Identification of Thyroid Gland Activity in Radioiodine Therapy

Ladislav Jirsa*

Institute of Information Theory and Automation, The Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Praha 8, Czech Republic, +420 2 6605 2337

Ferdinand Varga

Simulation Education Center, Jessenius Faculty of Medicine, Comenius University, Novomeského 7a, 036 01 Martin, Slovak Republic

Anthony Quinn

Department of Electronic and Electrical Engineering, Trinity College Dublin, the University of Dublin, Dublin 2, Republic of Ireland

Abstract

The Bayesian identification of a linear regression model (called the biphasic model) for time dependence of thyroid gland activity in ^{131}I radioiodine therapy is presented. Prior knowledge is elicited via hard parameter constraints and via the merging of external information from an archive of patient records. This prior regularization is shown to be crucial in the reported context, where data typically comprise only two or three high-noise measurements. The posterior distribution is simulated via a Langevin diffusion algorithm, whose optimization for the thyroid activity application is explained. Excellent patient-specific predictions of thyroid activity are reported. The posterior inference of the patient-specific total radiation dose is computed, allowing the uncertainty of the dose to be quantified in a consistent form. The relevance of this work in clinical practice is explained.

Keywords: biphasic model, prior constraints, external information, Langevin diffusion, nonparametric stopping rule, probabilistic dose estimation

1. Radioiodine Therapy for Thyroid Gland Cancer

The thyroid gland [1] is located in the neck. It is an important component of the endocrine system. Specific thyroid cells bind and accumulate free

*Corresponding author

Email addresses: jirsa@utia.cas.cz (Ladislav Jirsa), Ferdinand.Varga@jfmed.uniba.sk (Ferdinand Varga), aquinn@tcd.ie (Anthony Quinn)

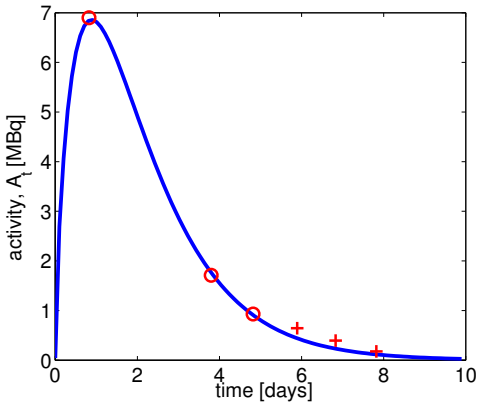

Figure 1: A typical patient activity curve, A_t, identified using 3 patient measurements (circles). The remaining measurements (crosses) are used to quantify prediction error.

iodine from the blood. Accumulated iodine is used in the synthesis of thyroid hormones. These hormones affect the body in the following ways: metabolic, thermoregulatory, growth and maturation.

While in 1987, when thyroid cancer affected about 5 in every 100 000 people in United States, 80 % of them female, in 2009 it was 14 in 100 000 [2]. In therapy, the thyroid is typically removed by surgery. However, it is impossible to remove the organ completely, owing to the proximity of the vocal chords, important arteries and nerves. Hence, in normal clinical practice, these remnants—along with any metastases (which, in common with the thyroid itself, are also iodine-accumulating)—are then destroyed by methods of nuclear medicine (radioiodine therapy).

Radioiodine therapy for thyroid gland cancer [3] exploits the fact that the gland selectively accumulates iodine from the blood. Nuclear decays in unstable (radioactive) ^{131}I release β-particles (electrons) which are absorbed by the thyroid tissue (as well as by other organs). Therapeutic administration of ^{131}I is typically in the activity range of 2–10 GBq[1], leading to radio-destruction of the thyroid tissue. The accompanying γ-particles (high energy photons) are not absorbed by the tissue and can therefore be detected outside the body. Typically, there is a preliminary diagnostic administration of ^{131}I, at an activity of 70 MBq, in order to assess the mass and disposition of the thyroid remnants, and to provide guidance in the design of the subsequent therapeutic administration.

The ^{131}I *activity*, A_t, of the thyroid, at a time t (days) following administration of ^{131}I, is defined as the mean number of nuclear decays (nuclear decay is a random Poisson-distributed process) occurring in the gland per second at time t. A typical activity curve is illustrated in Figure 1. It reveals the charac-

[1] 1 Giga-Becquerel (GBq) corresponds to 10^9 nuclear decays per second.

teristic *biphasic* (*i.e.* two-phase) behaviour, comprising the initial *uptake* phase, followed by the *clearance* phase. Note that the time-scale is far shorter than that for radio-destruction and elimination of the tissue by the immune system, which takes 3–6 months. Hence, the clearance is due dominantly to the radioactive decay of ^{131}I and metabolic elimination of the isotope by the thyroid. The key therapeutic quantity of interest is the *absorbed dose*, $\mathcal{D}$, defined as the total energy of the β-particles absorbed per unit mass of the thyroid:

$$\mathcal{D} = \mathcal{S}\xi, \qquad \xi = \int_0^{+\infty} A_t \ \mathrm{d}t. \tag{1}$$

Here, $\mathcal{S}$ is a known organ- and isotope-specific constant, provided by the MIRD methodology (Medical Internal Radiation Dose) [4].

1.1. The Measurement Process

The β-particles—and hence A_t—cannot be measured directly. However, the associated γ-particles (photons) released by the thyroid during one-second intervals around a measurement time, t, can be detected and counted by a scintillation probe at a specific range and direction [1, 5]. A matrix of such counts (*i.e.* a scintigram) is available if an array of such probes—known as a γ-camera—is used. The cumulative count in a Region-of-Interest (ROI) marked on the scintigram by the radiologist is then available at the measurement time, t. In standard radiological practice, the measured background count due to sources other than the thyroid itself is then subtracted, to yield an estimated count, $\hat{n}_t$, of particles from the thyroid. A calibration step then converts $\hat{n}_t$ into an estimate, d_t, of the thyroid activity, A_t, at the measurement time, t. The calibration is achieved using a source of known activity in the same geometrical arrangement as the patient and probe/camera. The calibration-adjusted estimate, d_t (MBq), is called the *measured activity* of the thyroid, and is the conventional statistic computed in standard radioiodine therapeutic practice. Details of this activity estimation procedure are provided in [6]. For a specific patient, the available data, D, are therefore the set of measurement times, t_i, and the associated measured activities, d_{t_i}:

$$D \equiv \{(t_i, d_{t_i})\}_{i=1}^{n},$$

where i is the discrete-time index and n is the number of data recorded for the specific patient[2].

[2]The maximum measured activity for each specific patient, which we denote by d_{m} (we omit any patient-specific index for the time being), can differ by several orders of magnitude within a population of patients, such as the one studied in Section 4.2. This is due to differences in administered activity of ^{131}I and metabolic variations between patients. For reasons of numerical stability in the Bayesian identification algorithm (Section 3), scaled measured activities, $d_{t_i}/d_{\mathrm{m}} \in (0, 1]$, are modelled for each patient. For notational simplicity, it is these scaled quantities that will be referred to as d_{t_i} in the sequel.

1.2. The Key Inference Tasks

The ability of thyroid remnants to accumulate iodine depends on the size of the remnants after surgery, the type of carcinoma, the patient's metabolism, the possible presence of metastases, *etc.* Therefore, *patient-specific* inference is of great clinical importance, both at the diagnostic and therapeutic stages.

Therefore, two key inference tasks are addressed in this paper:

1. Patient-specific sequential prediction of measured activity, d_t. There are two uses for these predictions: the first is to validate the parametric model that we will adopt for A_t in Section 2.1; and a second potential use is to provide a tool for quality assurance during logging of measured activities (*i.e.* if the recorded value differs significantly from the predicted one, a warning is generated).
2. Patient-specific inference of ξ and hence the absorbed dose, $\mathcal{D}$ (Section 1). This is the key therapeutic quantity determining the effectiveness of the radioiodine therapy and hence the patient's prognosis. In particular, we wish to quantify the uncertainty in $\mathcal{D}$, since this supports the radiologists in their planning of possible follow-up treatment for the patient. Furthermore, the thyroid acts as a radiation source during radioiodine therapy. β-particles from the thyroid irradiate the blood, while the associated γ-particles irradiate remote organs. Inference of $\mathcal{D}$ allows the radiologist to assess the levels of such irradiation. Note that distributions of non-patient-specific dose have been proposed in the radiation protection literature [7, 8]. Recently, the EANM Dosimetry Committee Series, *Standard Operational Procedures for Pre-Therapeutic Dosimetry* [9], provided guidelines on the assessment of patient-specific absorbed dose, but this was non-probabilistic. To our knowledge, no reference, beyond the work reported here, provides a patient-specific probabilistic inference of dose in radioiodine therapy.

A difficult inference regime is implied for the following reasons:

1. for economic reasons, and to avoid possible distress to patients, only a small number, $2 \leq n \lesssim 9$, of non-uniformly sampled measurements, d_{t_i}, are available per patient;
2. these measured activities are subject to considerable uncertainty (noise), due to imprecise calibration of the measurement system and uncertain background radiation levels.

The poor quality, and small quantity, of the available data point to the need for a Bayesian approach to the tasks above, as, succesfully, in similar situations, e.g. [10].

1.3. Structure of the Paper

In Section 2.1, the biphasic linear regression model for A_t is introduced, for which an elegant Bayesian conjugate framework is available (Section 3). A key benefit of the Bayesian approach in this case is that it provides the opportunity

to improve the patient-specific inference using an available database of measured activities for a large population of patients. In Section 4, we use these historic data, as well as known parameter constraints, to construct a suitable prior for the biphasic model parameters. The posterior inference is deduced in Section 5, and problems associated with its evaluation are outlined. Selection and tuning of an appropriate stochastic sampling algorithm for approximation of the exact inference is outlined in Section 6. The resulting activity prediction and dose inference are assessed for a population of actual patients in Section 7. The impact of the work on current clinical practice, and prospects for future work in the area, are discussed in Section 8.

