Age Distribution Inference of Blood cells

Aodong Li

al5350@nyu.edu

1 Recent Progress - April 17

1.1 Specify the linear system

New signal comes in continuously at any time. So if we want to get the correct output sequence, we have to eliminate the previous signal effects, through which we may be able to do the deconvolution pointwise to the signal.

To simplify the notation, let $H(t)$ be $HbA1c(t)$ and $F(t)$ be $1-\tilde{F}_{RBC_{age}}(t;\theta)$ where $\tilde{F}_{RBC_{age}}(t;\theta)$ is the cdf of the age of blood cells and controlled by parameter θ. At this moment, we assume that θ does not change over time and is an unknown constant.

Now suppose we can measure multiple Hs and Gs. The best case would be one measurement of G and one measurement of H at the same time.

The following equations hold up to a constant and is an approximation of formula 5. (We leave out of the constant glycation rate.)

$$\begin{aligned}
H(0) &= G(0)F(0) + G(-1)F(1) + ... + G(-n)F(n) \\
H(1) &= G(1)F(0) + G(0)F(1) + ... + G(-(n-1))F(n) \\
&... \\
H(n) &= G(n)F(0) + G(n-1)F(1) + ... + G(0)F(n)
\end{aligned}$$

which is essentially a linear system and is equivalent to the matrix form,

$$\begin{bmatrix} & G & \end{bmatrix} \begin{bmatrix} F \end{bmatrix} = \begin{bmatrix} H \end{bmatrix}$$

where $1 = F(0) \geq F(1) \geq ... \geq F(n) \geq 0$ as a constraint.

The only problem left is to specify the support n of the age distribution. So we can construct a linear programming (LP) with the objective function to minimize the distance between $F(n)$ and 0.

1.2 Specify the linear programming problem

The constraint $1 = F(0) \geq F(1) \geq ... \geq F(n) \geq 0$ can be rewritten as

$$\begin{aligned} F(0) &= 1 \\ F(0) - F(1) &\geq 0 \\ F(1) - F(2) &\geq 0 \\ &... \\ F(n-1) - F(n) &\geq 0 \\ F(n) &\geq 0 \end{aligned}$$

which is equivalent to the form $AF \geq 0$ and $F(0) = 1$ where

$$A = \begin{bmatrix} 1 & -1 & 0 & ... & 0 \\ 0 & 1 & -1 & 0 & ... \\ ... & & & & \\ 0 & 0 & ... & 0 & 1 \end{bmatrix}.$$

The final constrained problem is

$$\begin{aligned} &\min_F (c^\top F)^2 \\ \text{subject to } &GF = H \\ &AF \geq 0 \\ &F(0) = 1 \end{aligned}$$

where $c = (0, ..., 0, 1)^\top \in \mathbf{R}^n$, $A \in \mathbf{R}^{n \times n}$, $G \in \mathbf{R}^{n \times n}$, $F \in \mathbf{R}^n$, and $H \in \mathbf{R}^n$.

2 Recent Progress - April 2

2.1 Redefine the problem

From my point of view, the original formula 5 does not give a direct computation of $HbA1c$, but involves the related factors, i.e., $G(t)$ and $f_{RBC_{age}}(t)$, and how they affect the measurement $HbA1c$ together.

Based on the formula 5 and further assumptions, we try to define a direct computation formula of $HbA1c$.

First define a new factor $v(t) = f(G(t))$, the glycation rate – the added percentage of glycation per day. Specifically, one of its valid units can be %/d.

Thus the new formula that can directly compute $HbA1c$ is

$$HbA1c(t) = \int_0^\infty v(t-a)\left[1 - \int_0^a f_{RBC_{age}}(\tau)\mathrm{d}\tau\right]\mathrm{d}a. \tag{1}$$

2.2 Specify the glycation rate

Assumptions:

1. Suppose the glucose concentration G contributes linearly to the glycation rate v.
2. Suppose the oldest cells' life is 120 days.
3. Suppose the oldest cells reach a 20% level of glycation in 250 mg/dL glucose concentration, which is the inner environment of a person with type 1 diabetes.

