MODEL OF ^{131}I BIOKINETICS IN THYROID GLAND AND ITS IMPLEMENTATION FOR ESTIMATION OF ABSORBED DOSES

Ladislav Jirsa*, Ferdinand Varga**, Miroslav Kárný* and Jindřiška Heřmanská***

* Institute of Information Theory and Automation, Department of Adaptive Systems, Czech Academy of Sciences, Prague, CZ
** 2nd Medical School, Institute of Biophysics, Charles University, Prague, CZ
*** Clinic of Nuclear Medicine and Endocrinology, Motol Hospital, Prague, CZ

jirsa@utia.cas.cz

Abstract: In nuclear medicine, estimation of doses absorbed in target tissues serves for both medical and radio-hygienical purposes. Dose is proportional to time integral of the target tissue activity. Therefore, a suitable specific model describing ability of the considered tissue to absorb radioactive substance, particularly thyroid gland and ^{131}I, is essential. The paper shows the bi-phasic model of time-activity course as a linear regression model for logarithm of activity with normal noise, estimated by Bayes methodology. We present algorithmic solution of the model identification using MCMC sampling, formulation of hard constraints of the prior probability density function (*pdf*) for unknown model parameters (regression coefficients) and testing of *pdf* of the time integrals for each parameter sample. The identification was validated on predictive ability of the model. We used 2 355 data sequences, each containing 4–9 pairs of time-activity mesured within a few days. The model was identified using 3 data pairs, activity of the 4th measurement was predicted. After omitting 0.81 % of data with gross measurement errors, the mean of relative prediction error is -0.0004, median -0.0544 and standard deviation 0.42. Because of prior information, even 2 measurements give reliable predictions in cases inspected. Distribution of the time integral logarithms is approximated by the normal *pdf*.

Introduction

Use of radioactive iodine ^{131}I is one of the steps in diagnosis and treatment of thyroid gland carcinoma, among biochemical, endocrinological, histological and other examinations. ^{131}I is used for imaging and destruction of thyroid-like tissues accummulating iodine, either as elimination of thyroid remnants after removing the tumor by surgery or radiodestruction of tumor re-appearence or metastases [1].

The ^{131}I is administered to a patient orally as a water solution of sodium- or potassium-iodide salt. The radiologic procedure consists of two major steps:

- *Diagnosis:* A low (tracer) activity about 50–100 MBq is administered to a patient. Its distribution over the organism is monitored by a whole-body imaging using γ-camera and/or measurement of activity in selected regions. The measurements can be repeated in time to obtain activity sequence. This stage informs about accummulating sites, their size, activity kinetics etc. and helps in decision about subsequent treatment. The γ-radiation produced by ^{131}I is detected.
- *Therapy:* A high administered activity of 2–10 GBq causes destruction of accummulating tissues by their huge impact by β-radiation.

The activity decided for therapy must be balanced for sufficient destructive effect and, at the same time, for minimizing secondary radiation risks for the patient and medical staff as well. Individual response to ^{131}I activity administrated to the patient's organism requires individual mathematical processing of activity values in a tissue measured after the administration. The information gained is used for estimation of the absorbed radiation dose, for the treatment decision support, etc.

The standard methodology for doses estimation is MIRD [2] (Medical Internal Radiation Dose). It requires time integral of activity in thyroid gland. Due to limited number of diagnostic thyroid activity measurement A in time t practically available (usually 3 (A,t)-pairs within two days) and uncertainty of the data, adequate mathematical modelling of $A(t)$ must be used.

The bi-phasic model of activity A in time t was published e.g. in [3]. Although its dominance over classical mono-exponential model was obvious, its identification was very sensitive to quality and amount of measured data. Therefore, a robust approach had to be developed so that all the data available in clinical practice could be used.