2. Modelling of ^{131}I Activity

The uptake and clearance of ^{131}I by the thyroid is a topic in pharmacokinetics (PK), *e.g.* [11]. PK models have been proposed for quantifying the dose associated with inhalation [12] or ingestion [7] of ^{131}I, and for assessing its variability. In [8], the dose variability is evaluated and its distribution is assumed log-normal. In population PK, the individual pharmacokinetic parameters are studied across a patient population, *e.g.* [13]. However, we emphasize that the inference tasks which we defined in the previous Section are patient-specific, and so we do not concern ourselves with population PK models. Reported methods that are based on individual dosimetry, and on quantifying dose in individual ^{131}I-therapy patients (*e.g.* [4, 14]), do not provide measures of uncertainty. In contrast, in this paper, we develop a fully probabilistic, patient-specific inference of dose for the first time (Section 6).

Compartmental PK models for iodine activity, A_t, in the thyroid gland differ in the number of compartments and their purpose. The 1-compartment model is equivalent to a mono-exponential model for A_t (e.g. [15]), and so it omits the uptake phase (Figure 1). In our earlier work [15], the uptake phase was treated heuristically via a linear approximation. A 2-compartment model was used for the study of hyperthyroidism in [16], and it was also used in [9] as a reference model to evaluate precision of simpler methodologies. A 4-compartment model was used in [17] to model iodine metabolism, and a 6-compartment model was proposed in [18] to account for early uptake. Cyclic compartmental models, requiring more parameters, have also been proposed [19].

Recently, due to availability of a high computational power, physiologically-based pharmacokinetic models are frequently used, either for personalised medicine [20] or in population PK [21]. These models are typically sets of ordinary differential equations (possibly nonlinear) describing transport of a substance between organs (compartments) according to physiological processes. However, they tend to a high number of parameters, some of which can be correlated.

A simple 3-parameter linear regression model for A_t was proposed in [15]. This biphasic model was obtained as a functional approximation of A_t given by solution of a 4-compartment cyclic model [22] for ^{131}I, involving about 20 parameters. Its advantages are that (i) standard Bayesian methodology for recursive linear model identification [23] can be exploited; (ii) the model can be

identified even for the small number, n, of data encountered in clinical practice (Section 1.2); and (iii) good prediction of activity—even for these small datasets—was reported in [15], in contrast to the mono-exponential model whose predictions were highly sensitive to perturbations of the data, and to their number.

2.1. The Three-Parameter Biphasic Model for Thyroid Activity, A_t

The following 3-parameter biphasic model will be adopted:

$$\begin{aligned} \ln A_t &= a_1 + a_2 \ln(c\,t) + a_3 (c\,t)^{\frac{2}{3}} \ln(c\,t) - \frac{t}{T_{\rm p}} \ln 2 = \psi_t' a \ - \alpha t, \qquad (2) \\ \psi_t &\equiv \Big(1,\ \ln(c\,t),\ (c\,t)^{2/3} \ln(c\,t)\Big)'. \end{aligned}$$

Here, by convention, $t > 0$ is measured in days , $a = (a_1, a_2, a_3)'$ is a vector of unknown linear regression parameters[3] ($'$ denotes transposition), and ψ_t is the known regressor at time t. This model is an adaptation of the one first introduced in [15], to include a known time-scale factor, $c > 0$, whose value will be set in Appendix A. As we will see there, c will allow full exploitation of the biophysical requirements on the behaviour of the function A_t. The parameter-dependent term,

$$g_t \equiv \psi_t' a,$$

models the accumulation of ^{131}I by the thyroid, whereas the parameter-independent term, $-\alpha t$, $\alpha = \ln 2/T_{\rm p}$, models the radioactive decay (exponential) of the isotope itself, with $T_{\rm p}$ denoting the physical half-life of ^{131}I (8.04 days).

It was shown in [6] that the measured activity, $d_t > 0$ (Section 1.1), has an asymmetric distribution on a positive support, and is approximately log-normal with A_t as its first moment (mean). It follows that $\ln d_t$ has a Gaussian distribution, $\mathcal{N}(\mu, r)$. For $A_t \gg 0$, it follows that $\mu \approx \ln A_t$ [24]. The following approximate model for the measured activity, d_t, is therefore justified:

$$f(\ln d_t | A_t) = \mathcal{N}(\ln A_t, r). \qquad (3)$$

Here, $r > 0$ denotes the constant but unknown variance.

From (2), the implied parametric observation model is

$$\begin{aligned} x_t &\equiv \ln d_t + \alpha t \equiv \psi_t' a \ + e_t, \\ f(x_t | a, r) &= \mathcal{N}(\psi_t' a, r) = \frac{1}{\sqrt{2\pi r}} \ \exp\left\{ - \frac{(x_t - \psi_t' a)^2}{2r} \right\}. \qquad (4) \end{aligned}$$

$e_t \sim \mathcal{N}(0, r)$ is the additive residual representing the uncertainty (noise) in the background subtraction and calibration steps used to compute d_t (Section 1.1).

[3] Note that the model for the unscaled data of a specific patient (see Section 1.1) is trivially obtained by replacing a_1 by $a_1 + \ln d_{\rm m}$, where $d_{\rm m}$ is the maximum measured activity in the patient's data (see footnote 2).

It also quantifies the modelling error introduced by this simple 3-parameter model (2). The effect of unmodelled covariates—such as gender, age, metabolic factors, *etc.*—could be partially accounted for by introducing correlation in the process, e_t (*i.e.* a coloured innovations process [23]). The two main disadvantages of doing this are (i) the increased complexity of the model: the correlation structure would then need to be identified (*e.g.* in parametric form) for each patient, from just the 2 or 3 available data points; and (ii) the unavailability of a conjugate inference framework in this case (Section 3) [23]. For these reasons, we take e_t as independent and identically distributed (i.i.d.) at the distinct observation times, t_i. Note, finally, that since the γ-particles are released by *independent* nuclear decays in ^{131}I, and since the observation times, t_i, are distinct, the i.i.d. assumption is consistent with these aspects of the measurement process.

3. Bayesian Conjugate Inference for A_t

The conjugate distribution for the Normal observation model (4) is *Normal-inverse-Gamma* [25], $f(a, r|V, \nu) = \mathcal{N}i\mathcal{G}(V, \nu)$. Here, $\nu > 0$ is the *degrees-of-freedom* parameter, and V is the positive-definite *extended information matrix*, of dimension $(p+1)\times(p+1)$, where p is the length of a (*i.e.* $p = 3$ for the biphasic model (2)). For reasons of numerical stability and computational efficiency (see below), V is expressed via the LD-decomposition [23], as

$$V = L'\Lambda L,$$

where L is a lower triangular matrix with unit diagonal and Λ is a diagonal matrix with non-negative elements. L may be partitioned as

$$L = \begin{pmatrix} l_{11} & 0 \\ l_{a1} & L_{aa} \end{pmatrix},$$

where $l_{11} = 1$ is the $(1, 1)$ element. Similarly, Λ may be expressed via a partition into λ_{11} and Λ_{aa}. The $\mathcal{N}i\mathcal{G}$ distribution can then be expressed as follows:

$$f(a, r|V, \nu) \equiv f(a, r|L, \Lambda, \nu) \propto r^{-\frac{\nu}{2}} \exp\left\{-\frac{1}{2r}\left[(L_{aa}a - l_{a1})' \Lambda_{aa} (L_{aa}a - l_{a1}) + \lambda_{11}\right]\right\}.$$

The distribution is proper if $\nu > p + 2 = 5$, in which case the normalizing constant, ζ, is available in closed form [23].

The first moment and second central moment of a and r, respectively, are as follows [23]:

$$\begin{aligned} &\mathsf{E}[a] = L_{aa}^{-1} l_{a1} \;\equiv\; \hat{a}, && \mathsf{E}[r] = \tfrac{\lambda_{11}}{\nu-p-4} \equiv \hat{r}, \\ &\mathsf{cov}[a] = \hat{r}\; L_{aa}^{-1} \Lambda_{aa}^{-1} \left(L'_{aa}\right)^{-1}, && \mathsf{var}[r] = \tfrac{2\hat{r}^2}{\nu-p-6}. \end{aligned} \tag{5}$$

Finally, from (4), and noting the linear dependence of $\ln A_t$ on a (2), the log of the measured activity (3) is Student-t, yielding the following predictor[4]:

$$\begin{aligned} \mathsf{E}_{f(d_t|V,\nu)}\left[\ln d_t\right] &= \psi_t'\hat{a} - \alpha t, \\ \mathsf{var}_{f(d_t|V,\nu)}\left[\ln d_t\right] &= \hat{r}\left[1 + \psi_t' L_{aa}^{-1}\Lambda_{aa}^{-1}\left(L_{aa}'\right)^{-1}\psi_t\right] \equiv \hat{r}\left[1+\rho_t\right] \equiv \hat{r}_t. \end{aligned} \tag{6}$$

3.1. The Marginal Distribution of a

The marginal distribution of a is of the Student type [23],

$$f(a|L,\Lambda,\nu) \propto \left[1 + \lambda_{11}^{-1}\left(a-\hat{a}\right)' L_{aa}'\Lambda_{aa}L_{aa}\left(a-\hat{a}\right)\right]^{-\frac{1}{2}(\nu-2)}, \tag{7}$$

using (5). Once again, the normalizing constant, ζ, is available in closed form. The transformed variable,

$$a^* = T(a-\hat{a}), \qquad T = \sqrt{\frac{\nu-p-4}{\lambda_{11}}\,\Lambda_{aa}}\; L_{aa},$$

$\nu > p+4$, has zero mean and identity covariance matrix, a property which we will exploit in Section 6. Here, $\sqrt{\Lambda_{aa}}$ denotes the element-wise square-root.