The glycation rate can be specified as

$$v(t) = \frac{G(t)20\%}{250\mathrm{mg/dL} \times 120\mathrm{d}}. \tag{2}$$

So the final formula of $HbA1c$ is

$$HbA1c(t) = \int_0^\infty \frac{G(t-a)20\%}{250\mathrm{mg/dL} \times 120\mathrm{d}}\left[1 - \int_0^a f_{RBC_{age}}(\tau)\mathrm{d}\tau\right]\mathrm{d}a. \tag{3}$$

2.3 Simulations

Suppose the $G(t)$ is circulant, i.e., we extend the measurement sequence periodically due to the practical limitation. The simulation sequence is collected from PhysioNet[1].

Here we use $G[n]$ =[135, 140, 169, 215, 224, 201, 265, 252, 332, 325, 296, 240, 285, 294, 276, 273, 296, 286, 349], where the last entry is the latest measurement. And suppose the measurement interval is 1 day. The $HbA1c$ is computed as 10.5%[2] by prediction from the mean of glucose concentration $\mathbb{E}[G] \approx 255.42$mg/dL.

2.3.1 Uniform distribution

Suppose $f_{RBC_{age}}(t)$ follows from uniform distribution with mean M_{RBC},

$$f_{RBC_{age}}(t) = \frac{1}{2M_{RBC}} \qquad 0 \le t \le 2M_{RBC}.$$

For approximation, through binning the domain of $f_{RBC_{age}}(t)$, we have

$$\sum_{n=0}^{M_{RBC}-1} \tilde{v}(n) \times 1 \times \frac{2M_{RBC} - n}{2M_{RBC}} = HbA1c \tag{4}$$

where $\tilde{v}(n) = \frac{\tilde{G}(n)20\%}{250\text{mg/dL}\times 120\text{d}}$ and $\tilde{G}(n) = G(-n)$.

This results in $M_{RBC} \approx 58$ days, thus showing

$$f_{RBC_{age}}(t) = \frac{1}{116} \qquad 0 \le t \le 116.$$

Because we have the similar assumed value with G, the result is no wonder related to the oldest age of cells.

NOTE: under this distribution assumption, there is only one parameter that needs to be determined, so only one measurement of $HbA1c$ is needed.

[1] https://physionet.org/physiobank/database/mimic2cdb-ps/s20794/#234

[2] https://www.uptodate.com/contents/calculator-glycemic-assessment-using-conventional-or-si-units-for-hemoglobin-a1c

3 Questions - March 30

1. Should the G term in the formula be $G(t-a)$ instead of $G(-a)$ in the original word text?

$$HbA1c(t) = \int_0^\infty G(t-a)\left[1 - \int_0^a f_{RBC_{age}}(\tau)\mathrm{d}\tau\right]\mathrm{d}a \tag{5}$$

2. Is this formula equal to the practical measurement up to a constant? If yes, so how do we deal with this constant? (Suppose G is a constant, and M is the mean of a uniform distribution, then H = GM where the true mean $M_{age} = \frac{M}{\alpha}$. Ideally set the maximum of the support to be 120 days and then do a scale).

3. Can we use monte carlo integration to solve this problem? Is there any relationship between monte carlo integration with sampling theorem?

4. How do we define the measurement at t=0, i.e., HbA1c(0)?

5. Is this formula equal to the practical measurement up to a constant?

6. When we mention to "measure multiple HbA1c", are we referring to multiple measurements on the same blood sample or multiple measurements on different blood samples?

7. At this time, if we take G(t) as a constant and assume the shape of f, it seems solvable for uniform assumption and finite exponential assumption. But in practice, because both G and f are changing, and every time there are new signals entering, I can not see the process of taking deconvolution method, can you give me some more hints on this method? Or do we need to take other assumptions?