Probabilistic modelling in combination with Bayes methodology appears to be suitable for solution of this task. Its strength is in respecting random nature of the data and in possibility to add expert knowledge (prior information) on the model parameters to be estimated. It proved to be successfully applicable to estimation tasks with a few noisy data, e.g. [4], [5].

There is a developed theory about prior information in linear regression models (e.g. [6], [7]). However, this

ISSN: 1727-1983

practical estimation task is specific by a very low number of data potentially (and quite frequently) loaded with other measurement errors than random. Bayes methodology furthermore leaves subjective space for the formulation of prior information. Therefore, a careful approach must be adopted to balance between the prior information improving the identification and a limited information carried by a few pieces of data that must not be distorted by any subjective influence.

Materials and Methods

The **aim** is to estimate probaility density function $f(\xi|\text{data, prior})$, where

$$\xi = \int_0^{+\infty} A(t)\ \mathrm{d}t. \tag{1}$$

Data are represented by a measured sequence $\{(t_i, A_{t_i})\}_{i=1}^{n}$, where $2 \leq n \lesssim 9$. Prior information will be discussed below.

The **bi-phasic model** of $A(t)$ [3] is a linear regression model with normal noise

$$\ln A(t) = k_1 + k_2 \ln t + k_3 t^{\frac{2}{3}} \ln t - \frac{t}{T_\mathrm{p}} \ln 2, \tag{2}$$

where $\vartheta \equiv (k_1, k_2, k_3)'$ is a vector of unknown parameters and T_p is a physical half-life of ^{131}I (8.04 days). Unit of activity is MBq, unit of time is day, $t > 0$.

The equation (2) can be formally rewritten as

$$d_t = \psi_t' \vartheta, \tag{3}$$

where $'$ means transposition, $d_t = \ln A(t) + t/T_\mathrm{p} \ln 2$ and $\psi_t = (1,\ \ln t,\ t^{2/3} \ln t)'$. Let us denote data vector $\Psi_t = (d_t,\ \psi_t')'$.

According to a general theory of Bayesian estimation of the linear regresion model with normal noise [8], [9], the **posterior *pdf*** of ϑ is

$$f(\vartheta|L,D,\nu)\ =\ \mathscr{I}^{-1}(L,D,\nu) \times \left[1 + \left({}^{\lfloor d}D\right)^{-1} \left(\vartheta - \hat{\vartheta}\right)' {}^{\lfloor \psi}L' {}^{\lfloor \psi}D {}^{\lfloor \psi}L \left(\vartheta - \hat{\vartheta}\right)\right]^{-\frac{1}{2}(\nu + \mathring{\psi})}, \tag{4}$$

where $\mathscr{I}^{-1}(L,D,\nu)$ is a normalizing constant, $\hat{\vartheta} = {}^{\lfloor \psi}L^{-1}\, {}^{\lfloor d\psi}L$ is the least squares estimate of $\mathscr{E}\vartheta$, $\mathring{\psi} = 3$ is length of the regression vector and $V \equiv L'DL$ and ν are sufficient statistics evolving in time t like

$$\begin{aligned} V_t &= V_{t-1} + \Psi_t \Psi_t' \\ \nu_t &= \nu_{t-1} + 1, \end{aligned} \tag{5}$$

where V_0 and ν_0 are prior statistics. V is called extended information matrix. The decomposition $V = L'DL$, where L is a lower triangular matrix with unit diagonal and D is a diagonal matrix, is used because of numerical stability and computational comfort. Fast algorithms for direct update of L and D without explicit construction of V can be found e.g. in [9]. The matrix V, as well as L and D, is decomposed into a scalar ${}^{\lfloor d}V$, a $\mathring{\psi} \times \mathring{\psi}$-matrix ${}^{\lfloor \psi}V$ and a $\mathring{\psi} \times 1$-column vector ${}^{\lfloor d\psi}V$:

$$V = \begin{pmatrix} {}^{\lfloor d}V & {}^{\lfloor d\psi}V' \\ {}^{\lfloor d\psi}V & {}^{\lfloor \psi}V \end{pmatrix}. \tag{6}$$