3.2. The Conjugate Update

Let the prior also be the conjugate Normal-inverse-Gamma distribution, *i.e.* $f(a,r|\overline{V},\overline{\nu}) = \mathcal{N}i\mathcal{G}(\overline{V},\overline{\nu})$, where $\overline{V}$ and $\overline{\nu}$ are prior statistics. From (4), we define the *extended regressor* at observation time, t_i:

$$\Psi_{t_i} \equiv (x_{t_i}, \psi_{t_i}')'. \tag{8}$$

The posterior distribution is then $f(a,r|D) = \mathcal{N}i\mathcal{G}(V_n,\nu_n)$, where

$$\begin{aligned} V_n &= \overline{V} + \sum_{i=1}^{n}\Psi_{t_i}\Psi_{t_i}', \\ \nu_n &= \overline{\nu} + n, \end{aligned}$$

and $V_n = L_n'\Lambda_n L_n$, as above. To avoid the effects of rounding errors, L_n and Λ_n are, in fact, updated directly via the Ψ_{t_i}, ensuring positive-definiteness of V_n [23].

Note that $\lambda_{11,n}$, the $(1,1)$-element of Λ_n, is an offset least-squares remainder,

$$\lambda_{11,n} = \overline{\lambda}_{11} + \sum_{i=1}^{n}(x_{t_i} - \psi_{t_i}\hat{a})'(x_{t_i} - \psi_{t_i}\hat{a}),$$

where $\overline{\lambda}_{11}$ is the offset from $\overline{V}$.

[4] The unscaled log-data are predicted by adding $\ln d_m$ to the quantity in (6) (see footnote 2).

4. Construction of the Parameter Prior

In this thyroid activity context, prior information about the parameters, $\Theta = (a', r)'$ (4), is available from two independent sources (represented by Jeffreys' notation):

$\mathcal{I}_c$, a set of constraints specified by the radiologist, in order that any activity curve, A_t, be physically realizable (Figure 1), as explained in Section 4.1 below. This will be expressed via an appropriate prior, $f(a|\mathcal{I}_c)$. Since no information on the magnitude of r is available in advance, no prior constraints are imposed on r, beyond $r > 0$.

$\mathcal{I}_0$, an archive of measured thyroid activities for members of a population of ^{131}I-therapy patients; in Section 4.2, this will be merged into the conjugate, data-informed prior,

$$f(a, r|\mathcal{I}_0) = f(\Theta|\mathcal{I}_0) = \mathcal{N}i\mathcal{G}(V_0, \nu_0), \tag{9}$$

where $\mathcal{I}_0$ is merged via the prior parameters, V_0 and ν_0.

4.1. Hard Parameter Constraints, $\mathcal{I}_c$: Physical Properties of A_t

We consider prior limitations on the parameters, a, of the biphasic model (2), imposed by the following prior physiological constraints on the activity of the thyroid A_t, (see Figure 1):

1. $A_t \to 0^+$ as $t \to 0^+$, and as $t \to +\infty$;
2. A_t achieves a unique global maximum at some $t_\mathrm{m} > 0$;
3. medical experience [15] dictates that $t_\mathrm{m} \in (t_\mathrm{l}, t_\mathrm{u})$, where $t_\mathrm{l} = 4$ hours (0.167 days) and $t_\mathrm{u} = 72$ hours (3 days);
4. for some $t_\mathrm{h} > t_\mathrm{m}$, then A_t decreases for $t > t_\mathrm{h}$ faster than the decrease caused by physical decay of ^{131}I (the latter being represented by the term, $-\alpha t$, in (2)).

The resulting inequalities (see Appendix A), along with $a_3 < 0$ from (A.1), confine a to a convex domain, $\mathbb{A}$, via a linear matrix inequality, as follows:

$$a \in \mathbb{A} \equiv \{a \mid Ma < b\}, \qquad M = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 4.8687 \\ 0 & -1 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 0.2586 \\ -0.0144 \end{pmatrix}. \tag{10}$$

Here, '<' denotes element-wise inequalities. The prior,

$$f(a|\mathcal{I}_c) \propto \chi_{\mathbb{A}}(a), \tag{11}$$

is a conservative quantification of this prior knowledge, $\mathcal{I}_c$. Here, χ denotes the indicator function on the set. Since $\mathbb{A}$ has infinite measure, the prior (11) is improper.

4.2. Historic Data, $\mathcal{I}_0$: the Patient Archive

There exists an archive of activity measurements for a large population of thyroid cancer patients treated with ^{131}I at Motol Hospital, Prague, Czech Republic. From this archive, 3 876 datasets, D_j, $j = 1, \ldots, 3\,876$, were chosen, each containing a variable number, $2 \leq n_j \leq 10$ of data pairs, $\left\{(t_i^j, d_{t_i}^j)\right\}_{i=1}^{n_j}$ (Section 1.1). We emphasize that the task in our work is to infer the activity, A_t, of a specific (new) patient. However, this historic data constitutes *external information*, $\mathcal{I}_0$, which can be exploited in the patient-specific inference. This external information is represented by statistics V_0 and ν_0. These statistics, together with $\overline{V}$ and $\overline{\nu}$ (see Section 3.2) are described in Appendix B.

The merging of historic data proposed above avoids the need for population modelling of the patients and has proved to be a convenient means of initializing the identification of the biphasic model. A formal optimization with respect to $\overline{\nu}$ and ν_0 would require evaluation of the predictive distribution of D as a function of these quantities, but would be unwieldy. We will see in Section 7 that the merging achieved above is satisfactory, in the sense that identification of patient-specific biphasic parameters is greatly enhanced using these values of V_0 and ν_0.

5. The Posterior Inference

The posterior inference of thyroid activity parameters (2) for a specific patient, given prior constraints, $\mathcal{I}_c$, and external information from the patient archive, $\mathcal{I}_0$, is given by

$$\begin{aligned} f(a, r|D, \mathcal{I}_0, \mathcal{I}_c) &\propto f(a, r|\mathcal{I}_c)\ f(a, r|\mathcal{I}_0) \prod_{i=1}^{n} f(x_{t_i}|a, r) \\ &= \prod_{i=1}^{n} \mathcal{N}_{x_{t_i}}(\psi'_{t_i} a, r)\ \mathcal{N}i\mathcal{G}_{a,r}(V_0, \nu_0)\ \chi_{\mathbb{A}}(a) \\ &\propto \mathcal{N}i\mathcal{G}_{a,r}(V_n, \nu_n)\ \chi_{\mathbb{A}}(a). \end{aligned}$$

V_0 and ν_0 are given in Section 4.2, and the posterior statistics, V_n and ν_n, are calculated from these via the conjugate updates in Section 3.2. Recall (Section 1.2) that our aim is to predict patient-specific activity and to infer dose, ξ. These are consistently addressed via the associated marginal in a,

$$f(a|D, \mathcal{I}_0, \mathcal{I}_c) \propto f(a|L_n, \Lambda_n, \nu_n)\ \chi_{\mathbb{A}}(a), \tag{12}$$

where $f(a|L_n, \Lambda_n, \nu_n)$ is given by (7). Now, the normalizing constant is not available in closed form, owing to the domain restriction imposed by $\chi_{\mathbb{A}}(a)$.

The following difficulties emerge:

1. From (2), the patient's posterior mean log-activity curve is given by

$$\mathsf{E}_{f(a|D,\mathcal{I}_0,\mathcal{I}_c)}[\ln A_t] = \psi'_t \hat{a}_c - \alpha t.$$

Here, the expectation is with respect to the constrained distribution (12), whose required moments—such as $\hat{a}_c$ or $\mathsf{cov}_c[a]$ (where subscript 'c' denotes a constrained moment)—are, again, unavailable in closed form, because of the domain restriction, $\chi_{\mathbb{A}}(a)$.

2. The transformed distribution, $f(\xi|D, \mathcal{I}_0, \mathcal{I}_c)$, via the surjective mapping $a \to \xi(a)$ implied by (1) and (2), is unavailable in closed form, since the integral in (1) cannot be evaluated analytically.

These difficulties necessitate an approximation of $f(a|D, \mathcal{I}_0, \mathcal{I}_c)$. We adopt a stochastic sampling technique, as described next. Similar approach to numerical transformation of distributions was used e.g. in [26].

6. Stochastic Sampling from the Posterior Inference

Stochastic samples are drawn—in a manner to be described next—from the transformed posterior density, $f(a^*|D, \mathcal{I}_0, \mathcal{I}_c)$, under the transformation in Section 3.1. The transformed support, $\mathbb{A}^*$ (10), is the solution space of $M^*\ a^* < b^*$, with $M^* = MT^{-1}$ and $b^* = b - M\hat{a}$. Here, $\hat{a}$ is the unconstrained posterior mean (5). As explained in Section 3.1, the unconstrained distribution (Student), $f(a^*|D, \mathcal{I}_0)$, has zero mean and identity covariance matrix, and so the posterior distribution (12) is now completely specified by $\mathbb{A}^*$ and ν_n. This greatly reduces the number of matrix multiplications required when drawing a proposal sample, reducing the run time.

The Langevin diffusion algorithm [27, 28] is well adapted to sampling from a low-dimensional, heavy-tailed distribution such as ours'. The algorithm differs from the Random Walk Metropolis-Hastings (RWMH) sampler via a deterministic shift of the proposed point in the direction of maximal gradient of the sampled distribution. As shown in [27], the Langevin diffusion, when optimally tuned, exhibits an acceptance rate of 57.4 %, which is more than twice that of the RWMH algorithm (23 %), therefore achieving faster convergence.

Each i.i.d. realization of $a^{*(i)}$ is inverse-transformed to $a^{(i)}$ (Section 3.1), and substituted into (2). The equivalent realization from $f(\xi|D, \mathcal{I}_0, \mathcal{I}_c)$ is obtained by numerical evaluation of the integral (1), using the QUANC8 algorithm [29].

