Characterization of myeloproliferative disorders detection time through an age-dependent mutation rate

Ana Fernández Baranda

Under the supervision of Sylvie Méléard and Vincent Bansaye

CMAP, Ecole Polytechnique

Table of contents

Biological context

Myeloproliferative Neoplasms (MPN): family of cancers affecting blood cells.

JAK2V617 mutation related MPNs: Essential Thrombocytemia and Vaquez disease.

Goal

Estimate the distribution of the age of detection for these MPNs.

Data

Ages of detection of both diseases in the presence of JAK2V617 mutation in the Côte d'Or region.

Age	PV JAK2+	ET JAK2+	Total JAK2+ cases
0 to 4	0	0	0
5 to 9	0	0	0
10 to 14	0	0	0
15 to 19	0	0	0
20 to 24	0	0	0
25 to 29	0	0	0
30 to 34	1	4	5
35 to 39	0	4	4
40 to 44	1	4	5
45 to 49	5	4	9
50 to 54	2	7	9
55 to 59	8	11	19
60 to 64	10	6	16
65 to 69	13	15	28
70 to 74	12	14	26
75 to 79	14	29	43
80 to 85	15	30	45
More than 85	11	23	34
Total number	92	151	243

Data

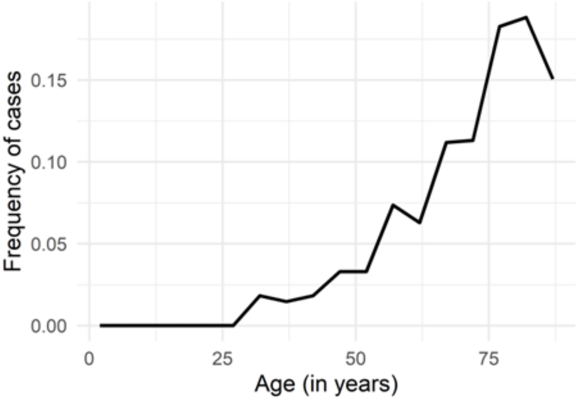


Table of contents

Modelisation: time to detection T_M

Two independent elements:

- T_1: active mutation time;
- T_2: MPN growing time.

Time to detection

$T_M = T_1 + T_2$

Table of contents

Models with constant active mutation rate

Assumptions:

- Mutations from stem cells occur at a constant rate τ.
- Each mutant cell has a probability p of eventually having its population reach detection size.

$\implies T_1 \sim Exp(\delta)$.

$\delta = \tau p$: **active mutation rate**.

Simplest case: constant MPN growing time

$T_2 = \alpha$ constant.

Distribution of T_M

$$f_M(t) = \delta e^{-\delta(t-\alpha)}.$$

- **Estimation** of δ and α through least squares.
- **Goodness of fit test**.

Chi-squared goodness of fit test

Testing the hypotheses:

$$\mathcal{H}_0 = \text{the data follows the model distribution,}$$
$$\mathcal{H}_1 = \text{the data does not follow said distribution.}$$

Test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

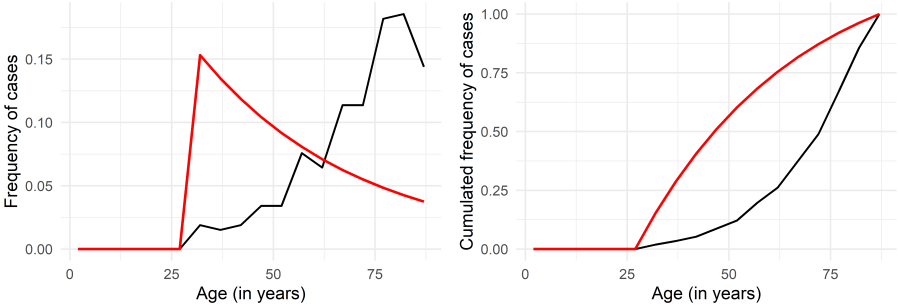
with

- O_i: frequency in bin i,
- $E_i = N(F(X_u^i) - F(X_l^i))$,

Null hypothesis is rejected if

$$\chi^2 \geqslant \chi^2_{1-\alpha,k-m}.$$

Constant MPN growing time: results



Model was **rejected**.