Covariance of ϑ is given by

$$\mathrm{cov}(\vartheta|L,D,\nu) = \frac{{}^{\lfloor d}D}{\nu - 2}\ {}^{\lfloor \psi}L^{-1}\, {}^{\lfloor \psi}D^{-1} \left({}^{\lfloor \psi}L'\right)^{-1}. \tag{7}$$

The *pdf* (4) is a marginal of a joint *pdf* of ϑ and model noise (Gauss-inverse-Wishart (*GiW*) *pdf*). The notation used in (4) guarantees existence of first two moments of noise in *GiW*.

The **prior information** can be formulated as constraints for the parameter vector ϑ and as values for the prior statistics V_0 and ν_0.

Let us now formulate **constraints for** ϑ. The function $A(t)$ must fulfil the following requirements:

(1) $A(t) = 0$ for $t = 0$ and $t \to +\infty$,
(2) $A(t)$ has a single global maximum $A(t_\mathrm{max})$ in t_max,
(3) $t_\mathrm{d} < t_\mathrm{max} < t_\mathrm{u}$, where, according to medical experience [10], $t_\mathrm{d} = 4$ hours (0.167 days) and $t_\mathrm{u} = 72$ hours (3 days),
(4) given $t_1 > t_\mathrm{max}$, $A(t)$ decreases for $t > t_1$ faster than decrease caused by a simple physical decay, i.e. by the term $-\frac{t}{T_\mathrm{p}} \ln 2$.

The first two requirements are fulfilled if $k_2 > 0$ and $k_3 < 0$. The straightforward solution of the rest of the requirements is exactly analytically impossible because of form of (2). We will present an approximate solution by analyzing $g(t)$ as $\ln A(t) = g(t) - t/T_\mathrm{p} \ln 2$. Putting the first derivative of $g(t)$ equal zero gives

$$k_2 + k_3 t^{\frac{2}{3}} \left(\frac{2}{3} \ln t + 1\right) = 0 \tag{8}$$

with solution denoted t_1b. Requiring $t_\mathrm{d} < t_\mathrm{1b} < t_\mathrm{u}$ and considering conditions above, we get

$$-k_3 t_\mathrm{d}^{\frac{2}{3}} \left(\frac{2}{3} \ln t_\mathrm{d} + 1\right)\ <\ k_2\ <\ -k_3 t_\mathrm{u}^{\frac{2}{3}} \left(\frac{2}{3} \ln t_\mathrm{u} + 1\right). \tag{9}$$

The analysis says that for $k_3 < 0 < k_2$ and $t < t_0 \equiv \exp(-3/2)$ days ≈ 5 hours 21 mins, $g(t)$ is always increasing, therefore, t_0 must be substituted into (9) instead of t_d. To include the decay term, $t/T_\mathrm{p} \ln 2$ with corresponding values of t_d and t_u must be added to the leftmost and rightmost side of (9), i.e.

$$0.019 < k_2 < -3.6\, k_3 + 0.26. \tag{10}$$

All these conditions can be written in the linear form

$$A\ \vartheta < b,$$

$$A = \begin{pmatrix} 0 & 1 & 3.6 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0.26 \\ -0.019 \\ 0 \end{pmatrix}. \tag{11}$$

ISSN: 1727-1983

This constraint ensures the requirements (1)–(4) fulfilled. The constrained posterior *pdf* $\tilde{f}(\vartheta|L,D,\nu)$ is obtained from (4):

$$\begin{aligned}\tilde{f}(\vartheta|L,D,\nu) &= f(\vartheta|L,D,\nu)\,\chi(\vartheta) \qquad (12)\\ \chi(\vartheta) &= 1 \quad \text{if } A\,\vartheta < b,\\ &\quad 0 \quad \text{otherwise.}\end{aligned}$$

The modified normalization constant neednot be considered, as shown below.