6.1. Tuning the Langevin Sampler

When the sampler is tuned appropriately, posterior moments and confidence intervals of ξ can be evaluated for a specific patient in the order of 0.1 second using C++ on a standard PC. Hence, this inference procedure is suitable for use in clinical practice. The salient features of this tuning are now outlined.

6.1.1. Initialization

The chain is initialized at the Maximum *a Posteriori* (MAP) estimate, once again found by constrained optimization of the quadratic denominator (7). Since the hard constraints (10) are linear, quadratic programming is used whenever $\hat{a^*} \notin \mathbb{A}^*$ (10), where $\hat{a^*}$ is the unconstrained transformed posterior mean, equal to zero, as explained above. In practice, $\hat{a^*} \in \mathbb{A}^*$ iff all the elements of b^* are positive.

6.1.2. Step-Size

The step-size of the Markov chain (MC) can be derived analytically in the Langevin diffusion case, if the posterior distribution belongs to the exponential family, if it can be factorized into univariate factors, and if it has unbounded support [27]. However, the posterior (12) does not satisfy any of these requirements.

Instead, the patient archive of 3 876 data sequences (Section 4.2) is used to generate a population of optimal MC step-sizes empirically. For each patient, the criterion of maximum first-order efficiency η [27] is used to search for the optimal step-size:

$$\eta = \frac{1}{N-1} \sum_{i=2}^{N} \left(|x_i - x_{i-1}|^2 \right).$$

Here, N is the number of drawn samples x_i, and $|x - y|$ denotes the Euclidean distance between the points x and y. In the case of the unconstrained posterior distribution (7), the acceptance rate for proposed samples is over 50 % when using the optimum step-size in terms of η. This is in agreement with [27]. The acceptance rate decreases as the mass of $f(a|L_n, \Lambda_n, \nu_n)$ is limited by the prior support, $\chi_{\mathbb{A}}(a)$. It was observed that the acceptance rate is never less than 10 % for any case of $\hat{a}$ (5) and $\mathbb{A}$. The magnitude of the step-size in a^*-space is approximately 1.6 when $M\hat{a} \ll b$. However, if $\hat{a} \notin \mathbb{A}$, then the step-size, optimized in terms of η above, can be as much as 10^6.

6.1.3. Burn-In

The burn-in stage of the MC run is used for finer adjustment of the step-size given by the rule above. After drawing 200 samples, the acceptance rate is estimated. If it is higher than 57 %, the step-size is multiplied by $\sqrt{2}$. If it is lower than 10 %, the step-size is divided by $\sqrt{2}$. The procedure is repeated until the acceptance rate is stabilized between 10 % and 57 %. In the majority of cases, no adjustment is necessary, but no more than two such adjustments are made in any case.

6.1.4. Stopping Rule

Stochastic sampling from $f(a|D, \mathcal{I}_c, \mathcal{I}_0)$ is terminated using the nonparametric Bayesian stopping rule proposed in [30]. The number of i.i.d. samples at stopping satisfies

$$N = \min \left\{ k : \mathsf{KLD} \left[\mathcal{D}_k || \mathcal{D}_{k-1}; \mathbb{P}_k \right] < \epsilon \right\}.$$

Here, $\mathcal{D}_k$ denotes the Dirichlet measure induced by the first k i.i.d samples. $\mathsf{KLD}[\cdot]$ denotes the Kullback-Leibler divergence between consecutive Dirichlet measures on the partition, $\mathbb{P}_k$, of the parameter space, $\mathbb{A}$, using the k i.i.d. samples as vertices. ϵ denotes the maximum permitted divergence at stopping [30].

For $\epsilon = 0.002$, the average value of N is $\bar{N} = 4\,529$, across 700 data sequences in the patient archive. The standard deviation is 540. The histogram of N is illustrated in Figure 2 for this set of 700 patients. For each of the 700 patients,

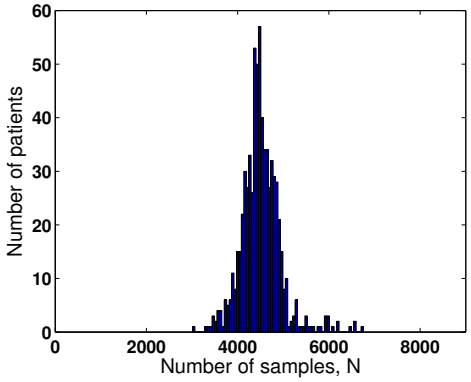

Figure 2: Histogram illustrating variability of the number, N, of i.i.d. samples at stopping, across a population of patients (700 patient cases, $\epsilon = 0.002$).

$i = 1, \ldots, 700$, two empirical distributions of ξ (1) are constructed: (i) $f_{ri}(\xi)$, the reference, using $N = 50\,000$ samples, and (ii) $f_{\epsilon i}(\xi)$, using N_i samples, where N_i satisfies the stopping rule above. The medians, m_{ri} and $m_{\epsilon i}$, were evaluated in each case and the relative error $(m_{\epsilon i} - m_{ri})/m_{ri}$, was calculated, $i = 1, \ldots, 700$. Finally, the mean and the standard deviation of these relative errors was calculated. The same procedure was applied to the lower bound, upper bound and length of the symmetric 95% confidence intervals of $f_{ri}(\xi)$ and $f_{\epsilon i}(\xi)$, $i = 1, \ldots, 700$. None of the means and standard deviations of these relative errors was greater than 0.035. We conclude that the stopping rule yields an accurate approximation of $f(\xi|D, \mathcal{I}_0, \mathcal{I}_c)$.

7. Performance Study: Influence of the Priors

We now consider the influence of the hard parameter constraints, $\mathcal{I}_c$ (Section 4.1), and the external information from the patient archive, $\mathcal{I}_0$ (Section 4.2), on the inference of thyroid activity for a specific patient. Thus, in Figure 3, we plot $f(a_2|D, \mathcal{I}_0, \mathcal{I}_c)$, which is the marginal of the parameter a_2 (2) implied by (12), for the patient whose data are illustrated in Figure 1. Note that $f(a_2|D, \mathcal{I}_0)$ is almost identical to $f(a_2|D, \mathcal{I}_0, \mathcal{I}_c)$, and so it is not shown in Figure 3. However, $f(a_2|D)$ which ignores both forms of prior information, and $f(a_2|D, \mathcal{I}_c)$ which ignores the external information from the patient archive, $\mathcal{I}_0$, are shown in Figure 3. Similar behaviour is observed in the respective marginals for a_3, while a_1 is unconstrained (10).

Note that, for this patient case, $\hat{a} \notin \mathbb{A}$, where $\hat{a}$ is the unconstrained posterior mean (5). This is found to be the case in about 41 % of the patients in the archive (see Table 1). In contrast, the posterior mean of $f(a|D, \mathcal{I}_0)$ is well within $\mathbb{A}$ in this case, as occurs in about 99.5 % of patients (Table 1).

We note the following:

(i) In most patient cases, the hard constraints, via $\chi_{\mathbb{A}}(a)$, have little impact on

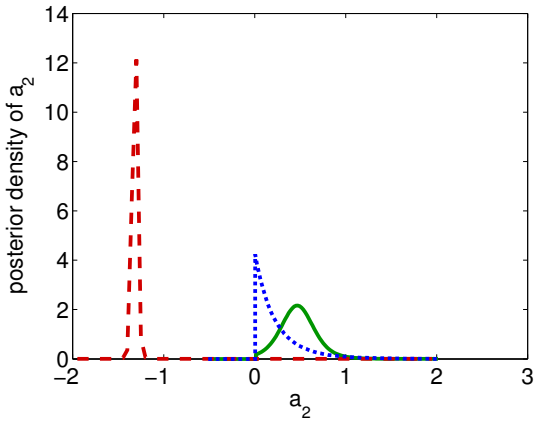

Figure 3: Marginal posterior inference of a_2 for the patient data in Figure 1 (*i.e.* $n = 3$ activity measurements). **Solid line:** the complete regularized inference, $f(a_2|D, \mathcal{I}_0, \mathcal{I}_c)$, from (12). **Dashed line:** unregularized inference, $f(a_2|D)$. **Dotted line:** inference, $f(a_2|D, \mathcal{I}_c)$, constrained via $\chi_{\mathbb{A}}(a)$, but without the data-informed prior, $\mathcal{NiG}(V_0, \nu_0)$ (9). Note that $f(a_2|D, \mathcal{I}_0)$ is almost identical to $f(a_2|D, \mathcal{I}_0, \mathcal{I}_c)$, differing only in respect of the truncation at $a_2 = 0$. It is therefore not illustrated.

the value of the point estimate, $\hat{a}$, once $\mathcal{I}_0$ is taken into account. In this sense, the external information is seen to 'regularize' the inference of a. In conclusion, for most of the patient cases,

$$f(a|D, \mathcal{I}_0, \mathcal{I}_c) \approx f(a|D, \mathcal{I}_0);$$

i.e. a is approximately conditionally independent of $\mathcal{I}_c$ *a posteriori*, given $\mathcal{I}_0$. (ii) Since the distribution of a is heavy-tailed, a relatively diffuse truncated distribution, $f(a|D, \mathcal{I}_c)$, is typically implied in the case when $\mathcal{I}_0$ is ignored (see Figure 3). In the rare cases when $\hat{a} \in \mathbb{A}$ (here, $\hat{a}$ is the mean of the unregularized inference, $f(a_2|D)$ (5)), the optimum step-sizes are between 1 and 2 and the acceptance rates are between 35 % and 50 %. Recall, from Section 6.1.2, that in the frequent cases when $\hat{a} \notin \mathbb{A}$ (*e.g.* Figure 3), the optimum step-sizes can increase to as high as 10^6, and the acceptance rates can drop to as low as 10 %. Hence, the external information from the patient archive, $\mathcal{I}_0$, greatly improves the performance of the Langevin sampler and stabilizes the optimum step-size.