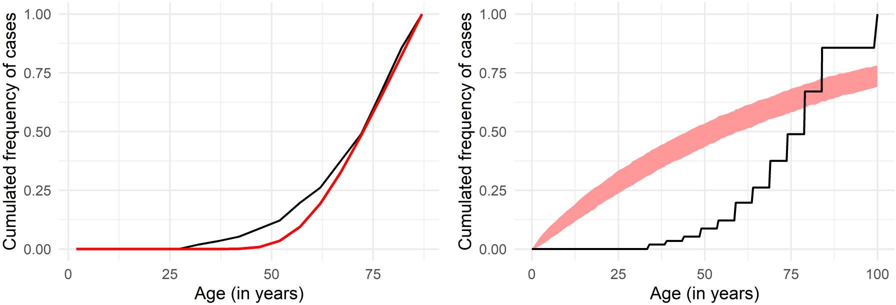
Two possible solutions

- **Random MPN growing time**:

$$T_2 \sim \mathit{lognormal}(\mu, \sigma^2)$$

- **Individual variability**: each individual has a different active mutation rate δ_i, $i = 1, 2, \ldots, N$, considered as a sample of a $\mathit{lognormal}(m, s^2)$

Estimation of both models



Both models were **rejected**.

Table of contents

Age dependency on active mutation rate

Age dependency of δ:

$$\delta(t) = A\exp(kt)$$

.

$$\implies f_1(t) = A\exp\left(\frac{A}{k}\right)\exp\left(kt\right)\exp\left(-\frac{A}{k}\exp(kt)\right).$$

Two age-dependant models

Model 1: $T_2 = \alpha$ constant.

Model 1: distribution of T_M

$$f_M(t) = A \exp\left(\frac{A}{k}\right) \exp\left(k(t-\alpha)\right) \exp\left(-\frac{A}{k}\exp(k(t-\alpha))\right).$$

Model 2: $T_2 \sim \mathit{lognormal}(\mu, \sigma^2)$.

Model 2: distribution of T_M

$$f_M(t) = \int_0^t A \exp\left(\frac{A}{k}\right) \exp\left(ks\right) \exp\left(-\frac{A}{k}\exp(ks)\right) \times \frac{1}{(t-s)\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log(t-s)-\mu)^2}{2\sigma^2}\right) ds$$

Estimation: generalized EM algorithm

$X \sim f_X(x;\theta)$: **observed data** of ages of detection .

$Z \sim f_Z(z;\theta)$: **missing data** (values of T_2).

$(x, z) \sim f_{X,Z}(x, z;\theta)$: **complete data**.

Goal

Find θ that maximizes

$$\log L_{X,Z}(\theta) = \log f_{X,Z}(x, z;\theta).$$

Generalized EM algorithm

θ_0: initial value.

Iteration $k+1$:

- **E-step**: computing

$$Q(\theta;\theta_k) = \mathbb{E}_{\theta_k}(\log L_{X,Z}(\theta)|x).$$

- **M-step**: choose θ_{k+1} such that

$$Q(\theta_{k+1};\theta_k) \geqslant Q(\theta_k;\theta_k).$$

M-step

(Rai and Matthews, 1993)

θ_{k+1}: **one Newton-Raphson step** from θ_k over the function $Q(\theta_{k+1}; \theta_k)$, that is

$$\theta_{k+1} = \theta_k + a_k \delta_k,$$

where

$$\delta_k = -\left[\frac{\partial^2(Q(\theta;\theta_k))}{\partial\theta\partial\theta^T}\right]^{-1}\Bigg|_{\theta=\theta_k}\left[\frac{\partial(Q(\theta;\theta_k))}{\partial\theta}\right]\Bigg|_{\theta=\theta_k}$$

and $0 < a_k \leqslant 1$.

Choice of a_k

- **First**: staying in the parameters space and nondecreasing likelihood.

  Start with $a_k^0 = 1$

  $a_k^{j+1} = \frac{a_k^j}{2}$.

  $\implies a_k^{j^*}$.

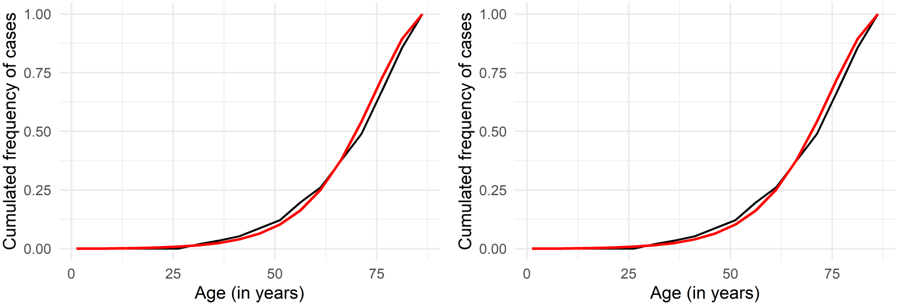
- **Then**: backtracking line search with Armijo condition.