Let us now focus on **prior statistics** V_0 and ν_0. The element ${}_{\lfloor d}D$, which is the least-squares remainder, is proportional to the model noise and to covϑ in (7). To keep the prior covariance high, either ${}_{\lfloor d}D$ must be high, which will negatively influence the noise even after real data update, or terms of ${}_{\lfloor \psi}D$ must be low. This is in agreement with [7], where *fictitious data* used to create $V_0 \equiv L_0' D_0 L_0$ are scaled by a coefficient, representing *weight* or *belief* in these data, and the term ${}_{\lfloor d}D$ (scaled as well) is then additionally increased by assumed (i.e. usual) noise. As the fictitious data, we used three pairs of average data: two within the first day after administration and the third one after several days when activity of thyroid gland is almost zero. The vectors of these fictitious data Ψ_{fict} were multiplied by a scale factor 0.01 to make the prior flat enough but to keep a prior point estimate of ϑ in a range leading to meaningful courses of $A(t)$. The term ${}_{\lfloor d}D_0$ was additionally increased by 3×0.0015 which was an average noise contributed by one data pair. The value of ν_0 was chosen 1.05 as best performing. However, topic of prior information is still a matter of research.

Now, the **algorithmic solution** of the model identification will be described. For convenience and numerical reasons, the parametric space of ϑ was linearly transformed into ϑ^* so that the (unconstrained) posterior *pdf* (4) had zero mean and unit covariance

$$T = \sqrt{\frac{\nu-2}{{}_{\lfloor d}D}}\;{}_{\lfloor \psi}D\;\;{}_{\lfloor \psi}L, \qquad (13)$$

$$\vartheta^* = T(\vartheta - \hat{\vartheta}). \qquad (14)$$

where $\sqrt{{}_{\lfloor \psi}D}$ means square root by elements.

The constraint (11) was transformed

$$A^* = AT^{-1}, \qquad (15)$$

$$b^* = b - A\hat{\vartheta}, \qquad (16)$$

so that $A^*\,\vartheta^* < b^*$ is corresponds to (11).

The transformed *pdf* (12) was sampled using Langevin diffusion algorithm [11]. This improvement of random walk Metropolis-Hastings algorithm shifts the proposed sample deterministically towards the gradient which increases the acceptance rate and makes convergence faster. The normalizing constant $\mathscr{I}^{-1}(L,D,\nu)$ in (4) is not required in the sampling algorithm.

However, the analysis of MC properties including derivation of optimum chain step size [12] assumes membership of the posterior *pdf* in exponential family, existence of one-dimensional factors of the *pdf* and unconstrained domain. Unfortunately, all of these assumptions are violated in this task.

A rule for the optimum step size was constructed experimentally by numerical analysis of 3 876 data sequences. The rule is quite complex, although algebraically simple. We observed that the optimum step size depends on the statistics ν and distance of the unconstrained posterior mean from the planes (11) in the transformed space. Testing of the heuristic rule proved performance of MCs close to their optimum.

Initial point of MC was chosen by optimization of the quadratic form in the denominator of (4) with the constraints (11). 5 000 samples were found sufficient after 500 of burn-in. Each parameter sample ϑ_j^* was, after the inverse transformation, substituted into (2) and integral ξ_j (1) was computed from 0 to 70 days using the algorithm QUANC8 [13]. The samples ξ_j created a histogram, distribution of which was tested.

Kolmogorov-Smirnov test of normality was applied to the samples ξ_j and $\ln \xi_j$ for each data sequence. Also, skewness for both sets of samples was computed.

Results

First, prediction ability of the model was tested both with and without the prior information. 2 355 data sequences that contained at least 4 data pairs were selected. 3 data pairs were used to identify the model and the 4th one, usually following after 1–3 days, was predicted. This choice is justified by the fact that after a diagnostic administration, usually not more that 3 measurements are performed.