7.1. Statistical Study of Activity Prediction

Next, the influence of the priors on the prediction of measured activity is studied. The same set of 2 355 data sequences as was used in Section Appendix B was used here, each containing at least 4 measurement pairs. For each data sequence (*i.e.* patient case), the log of the measured activity at the 4th measurement time, t_4, is predicted via $\mathsf{E}_{f(d_{t_4}|V_n,\nu_n)}[\ln d_{t_4}]$ (6), given the first $n = 3$ measurements. The following four predictions are generated for each of the 2 355 patients:

(a) Prior knowledge $\mathcal{I}_c$ and $\mathcal{I}_0$ are ignored (*i.e.* V_n and ν_n are initialized via $\overline{V}$ and $\overline{\nu}$ respectively (Section Appendix B)). In this case, about 41 % of the predictions (Table 1) must be rejected, since the inferred mean activity curve (2) is physically impossible (*i.e.* $\hat{a} \notin \mathbb{A}$ in these cases, as discussed in the previous Section). Clearly, this diffuse prior assumption is unacceptable for inference with typical patients.

(b) $\mathcal{I}_c$ is active, but $\mathcal{I}_0$ is ignored (*i.e.* initialization as in (a) above). By definition, all predictions are now accepted.

(c) $\mathcal{I}_0$ is active, but $\mathcal{I}_c$ is ignored (*i.e.* V_0 is constructed via external information from the patient archive, as explained in Section Appendix B, and so V_n and ν_n are initialized as $\overline{V} + V_0$ and $\overline{\nu} + \nu_0$ respectively). In this case, only 0.5 % of the predictions need to be rejected (Table 1) as physically impossible.

(d) Both $\mathcal{I}_c$ and $\mathcal{I}_0$ are active (*i.e.* initialization as in (c) above). Once again, by definition, all predictions are accepted.

For each of the 2 355 patients, the prediction error, *i.e.* $\mathsf{E}_{f(d_{t_4}|V_n,\nu_n)}[\ln d_{t_4}] - \ln d_{t_4}$, is evaluated (where, once again, the argument of $\mathsf{E}[\cdot]$ is to be understood as a random variable, while d_{t_4} is the available fourth measurement in each case (footnote 4)). The mean, median and standard deviation of this quantity across the 2 355 patients are recorded in Table 1 for each of the cases (a)–(d) above. We note a major improvement in activity prediction when both $\mathcal{I}_c$

prior	initialization		posterior	mean	median	st. dev.	% valid
(a)	$\overline{V}$	$\overline{\nu}$	(7)	−0.2333	−0.1464	0.7118	59.0
(b)	$\overline{V}$	$\overline{\nu}$	(12)	−0.1989	−0.1454	0.6553	100.0
(c)	$\overline{V} + V_0$	$\overline{\nu} + \nu_0$	(7)	−0.0008	−0.0340	0.4713	99.5
(d)	$\overline{V} + V_0$	$\overline{\nu} + \nu_0$	(12)	0.0000	−0.0348	0.4727	100.0

Table 1: Statistics of the prediction error in measured log-activity for the four prior knowledge structures, (a)–(d), listed in the text, over a population of 2 355 patients. The "% valid" column gives the percentage of data sequences yielding valid predictions.

(prior constraints) *and* $\mathcal{I}_0$ (extended information) are exploited. For example, the mean and median errors are reduced by a factor greater than 4 compared to the unregularized case (a). Most of this improvement is achieved via $\mathcal{I}_0$ (case (c)) alone, as discussed in Section 7. The modest extra improvement between cases (c) and (d), and the robustness of the predictions in case (d) (see the "% valid" column), recommend the conditioning of patient-specific inferences on both $\mathcal{I}_0$ and $\mathcal{I}_c$ (12).

The prediction study was repeated for prediction of activities, d_t, using the log-normal mean, $\hat{d}_{t_4} = \exp(\psi' \hat{a}_c - \alpha t + \hat{r}_{t_4}/2)$ [24], where $\hat{r}_{t_4}$ is the predictive variance at time t_4. The results were similar to those in Table 1.

7.2. Statistical Study of Predictive Variance

Next, the influence of prior information, $\mathcal{I}_0$ and $\mathcal{I}_c$, on predictive variance, $\hat{r}_t = \hat{r}[1 + \rho_t]$ in (6), was examined. Data containing 3 355 sequences, each with at least 4 activity measurements, were selected, as in Section 7.1. From these, three populations were selected, containing, respectively,

P1: all 2 355 sequences,

P2: 2 344 sequences for which $\chi_{\mathbb{A}}(\hat{a}) = 1$, identified using 3 measurements with external information $\mathcal{I}_0$,

P3: 1 389 sequences for which $\chi_{\mathbb{A}}(\hat{a}) = 1$, identified using 3 measurements without external information $\mathcal{I}_0$.

For all sequences, the unconstrained posterior mean, $\hat{a}$ (5), was used for classification into the populations P2 and P3. For the sequences in P1 in case (iv) below, whenever $\chi_{\mathbb{A}}(\hat{a}) = 0$, then the MAP estimate, $\hat{a}_{\text{MAP}}$, obtained by quadratic programming (see Sections Appendix B and 6.1.1), was used to guarantee a physically meaningful A_t. As noted in the previous section, statistics for the

case	initialization	population	mean	median	std. dev.	min.	max.
(i)	$\overline{V} + V_0, \overline{\nu} + \nu_0$	P1	1.572	1.134	1.282	0.301	9.191
(ii)	$\overline{V} + V_0, \overline{\nu} + \nu_0$	P2	1.574	1.134	1.284	0.301	9.191
(iii)	$\overline{V} + V_0, \overline{\nu} + \nu_0$	P3	1.543	1.140	1.228	0.304	9.191
(iv)	$\overline{V}, \overline{\nu}$	P1	15.277	6.793	97.868	0.999	4 637.600
(v)	$\overline{V}, \overline{\nu}$	P3	12.602	6.773	21.947	1.018	259.620

Table 2: Sampling statistics of $\rho_{t_4} = \psi'_{t_4} L_{aa}^{-1} \Lambda_{aa}^{-1} (L'_{aa})^{-1} \psi_{t_4}$ (6), for the combinations (i)–(v) of prior information and the selected patient populations listed in the text.

cases (i)–(iii) in Table 2, where $\mathcal{I}_0$ is used, do not differ significantly, whereas, in (iv) and (v), the absence of $\mathcal{I}_0$ increases predictive variance greatly, particularly in case (iv), where nearly 1 000 of the estimated activity curves are based on $\hat{a}_{\text{MAP}}$ above, yielding poor prediction. The impact of $\mathcal{I}_0$ on the quality of prediction is particularly evident when comparing cases (iii) and (v), as they involve the same populations of patients.

For the next study, the population P3 was used. For each data sequence in the population, quantities $\rho_{t_4}^+$ (using external information $\mathcal{I}_0$) and $\rho_{t_4}^-$ (without $\mathcal{I}_0$) were evaluated, and combined as shown in Table 3. These results demonstrate that external information $\mathcal{I}_0$ decreases predictive variance (6) approximately fourfold in the mean.

7.3. The Posterior Distribution of ξ

The empirical approximation of $f(\xi|D, \mathcal{I}_0, \mathcal{I}_c)$ computed via the Langevin diffusion-based sampler (Section 6) is illustrated in Figure 4 for the specific patient data shown in Figure 1 ($n = 3$). There is evidence in the literature to

expression	mean	median	std. dev.	min.	max.
$(1+\rho^{-}_{t_4})/(1+\rho^{+}_{t_4})$	4.310	3.617	3.185	1.103	76.707

Table 3: Sampling statistics of the ratio of predictive variance terms (6), excluding and including external information, $\mathcal{I}_0$, and based on the patient population, P3, defined in the text.

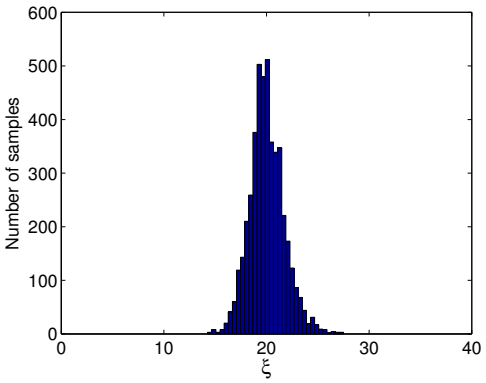

Figure 4: Empirical approximation (with binning) of $f(\xi|D,\mathcal{I}_0,\mathcal{I}_c)$, for the patient data in Figure 1. Computation was via a Langevin diffusion sampler (Section 6) with $\epsilon = 0.002$, giving $N = 4\,600$ at stopping.

support a log-normal distribution of ξ across a patient population. For example, a theoretical thyroid mass distribution was used to support such a claim in [7], and sources of uncertainty were assumed log-normal in [8]. It is therefore of interest to examine the log-normality of our patient-specific dose inference above.

Our investigations concerning log-normality of $f(\xi|D,\mathcal{I}_0,\mathcal{I}_c)$ were partly reported in [31]. The accumulated evidence is now summarized:

(i) Bayesian binary hypothesis testing between a log-normal and normal model for $f(\xi|\cdot)$ was undertaken for many patient cases. This supported the former against the latter, but did not consider other alternatives.

(ii) A Kolmogorov-Smirnov (KS) test of normality was performed on samples from $f(\xi|\cdot)$ and $f(\ln\xi|\cdot)$. The average KS statistic, across a large sample of patients in the database, was too large to support normality of either $f(\xi|\cdot)$ or $f(\ln\xi|\cdot)$. This was probably due to an insufficient number of samples drawn from $f(\xi|\cdot)$.

