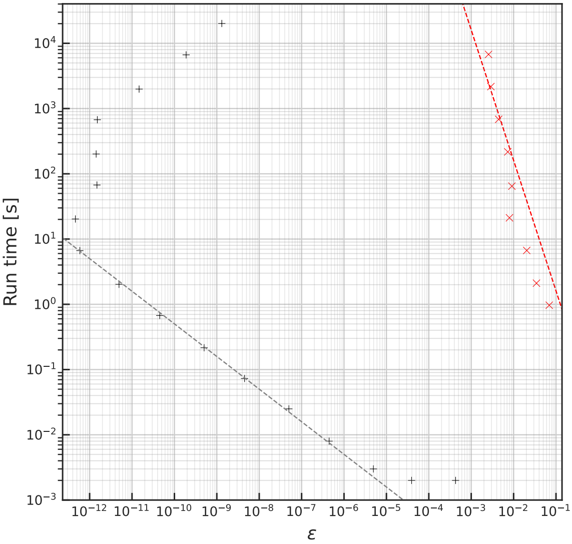
Random sampling vs fast forward method:
Monte Carlo: ≈ 5000s Fast forward: 4ms

# Fast forward method

## Error analysis

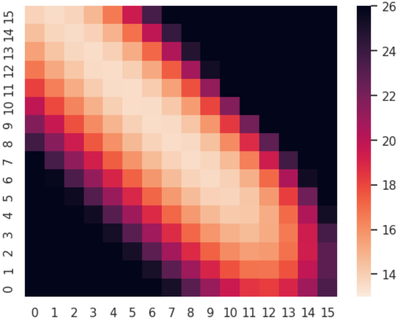


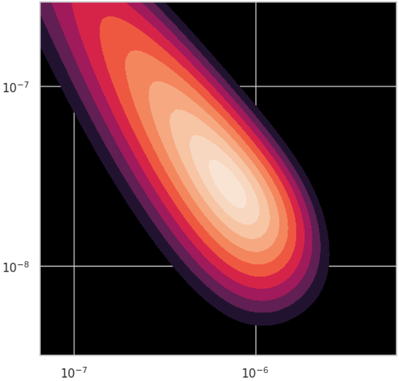
# How do they compare?



# Fast forward method

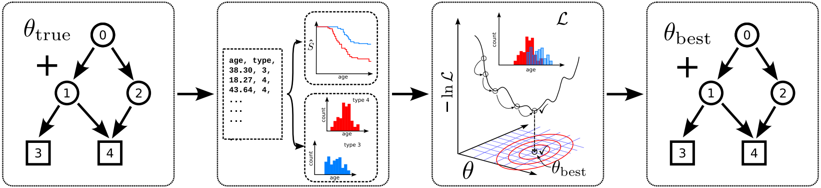
Efficiently computing $S_k(a)$ means we can evaluate likelihood functions directly, just sampling them.....





# Fast forward method

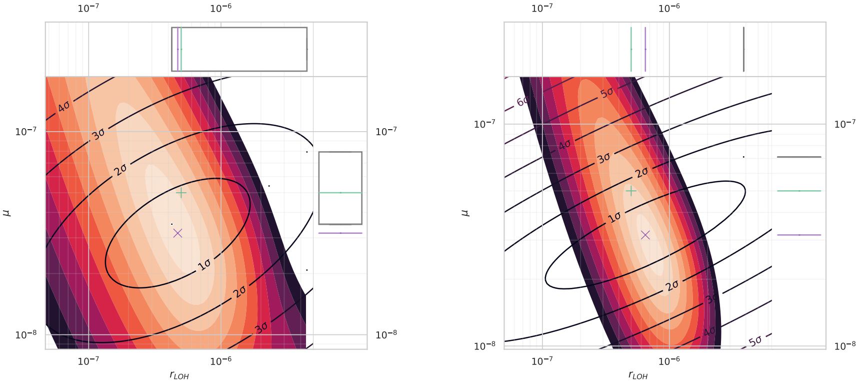
## Parameter inference



```
age, type,
38.30, 3,
18.27, 4,
43.64, 4,
...
...
...
```

# Fast forward method

## Parameter inference



# Thank you!

What is my message?

- Don't study continua – study PROBABILITIES!
- Age structure is important & informative, genes are discrete

Where next?

- Run these analyses on real studies!
- Combine genomic and age data TOGETHER
- Ongoing: sequencing schwannoma tumours in NF2 and oesophageal cancer in Barrett's cases

# Acknowledgements

CDMRP
DEPARTMENT OF DEFENSE
CONGRESSIONALLY DIRECTED
MEDICAL RESEARCH PROGRAMS

funded this research

NHS
Manchester University
NHS Foundation Trust

W

The University of Washington

ACED
INTERNATIONAL
ALLIANCE FOR
CANCER EARLY
DETECTION

MANCHESTER
1824
The University of Manchester

All my collaborators...

- Ivana Božić
- Gareth Evans
- Miaomiao Gao
- David Wedge
- Miriam Smith
- Marian Love
- Joshua Hellier