Without the prior constraints, 40 % of data sequences had to be excluded as they did not lead to estimates matching the requirements for physical behaviour of $A(t)$. No prior statistics were used in this case. Although the prediction is the best, number of outlied predictions (relative error >3) is high. Then, the prior constraints were considered, either with empty or non-empty prior statistics. As expected, all the data sequences led to meaningful estimates. The case with prior statistics performs lower both relative prediction error and its standard deviations. These results are shown in the Table 1. Standard deviations are similar in all the cases, although the best in case 3.

Table 1: Relative prediction errors in cases: 1) no prior constraints, 2) prior constraints and empty prior statistics, 3) prior constraints and statistics

#	mean	median	st.dev.	data	outliers
1)	0.0576	−0.0066	0.475	1 403	2.28 %
2)	−0.0968	−0.1456	0.431	2 355	0.85 %
3)	−0.0004	−0.0544	0.416	2 355	0.81 %

ISSN: 1727-1983

Next, distribution $f(\xi)$ in (1) was tested. Kolmogorov-Smirnov test of normality did not give satisfactory significance level of null hypothesis (i.e. data are normal) neither for ξ nor for $\ln \xi$. Another approximate test was done by comparing skewness of both $f(\xi)$ and $f(\ln \xi)$ after excluding samples out of $\hat{\xi} \pm 3\sigma$. The comparison is shown in Table 2.

Table 2: Skewness of distributions $\{f(\xi_j)\}$ and $\{f(\ln \xi_j)\}$

$f(\cdot)$	**mean**	**median**	**st.dev.**
ξ_j	1.66	0.84	3.53
$\ln \xi_j$	0.29	0.24	0.61

Although these results correspond to the non-significant levels of the Kolmogorov-Smirnov test, it is obvious that normal *pdf* (with zero skewness) describes better the distribution of $\ln \xi$. Practical experience shows that such an approximation is sufficient with respect to existing uncertainty.

Figure 1 shows an example of $A(t)$ identified on 2 initial data pairs, other data are predicted.

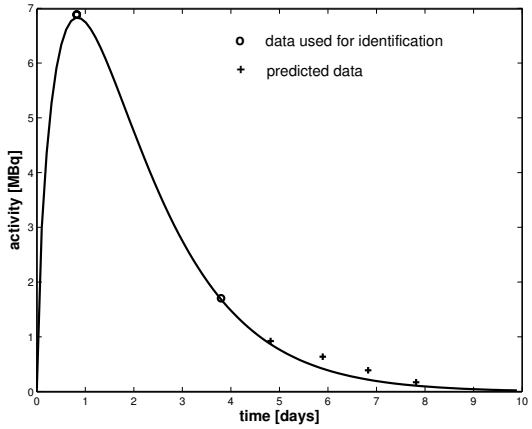

Figure 1: Example of $A(t)$ (one sample) identified on 2 data pairs

Discussion

The bi-phasic model (2) has been so far considered as the best known tool for modelling of activity kinetics in thyroid gland. The Bayes methodology used for its probabilistic identification has advantage in decreasing uncertainty by prior information. However, the freedom in its construction requires awareness, on the other hand, original and specific approaches can be applied.

It was shown that the values of prior statistics, chosen in accordance to the recent theory, can improve predictive abilities of the model. Although standard deviations of relative prediction errors seem high (above 40 % of activity magnitude), we must take into accout limited quality and relatively high uncertainty of measured data. Radioactive decay is a random process which becomes significant especially in case of low activities in diagnostic administration and late measurements.

Analytical approximation of the resulting distribution $f(\xi)$ is another subject of more testing. Nevertheless, its log-normal form seems to be sufficient in practice.

Algorithmic solution appears robust and stable. Unfortunately, impossibility of analytic transformation of $f(\vartheta)$ into $f(\xi)$ requires numerical transformation which was outlined in the paper. Although the numerical procedure seems complicated, on contemporary PCs one identification takes 1–2 seconds in MATLAB and fractions of seconds in C++, which makes possible to use it on-line in clinical practice.