(iii) For each of 700 patients drawn from the database, a log-normal model was fitted to the empirical approximation of $f(\xi|\cdot)$, generated, as always, via the Langevin diffusion-based sampler (Figure 4). The median, and the lower and upper bounds of the 95 % confidence interval, were calculated

for the empirical approximation, and averaged over the 700 cases. The same was done for the log-normal fit. Pairwise comparison of these three averaged statistics, between the empirical and parametric cases, agreed to within 2 %, providing good support for a log-normal model of ξ.

(iv) Finally, the *skewness* of both the empirical approximations, $f(\xi|D,\mathcal{I}_0,\mathcal{I}_c)$ and $f(\ln\xi|D,\mathcal{I}_0,\mathcal{I}_c)$, were quantified. The rationale is that ξ should exhibit positive skewness if it is, indeed, approximately log-normal, while $\ln\xi$ (which is therefore approximately normal) should have skewness close to zero. These quantities were calculated for each of the 3 876 data sequences in the patient archive, and the statistics of the resulting empirical distributions of skewness were evaluated and compared, as summarized in Table 4. Note that the mean skewness of $f(\xi|D,\mathcal{I}_0,\mathcal{I}_c)$ is more that five

$f(\cdot\|D,\mathcal{I}_0,\mathcal{I}_c)$	**mean**	**median**	**st. dev.**
ξ	1.69	0.85	3.60
$\ln\xi$	0.28	0.23	0.62

Table 4: Statistics for the skewness of the empirical approximations of $f(\xi|D,\mathcal{I}_0,\mathcal{I}_c)$ and $f(\ln\xi|D,\mathcal{I}_0,\mathcal{I}_c)$ across a population of 3 876 patients.

times greater than that of $f(\ln\xi|D,\mathcal{I}_0,\mathcal{I}_c)$ and the latter is quite small. Again, this supports a log-normal model for $f(\xi|\cdot)$. Note also from Table 4 that the mean skewness of $f(\xi|D,\mathcal{I}_0,\mathcal{I}_c)$ is about twice its median skewness, suggesting that this distribution is heavily skewed for many of the patient cases.

This evidence, particularly in (iii) and (iv), supports the adoption of a log-normal model for patient-specific dose, $f(\xi|D,\mathcal{I}_0,\mathcal{I}_c)$. However, further work on formal parametric identification of ξ, via (1), (2) and (7), is warranted.

Finally, for each of the 3 876 data sequences in the patient archive, the standard deviation of $f(\ln\xi|D,\mathcal{I}_c)$ was computed (*i.e.* ignoring the external information from the patient archive, $\mathcal{I}_0$ (Section 4.2)). This was repeated for $f(\ln\xi|D,\mathcal{I}_0,\mathcal{I}_c)$, *i.e.* exploiting the external information. The average standard deviation in the latter case was found to be just 36 % of the former case. This underlines the major impact which the external information from the patient archive has in reducing uncertainty concerning the radiation dose delivered to a specific patient. This has practical significance in the design of a probabilistic dose advisory system based on $f(\xi|D,\mathcal{I}_0,\mathcal{I}_c)$ (see Section 8).

8. Discussion

The inference of biphasic model parameters for an individual patient's thyroid activity in ^{131}I-therapy is a challenging problem since the maximum number of measurements is typically three, while noise from the background and other

sources of uncertainty are typically high. In previous work [15], the biphasic model was shown to yield far better predictions of activity during the clearance phase than is possible for a monoexponential model, in addition to modelling the uptake phase of course. This, in turn, provides improved inference of dose via the integrated activity curve (1). In this paper, we have concentrated on the role of the biphasic model in thyroid ^{131}I-therapy, and have reported an optimized Bayesian framework for inference of its parameters.

8.1. Key findings

The following are the key findings of this work:
(i) The original biphasic model [15, 31] used a time-scale factor of $c = 1$. The optimization of c undertaken in this paper has allowed the expert information on A_t to be fully exploited, as described in Section 4.1. With $c = 1$, the increase of inferred A_t in the initial stage of accumulation (Figure 1) was too slow, especially for lower values of a_3. Modification of c to values higher than proposed in Section Appendix A does not significantly improve the model behaviour.
(ii) The hard constraints, $a \in \mathbb{A}$, on the model parameters, imposed via prior information, $\mathcal{I}_c$, have ensured physically realizable inferences of thyroid activity in ^{131}I-therapy.
(iii) The prior statistics, V_0, constructed by processing external information, $\mathcal{I}_0$, from the patient archive, have ensured excellent prior regularization in the sense that the model parameters are found to be *a posteriori* approximately conditionally independent of $\mathcal{I}_c$, given $\mathcal{I}_0$ (Section 7). Three practical benefits of merging $\mathcal{I}_0$, reported in this paper, have been (a) improved accuracy in the prediction of future measured activities (Section 7.1), (b) significantly increased acceptance rates for proposal samples in the Langevin diffusion sampler (Sections 6.1.2 and 7), and (c) greatly reduced uncertainty in the inference of patient-specific dose, ξ (Section 7.3).
(iv) The nonparametric Bayesian stopping rule (Section 6.1.4) can speed up the computation of the dose (ξ) distribution for a particular patient by up to 20 % compared to the use of a pre-specified sample size (being $\bar{N} + 2\sigma_N$, *i.e.* 4 529+1 080=5 609 samples, while ensuring a specified precision of the confidence interval bounds (Section 6.1.4)).
(v) Reliable probabilistic inference of dose, ξ, for individual patients has been achieved, quantifying its uncertainty. Its approximation by a log-normal distribution has been justified.

8.2. Clinical practice and research potential

This work has the following potentional impact on clinical practice at nuclear medicine clinics:
(i) The irradiated thyroid acts as a source of radiation for the patient's other organs. The Bayesian inference of dose delivered to the thyroid (Section 7.3) may be used directly in the inference of dose delivered by the thyroid to other organs during ^{131}I-therapy, in line with the MIRD methodology [4].
(ii) The prediction of the patient's thyroid activity at the next measurement

time (6) can be used to check for gross measurement or logging errors. A measured activity that diverges significantly from the predicted activity, using an appropriate criterion that has yet to be specified, would generate a warning to the operator.

The reported techniques would provide the means for retrospective studies for the following purposes:
(i) Quantification of *thyroid stunning*: there is empirical evidence that the relative maximum activity of the thyroid is reduced, and the rate of clearance increased (up to threefold), during therapeutic (high) administration of ^{131}I, as compared to the values observed at the preliminary diagnostic administration (Section 1). The accurate Bayesian prediction of activity during the clearance phase, using the biphasic model, is proving to be important in the quantitative study of this thyroid stunning phenomenon.
(ii) An *advisory system* for design of patient-specific optimized administrations of ^{131}I [32, 33]: the quantification of dose, ξ, and particularly its uncertainty, can be used to recommend an optimized administration of ^{131}I for a specific patient. It is hoped that an advisory system of this kind will contribute to the quality of ^{131}I-therapy for the patient and to radiation protection of the environment.

8.3. Possible extensions

The key aim of this work has been to demonstrate the success of the simple 3-parameter biphasic model (2) in prediction of measured activity and dose for individual patients undergoing thyroid ^{131}I-therapy. The numerical benefits of the associated conjugate framework for this linear-Gaussian model have been emphasized (Section 3). The paucity of data available for each patient discourages the introduction of extra parameters (Section 2.1). While these might, indeed, reduce the modelling error, e_t (4), a higher prediction error (6) would be inevitable (*i.e.* the influence of Ockham's razor). Nevertheless, the following three extensions do warrant consideration in the future:
(i) Note the large variability in measured activity across the patient archive (Figure B.5). Also, in Table 1, the standard deviation in the prediction error is relatively large compared to the mean, and variability is also indicated by the significant differences between the mean and median (columns 4 and 5 of the Table). The same is true of the estimates of ξ in Table 4. This points to the heterogeneity of the data in the patient archive. In reality, the response of an individual patient will depend on factors such as age, gender, weight and other patient-specific metabolic variables. There may be an advantage in introducing some of these as covariates in the model for measured activity in the thyroid. Informally, the patient archive might be partitioned into more homogeneous sub-groups, and the inference for an individual patient conditioned on the $\mathcal{I}_0$ calculated from the sub-group to which they belong (Section 4.2). More formally, a mixture of biphasic regression models might be used to analyze the patient archive.
(ii) The biphasic model (2) with nonlinear time-scale factor c can be written as a regression model without time-scaling, but with four linear parameters. Its

identification would yield a patient-specific inference of c, but at the cost of increased model complexity, as noted above.
(iii) Further work on the formal parametric identification of the dose distribution, $f(\xi|D, \mathcal{I}_0, \mathcal{I}_c)$ (Section 7.3), is required, to include testing of other possible skewed distributions on a positive support.

9. Conclusion

The reported inferences of thyroid activity and radiation dose can provide radiologists with important quantitative feedback concerning the impact of ^{131}I-therapy on individual patients in their care. The capacity to predict thyroid activity several days beyond the measurement times is important for model validation, and for quality assurance of the measurement procedure. The estimation of dose, and its uncertainty, at the diagnostic stage is important in inferring the irradiation of the patient's other organs, and in planning the subsequent therapeutic administration of ^{131}I. This paper has shown how a Bayesian conjugate inference framework has been crucial in exploiting external information available *in situ* from a patient archive and from expert opinion. Evidence of improved activity predictions and dose inference for the individual patient has been provided.

10. Acknowledgements

This work was partially supported by grants AV ČR 1ET 1007 50404 and MŠMT ČR 1M0572, and (third author) by grants 10/CE/11855 (Lero) and GA16-09848S. The authors acknowledge the valuable contribution made by Dr. Miroslav Kárný of the Department of Adaptive Systems, Czech Academy of Sciences, to the development of this work.