  Start with $a_k^{j^*}$.

  While $Q(\theta_k + a_k^i \delta_k; \theta_k) < Q(\theta_k; \theta_k) + 10^{-4} a_k \nabla Q(\theta_k; \theta_k)^T \delta_k$

  $a_k^{i+1} = 0.8 a_k^i$.

Age-dependency models estimations



Both models were **not rejected**.

Comparing the models: BIC

Bayesian Information Criterion (BIC)

$$\mathrm{BIC} = k \log n - 2 \log \hat{L}.$$

- k: number of parameters of the model,
- n: size of data sample,
- $\hat{L}$: maximized value of the likelihood function.

A lower BIC is preferred.

Comparing the models

Model 1: $T_2 = \alpha$ constant:

$$BIC_1 = 2079.048.$$

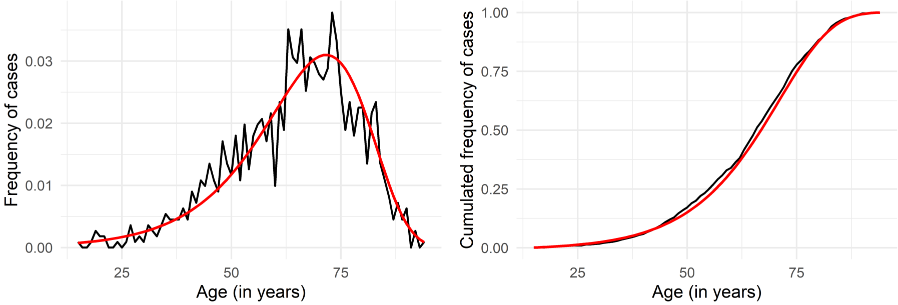
Model 2: $T_2 \sim \mathit{lognormal}(\mu, \sigma^2)$:

$$BIC_2 = 2085.299.$$

$\Longrightarrow$ the gain of adding variability is small and not compensated by the cost of adding a new parameter.

Model validation

French national registry of MPN (FIMBANK): 1111 individuals.



Model was **not rejected**.

Quantitative results

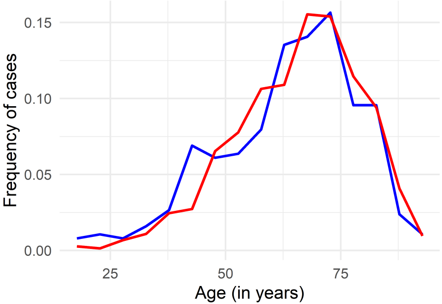
Under the age-dependent model:

- Time of emergence of the cancer (T_1): mean of 60 years.
- Time from tumor emergence to diagnosis (T_2): 8.7 years.

Comparing PV and ET

ET: majority of heterozygous cells.

PV: majority of homozygous cells (mitotic recombination needed).



Estimated time of emergence approximately 1.5 years higher for PV.

References I

- Snedecor, George W. *Statistical methods*. sixth. 1967.
- Rai, S.N. and D.E. Matthews. "Improving the EM algorithm". In: *Biometrics* 46 (1993), pp. 587–591.
- McLachlan, Geoffrey J. and Thriyambakam Krishnan. *The EM Algorithm and extensions*. Wiley- interscience, 2008.
- Lavielle, Marc. *Mixed Effects Models for the population Approach: Models, Tasks, Methods and Tools*. second. 2014.
- Mosca, Matthieu et al. "Inferring the dynamic of mutated hematopoietic stem and progenitor cells induced by IFN - α in myeloproliferative neoplasms". In: *Blood* 138 (2021), pp. 2231–2243.
- Van Egeren, Debra et al. "Reconstructing the Lineage Histories and Differentiation Trajectories of Individual Cancer Cells in Myeloproliferative Neoplasms". In: *Cell Stem Cell* 28 (2021), pp. 514–523.

References II

Mitchell, Emily et al. "Clonal dynamics of haematopoiesis across the human lifespan". In: *Nature* 606 (2022), pp. 343–350.