Conclusions

Robust and stable probabilistic identification of bi-phasic model of thyroid activity $A(t)$ after ^{131}I administration was presented. The model was tested using prediction of future data. Prior information guarantees physical meaningfullness of $A(t)$ and increases precision of the predicitons.

To estimate distribution of absorbed dose evaluated according to the MIRD methodology, the *pdf* $f(\xi)$ is directly applicable because it depends linearly on the dose.

It was observed that reliable predictions are given even after 2 measurements. Due to this, the model can be used for prediction of time interval when a required activity would be reached, which might be important for planning of a patient's stay in the clinic, or for checking correctness of data typed by staff into the software supporting activity measurements.

Further work would focus on improvement of prior information, better analysis of convergence and more exact analytical approximation of $f(\xi)$. Initial phase of $A(t)$ as well as overall performance might be improved by change of the time scale and, potentially, activity scale which is under consideration.

Use of the model can contribute to treatment quality, radiation protection and quality of future data.

Acknowledgements

This research was partially supported by the grants AV ČR 1ET 1007 50404 and MŠMT ČR 1M6 798 555 601.

References

[1] NĚMEC J.: *Treatment of Thyroid Diseases by Radioiodine*. Avicenum, Prague, 1974. In Czech.

[2] LOEVINGER R., BUDINGER T. F., and WATSON E. E.: *MIRD Primer for absorbed dose calculations*. The Society of Nuclear Medicine, New York, 1988.

[3] HEŘMANSKÁ J., KÁRNÝ M., ZIMÁK J., JIRSA L., ŠÁMAL M., and VLČEK P.: Improved prediction of therapeutic absorbed doses of radioiodine in the treatment of thyroid carcinoma. *Journal of Nuclear Medicine*, 42(7):1084–1090, July 2001.

[4] FONSECA C. M.: Bayesian estimation of the intensity of low-level radiation sources. *Jaderná energie*, 37:83–97, 1991.

[5] HEŘMANSKÁ J. and KÁRNÝ M.: Bayesian estimation of effective half-life in dosimetric applications. *Computational Statistics and Data Analysis*, 24(5):467–482, 1997.

[6] KÁRNÝ M., KHAILOVA N., NEDOMA P., and BÖHM J.: Quantification of prior information revised. *International Journal of Adaptive Control and Signal Processing*, 15(1):65–84, 2001.

[7] KRACÍK J. and KÁRNÝ M.: Merging of data knowledge in Bayesian estimation. In *ICINCO*, Barcelona, September 2005. IFAC.

[8] PETERKA V.: Bayesian system identification. In P. Eykhoff, editor, *Trends and Progress in System Identification*, pages 239–304. Pergamon Press, Oxford, 1981.

[9] KÁRNÝ M., BÖHM J., GUY T. V., JIRSA L., NAGY I., NEDOMA P., and TESAŘ L.: *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, London, 2005. ISBN 1-85233-928-4, pp. 552.

[10] HEŘMANSKÁ J.: *Bayesian Approach to Dosimetric Data Evaluation for Medical Use of* ^{131}I. Clinic of Nuclear Medicine, 2nd Medical Faculty, Charles University, Prague, 1993. Associated Professor Thesis, 103 pp. In Czech.

[11] ROBERTS G. O. and TWEEDIE R. L.: Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[12] ROBERTS G. O. and ROSENTHAL J. S.: Optimal scaling of discrete approximation to Langevin diffusions. *J. R. Statist. Soc.*, 60, Part 1(B):255–268, 1998.

[13] FORSYTHE G. E., MALCOLM M. A., and MOLER C. B.: *Computer Methods for Mathematical Computations*. Prentice Hall, 1977. Russian translation, Mir, Moscow 1980.

ISSN: 1727-1983