Appendix A. Hard Parameter Constraints, $\mathcal{I}_c$

Here, the prior limitations on the parameters, a, specified in the constraints 1–4, Section 4.1, are formalised.

Constraint 1: Zero Limits of A_t. It follows directly from (2) that constraint 1 is fulfilled if

$$a_3 < 0 < a_2. \tag{A.1}$$

Constraint 2: Unique Maximizer, t_m, of A_t. The biphasic model (2) of A_t is a continuously differentiable function, $\forall t > 0$. Furthermore, $g_t = \psi_t' a$ has a unique maximizer, $t_{\mathrm{m}g}$, if (A.1) is fulfilled. This is given by the solution of ${g_t}^{(1)} = 0$ (here, $\frac{\mathrm{d}^p g_t}{\mathrm{d}t^p} \equiv {g_t}^{(p)}$). It follows that $A_t = \exp(g_t - \alpha t)$ also has a unique maximizer, t_{m}, satisfying constraint 2 without any further requirements on a. Furthermore, $t_{\mathrm{m}g} > t_{\mathrm{m}}$, since $\alpha > 0$.

Constraint 3: Allowed Interval, (t_l, t_u), for the Maximizer, t_m. Since $t_{\rm m} < t_{\rm u}$, it follows that the first derivative ${A_t}^{(1)} < 0$ for $t \geq t_{\rm u}$.

$$a_2 \;<\; -\,a_3\,(c\,t_{\rm u})^{\frac{2}{3}}\left(\frac{2}{3}\ln(c\,t_{\rm u})+1\right)+\alpha t_{\rm u}.$$

Similarly, for $t \leq t_{\rm l}$,

$$a_2 \;>\; -\,a_3\,(c\,t_{\rm l})^{\frac{2}{3}}\left(\frac{2}{3}\ln(c\,t_{\rm l})+1\right)+\alpha t_{\rm l}. \tag{A.2}$$

(A.2) can be written as $a_2 + k a_3 > q$, where $q > 0$. If $k < 0$, then (A.1) and (A.2) are in contradiction for some values of a_3, in which case $t_{\rm m}$ cannot reach its lower limit, $t_{\rm l}$. To overcome this problem, the time-scale factor, c, in (2), can be chosen to ensure that $k \geq 0$. In particular, $k = 0$ if

$$c = \frac{1}{t_{\rm l}}\;\exp\left(-\frac{3}{2}\right) \equiv 1.3388\ \text{days}^{-1},$$

in which case (A.2) is simply replaced by $a_2 > \alpha t_{\rm l}$, and the upper bound in (A.1) becomes redundant.

Constraint 4: Faster Decrease of A_t than the Physical Decay, for $t > t_h$. ${g_t}^{(1)} < 0$ when $t > t_{{\rm m}g}$, in which case ${A_t}^{(1)} < -\alpha$, as required. Also, $t_{{\rm m}g} > t_{\rm m}$, and so constraint 4 is satisfied by choosing $t_{\rm h} = t_{{\rm m}g}$.

Constraint 1 may be extended to higher-order derivatives of A_t, *i.e.* ${A_t}^{(i)} \rightarrow 0^+$ for $i = 0, 1, \ldots, q$, as $t \rightarrow 0^+$, in order to capture the initial convexity in the accumulation of ^{131}I by the thyroid. The required modification of (A.1) is then $a_2 > q$. Nevertheless, the current choice, $q = 0$, still guarantees behaviour of A_t that is physically reasonable.

Appendix B. External Information, $\mathcal{I}_0$

A review of methods for merging external information in probabilistic inference is provided in [34]. In [35], a general Bayesian theory is elaborated for hierarchical models. In [36], the task is specialized to observation models, $m(\Psi, \Theta)$, belonging to the exponential family, with extended regressor, Ψ, and parameters, Θ. In that approach, $\mathcal{I}_0$ is expressed by (i) an externally supplied distribution, $M(\Psi)$, on Ψ and (ii) a probabilistic weight, w, quantifying the observer's belief in this external information. With these conditions, it was shown that $\mathcal{I}_0$ adapts the inference of Θ, as follows:

$$f(\Theta|D,\mathcal{I}_0) \;\propto\; f(\Theta|D)\exp\left(\nu_0\int M(\Psi)\ln m(\Psi,\Theta)\,{\rm d}\Psi\right),$$

where $\nu_0 = n\left(\frac{w}{1-w}\right)$, and n is the number of observations in the data sequence, D. In the special case of a normal linear regression model for observations (4), $\Theta = (a', r)'$ and the term modulating the posterior above has the

form $\mathcal{N}i\mathcal{G}(V_0, \nu_0)$ [36, 37], with

$$V_0 = \nu_0 \int M(\Psi)\ \Psi\Psi'\,\mathrm{d}\Psi,$$

for any supplied $M(\Psi)$. It remains, therefore, to construct[5] $M(\Psi)$ using the historic data from the patient archive, and to set an appropriate value for ν_0.

Construction of $M(\Psi)$. A scatterplot of measurement pairs, $(t_i^j, \ln d_{t_i}^j)$, from the patient archive is illustrated in Figure B.5, where j indexes the patients in the archive.

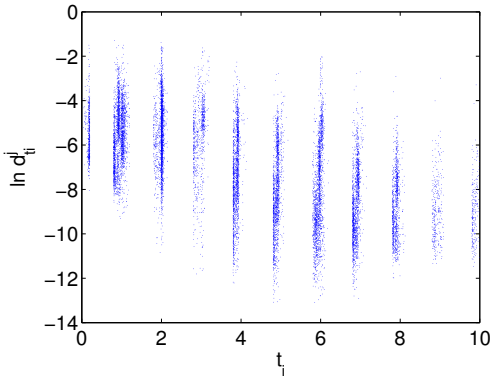

Figure B.5: A scatterplot of measurement pairs, $(t_i^j, \ln d_{t_i}^j)$, from the patient archive.

We note the following:
(i) Measurement times, t_i^j, are strongly clustered around integer times $t \in \{1, 2, \ldots, 10\}$, measured in units of days. This reflects the fact that patients are measured during regular clinic hours on the days immediately following administration of ^{131}I. About 5 % of measurement times *in toto* fell outside the intervals $\pm\Delta t$, $\Delta t = 0.2$ days, around these integer times, and all such measurement pairs, $(t_i^j, d_{t_i}^j)$, were removed (censored). The standard deviation of times in each resulting cluster was then found to be in the range 2–4$\times 10^{-2}$ days.
(ii) The uncensored measured log-activities, $\ln d_{t_i}^j$, in each cluster are assumed to be scattered normally. We evaluated the arithmetic mean, $\langle \ln d_t \rangle_k$, and standard deviation, $\hat{\sigma}_k$, of the $\ln d_{t_i}^j$ in each cluster, $k = 1, \ldots, 10$. From (4), we denote $\hat{x}_k = \langle \ln d_t \rangle_k + \alpha k$. The $\hat{\sigma}_k$ were found to be in the range $(0.8, 1.1)$, *i.e.*

[5] In the case where $M(\Psi) = N^{-1} \sum_{i=1}^{N} \delta(\Psi - \Psi_i)$ (*i.e.* the empirical distribution, where $\delta(\Psi - \Psi_i)$ is the distribution degenerate at Ψ_i), and $\nu_0 = N$ (*i.e.* $w = \frac{N}{n+N}$), then each externally processed regressor, Ψ_i, contributes an unweighted outer-product, $\Psi_i\Psi_i'$, to the posterior extended information matrix, V_n (Section 3.2), in agreement with standard results in nonparametric learning [30].

much larger than the deviations of measured times in each cluster (Figure B.5), as given in (i) above. This observation justifies our neglecting of the uncertainty in the time measurement, *i.e.* we assume $M(t_k) = \delta(t-k)$.
(iii) In the vast majority of patient cases, three measurements were taken in the days following diagnostic administration of ^{131}I. Hence, only the three clusters at $k = 1$, 2 and 10 were chosen, as representative of a typical patient.

From the foregoing, the externally supplied distribution, $M(\cdot)$, which summarizes the historic data from the patient archive, is the following mixture:

$$M(x_t, t) = \frac{1}{3} \sum_{k=1,2,10} \mathcal{N}(\hat{x}_k, \hat{\sigma}_k^2)\, \delta(t-k).$$

Since the mapping (2), (8), is bijective, we can replace Ψ_t by (x_t, t). Substituting $M(x_t, t)$ into the expression for V_0 above, we obtain

$$V_0 = \frac{\nu_0}{3} \sum_{k=1,2,10} \left(\hat{\Psi}_k \hat{\Psi}_k' + \hat{\sigma}_k^2 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \right),$$

where $\hat{\Psi}_k = \left(\hat{x}_k, 1, \ln(ck), (ck)^{2/3} \ln(ck)\right)'$. The method for choosing an appropriate value of ν_0 will be explained in the next Section.

Choice of $\overline{V}$, $\overline{\nu}$ and ν_0. The following constraints must be observed in order that $\mathcal{N}i\mathcal{G}(V, \nu)$, $a \in \mathbb{R}^p$, be proper (*i.e.* that its normalizing constant, ζ [25], exist) and for existence of its key moments (5):

Existence of	**Constraint**
ζ	$\nu > p + 2 = 5$
$\hat{r}$, $\mathsf{cov}[a]$	$\nu > p + 4 = 7$
$\mathsf{var}[r]$	$\nu > p + 6 = 9$

In the ^{131}I-therapy context, the minimal number of measurements is $n = 2$. From Section 3.2, we therefore note that if $\overline{\nu} = 7.05$, then $\nu_n \geq 9.05$ in the posterior distribution, guaranteeing that it is proper with finite moments, even in the absence of any external information, $\mathcal{I}_0$. We choose this conservative value of $\overline{\nu}$ to ensure maximal influence of the data in the posterior inference. This value also ensures that the proposed transformation, T, in Section 3.1, exists. In the absence of other sources of information, beyond $\mathcal{I}_0$, we set $\overline{V} = 10^{-6}\, I_4$, to ensure invertibility (here, I_4 is the 4×4 identity matrix).

Finally, we return to the issue of weighting the external information via ν_0, which corresponds to finding the weighting probability $w = \nu_0/(n + 7.05 + \nu_0)$ (Section 4.2). For this purpose, we select 2 355 normalized data sequences from the archive of 3 876 sequences (Section 4.2), each of which contains at least four measurement pairs. For each sequence, the marginal distribution of a (7),

via $\mathcal{N}i\mathcal{G}(V_3, 10.05 + \nu_0)$, using the first $n = 3$ measurements[6], was maximized over its support, $\mathbb{A}$, by constrained optimization of the quadratic denominator. This estimate, $\hat{a}_{\mathrm{MAP}}$, was used to predict the log of the measured activity, via (6), at the fourth measurement time, t_4, in the sequence, which typically follows after 1–3 days (Figure 1). The error in this predicted quantity, *i.e.* $\psi'_{t_4}\hat{a}_{\mathrm{MAP}} - \alpha t_4 - \ln d_{t_4}$, where d_{t_4} is the available 4th measurement in each case, was averaged over the 2 355 patient cases, and optimized with respect to ν_0. The value $\nu_0 = 0.21995$ was found to minimize this average prediction error and was used as the weighting parameter for the external information, $\mathcal{I}_0$.

References

[1] J. Harbert, W. Eckelman, R. Neumann, Nuclear Medicine. Diagnosis and Therapy, Thieme Medical Publishers, Inc., New York, 1996.

[2] L. T. Morris, R. M. Tuttle, L. Davies, Changing trends in the incidence of thyroid cancer in the United States, JAMA Otolaryngology-Head & Neck Surgery 142 (7) (2016) 709–711.

[3] PDQ Adult Treatment Editorial Board: Thyroid Cancer Treatment (PDQ), PDQ Cancer Information Summaries [Internet] (2016) Health Professional Version.
URL `https://www.ncbi.nlm.nih.gov/books/NBK65719/`

[4] R. Loevinger, T. F. Budinger, E. E. Watson, MIRD Primer for absorbed dose calculations, The Society of Nuclear Medicine, New York, 1988.

[5] H. M. Thierens, M. A. Monsieurs, K. Bacher, Patient dosimetry in radionuclide therapy: The whys and the wherefores, Nuclear Medicine Communications 26 (7) (2005) 593–599.

[6] L. Jirsa, Advanced Bayesian processing of clinical data in nuclear medicine., Ph.D. thesis, FJFI ČVUT, Prague (1999).
URL `http://library.utia.cas.cz/prace/20000056.pdf`

[7] D. M. Hamby, R. R. Benke, Uncertainty of the Iodine-131 ingestion dose conversion factor, Radiation Protection Dosimetry 82 (4) (1999) 245–256.

[8] D. W. Schafer, E. S. Gilbert, Some statistical implications of dose uncertainty in radiation dose-response analyses, Radiation Research 166 (2006) 303–312.

[6] This reflects the usual practice of taking no more than $n = 3$ measurements per patient. In this sense, the extra measurements available for these 2 355 patients may be viewed as test data.

[9] H. Hänscheid, C. Canzi, W. Eschner, G. Flux, M. Luster, L. Strigari, M. Lassmann, EANM Dosimetry Committee Series on Standard Operational Procedures for Pre-Therapeutic Dosimetry II. Dosimetry prior to radioiodine therapy of benign thyroid diseases, European Journal of Nuclear Medicine and Molecular Imaging 40 (7) (2013) 1126–1134.

[10] P. Gebouský, M. Kárný, H. Křížová, M. Wald, Staging of upper limb lymphedema from routine lymphoscintigraphic examinations, Computers in Biology and Medicine 39 (1) (2009) 1–7.

[11] L. Yuh, S. Beal, M. Davidian, F. Harrison, A. Hester, K. Kowalski, E. Vonesh, R. Wolfinger, Population pharmacokinetic/pharmacodynamics methodology and applications: a bibliography, Biometrics 50 (1994) 566–575.

[12] D. M. Harvey, R. P. Hamby, T. S. Palmer, Uncertainty of the thyroid dose conversion factor for inhalation intakes of ^{131}I and its parametric uncertainty, Radiation Protection Dosimetry 118 (3) (2006) 296–306.

[13] D. J. Lunn, N. Best, A. Thomas, J. Wakefield, D. Spiegelhalter, Bayesian analysis of population PK/PD models: General concepts and software, Journal of Pharmacokinetic and Pharmacodynamics 29 (3) (2002) 271–307.

[14] L. D. Marinelli, E. H. Quimby, G. J. Hine, Dosage determination with radioactive isotopes; II. Practical considerations in therapy and protection, American Journal of Roentgenology and Radiotherapy 59 (1948) 260–280.

[15] J. Heřmanská, M. Kárný, J. Zimák, L. Jirsa, M. Šámal, P. Vlček, Improved prediction of therapeutic absorbed doses of radioiodine in the treatment of thyroid carcinoma, Journal of Nuclear Medicine 42 (7) (2001) 1084–1090.

[16] F. Di Martino, A. C. Traino, A. B. Brill, M. G. Stabin, M. Lazzeri, A theoretical model for prescription of the patient-specific therapeutic activity for radioiodine therapy of Graves' disease, Physics in Medicine and Biology 47 (2002) 1493–1499.

[17] K. Weber, U. Wellner, E. Voth, H. Schicha, Influence of stable iodine on the uptake of the thyroid — model versus experiment, Nuklearmedizin 40 (2001) 31–37, in German.

[18] J. P. Bazin, P. Fragu, R. Di Paola, M. Di Paola, M. Tubiana, Early kinetics of thyroid trap in normal human patients and in thyroid diseases, European Journal of Nuclear Medicine 6 (1981) 317–326.

[19] B. K. Shah, Data analysis problems in the area of pharmacokinetics research, Biometrics 32 (1) (1976) 145–157.

[20] C. Hartmanshenn, M. Scherholz, I. P. Androulakis, Physiologically-based pharmacokinetic models: approaches for enabling personalized medicine, Journal of Pharmacokinetics and Pharmacodynamics 43 (5) (2016) 481–504.

[21] T. H. Kim, S. Shin, J. B. Bulitta, Y. S. Youn, S. D. Yoo, B. S. Shin, Development of a physiologically relevant population pharmacokinetic in vitro-in vivo correlation approach for designing extended-release oral dosage formulation, Molecular Pharmaceutics 14 (1) (2017) 53–65.

[22] V. Kliment, J. Thomas, Mathematical solution of the iodine retention and excretion model, Jaderná energie 32 (1988) 85–96.

[23] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, L. Tesař, Optimized Bayesian Dynamic Advising: Theory and Algorithms, Springer, London, 2006.

[24] E. L. Crow, K. Shimizu, Lognormal Distributions: Theory and Applications, Dekker, New York, 1998.

[25] J. M. Bernardo, A. F. M. Smith, Bayesian Theory, John Wiley & Sons, Chichester, 2002.

[26] L. Xiong, C.-K. Chui, Y. Fu, C.-L. Teo, Y. Li, Modeling of human artery tissue with probabilistic approach, Computers in Biology and Medicine 59 (2015) 152–1595.

[27] G. O. Roberts, J. S. Rosenthal, Optimal scaling of discrete approximation to Langevin diffusions, J. R. Statist. Soc. 60, Part 1 (B) (1998) 255–268.

[28] G. O. Roberts, R. L. Tweedie, Exponential convergence of Langevin distributions and their discrete approximations, Bernoulli 2 (4) (1996) 341–363.

[29] G. E. Forsythe, M. A. Malcolm, C. B. Moler, Computer Methods for Mathematical Computations, Prentice Hall, 1977.

[30] A. Quinn, M. Kárný, Learning for Nonstationary Dirichlet Process, International Journal of Adaptive Control and Signal Processing 21 (10) (2007) 827–855.

[31] L. Jirsa, F. Varga, M. Kárný, J. Heřmanská, Model of ^{131}I biokinetics in thyroid gland and its implementation for estimation of absorbed doses, in: Proceedings of the 3rd European Medical and Biological Engineering Conference, IFMBE, Prague, 2005, pp. 1–5.

[32] L. Jirsa, A. Quinn, Mixture analysis of nuclear medicine data: Medical decision support, in: R. Shorten, T. Ward, T. Lysaght (Eds.), Irish Signals and Systems Conference 2001. Proceedings, NUI Maynooth, Maynooth, 2001, pp. 393–398.

[33] A. Quinn, P. Ettler, L. Jirsa, I. Nagy, P. Nedoma, Probabilistic advisory systems for data-intensive applications, International Journal of Adaptive Control and Signal Processing 17 (2) (2003) 133–148.

[34] P. H. Garthwaite, J. B. Kadane, A. Q. O'Hagan, Statistical methods for eliciting probability distributions, Journal of the American Statistical Association 100 (470) (2005) 680–700.

[35] A. Quinn, M. Kárný, T. V. Guy, Fully probabilistic design of hierarchical Bayesian models, Information Sciences 369 (2016) 532–547.

[36] J. Kracík, M. Kárný, Merging of data knowledge in Bayesian estimation, in: J. Filipe, J. A. Cetto, J. L. Ferrier (Eds.), Proceedings of the Second International Conference on Informatics in Control, Automation and Robotics, INSTICC, Barcelona, 2005, pp. 229–232.

[37] M. Kárný, A. Bodini, T. V. Guy, J. Kracík, P. Nedoma, F. Ruggeri, Fully probabilistic knowledge expression and incorporation, Statistics and Its Interface 7 (4) (2014) 503–515.