Akademie věd České republiky
Ústav teorie informace a automatizace, v.v.i.

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

Sean Ernest Murray, Anthony Quinn

Bayesian Selective Transfer Learning for Patient-Specific Inference in Thyroid Radiotherapy

No. 2388 2020

GAČR 18-15970S

ÚTIA AV ČR, P.O.Box 18, 182 08 Prague, Czech Republic
Tel: (+420)266052422, Fax: (+420)286890378, Url: http://www.utia.cas.cz, E-mail: utia@utia.cas.cz

Any opinions and conclusions expressed in this report are those of the authors and do not necessarily represent the views of the Institute.

Contents

1 Abstract

This research report outlines a selective transfer approach for Bayesian estimation of patient-specific levels of radioiodine activity in the thyroid during the treatment of differentiated thyroid carcinoma. The work seeks to address some limitations of previous approaches [4] which involve generic, non-selective transfer of archival data. It is proposed that improvements in patient-specific inferences may be achieved via transferring external population knowledge selectively. This involves matching the patient to a similar sub-population based on available metadata, generating a Gaussian Mixture Model within the partitioned data, and optimally transferring a data predictive distribution from the sub-population to the specific patient. Additionally, a performance evaluation method is proposed and early-stage results presented.

2 Introduction

For patients suffering from differentiated thyroid cancer (DTC), treatment using radioactive iodine (RAI) ^{131}I is considered a standard practice and is used in approximately half of all newly diagnosed cancer cases [8]. As of 2012, incidence rates of people (80% of them female) affected by thyroid cancer in the USA is 1.75% [7], and the high prevalence of thyroid diseases in Asia has prompted continued development [13].

A key pharmacokinetic (PK) quantity of interest during treatment is the time-dependent activity of ^{131}I. This may be used to estimate the net radiation dose delivered to the thyroid, the inference of which is essential in patient prognosis and planning of further treatment [2, 12]. Dosimetry-based personalised treatment has been shown to prevent both sub-optimal administrations, entailing further RAI therapy, and excessive administration of ^{131}I, which increases the potential for radiation toxicity [6]. However, measurements of ^{131}I activity for a specific patient are typically of low quantity and quality, owing to the economics and the nature of the measurement process respectively. A Bayesian approach is thus adopted here as is previously done in [3, 4]. Treatment techniques place high demands on the accuracy of irradiation [8] and dose-prediction models have been shown to produce misleading results if they fail to account for radiation dose uncertainties [11]. Expressing degrees of belief as probabilities in a Bayesian framework provides a numerical measure of the uncertainty attached to each event [1], in addition to enabling the incorporation of externally available knowledge.

In [4], Jirsa et al. introduce a biphasic (uptake-clearance) linear-regression model for ^{131}I activity in a specific (target) patient. It is then shown that transferring externally available knowledge to a patient-specific model is effective in predicting ^{131}I activity. This external knowledge is in the form of archives of patient measurement records, and the same knowledge, namely a data-predictor in the form of a Gaussian Mixture Model (GMM), is transferred indiscriminately to a target patient. This report utilises the same biphasic model, but proposes a more nuanced transfer of external knowledge. This involves using available patient metadata to identify a subpopulation of similar patients within the archive. From which a more tailored GMM data predictor is transferred optimally to the patient at times that complement the target's own data, supplementing its local parameter estimation.

3 Parametric Model for Thyroid Activity Estimation

The following section outlines the parametric model and conjugate update for thyroid activity proposed in [4]. This defines how the target model processes its own measurements for activity estimation.

3.1 Biphasic Model for Thyroid Gland Activity

A model for thyroid gland activity during RAI treatment for DTC is presented by Jirsa et al. in [4]. It is an uptake-clearance (biphasic) log-normal linear regression model for thyroid activity, A_t (MBq), at time t (days).

$$\ln(A_t) = a_1 + a_2 \ln(ct) + a_3 (ct)^{2/3} \ln(ct) - \alpha t \tag{1}$$

$$= \psi_t' a - \alpha t \tag{2}$$

The biphasic model is parameterised by three shaping parameters $a \in \mathbb{R}^3$ and one variance estimate, r. The explanatory variables and constant, c, are grouped in the term $\psi_t \in \mathbb{R}^3$. The parameter-independent term, $-\alpha t$, accounts for the radioactive decay of the ^{131}I isotope.

The patient measured-activity, d_t, is log-normal, and it follows that the Wold observational model of the log-scaled activity measurements, $\ln(d_t)$, at time t, is normally distributed. The target patient's parametric observation model is thus:

$$x_t = \psi_t' a + e_t, \begin{cases} x_t & = \ln(d_t) + \alpha t \\ e_t & \overset{ciid}{\sim} \mathcal{N}(0, r) \end{cases} \tag{3}$$

$$f(x_t|a, r) \propto \mathcal{N}_{x_t}(\psi_t' a, r) \tag{4}$$

3.2 Normal Inverse-Gamma Conjugate Update of Parameters

The adopted conjugate form for estimation of a and r is the standard-form multivariate Normal Inverse-Gamma (NiG) distribution [4], parameterised via two sufficient statistics: the *extended information matrix* (EIM), $\overline{V}_i \in \mathbb{R}^{4\times4}$; and the degree-of-freedom, $\nu_i \in \mathbb{R}^+$.

$$f(a, r|\overline{V}_0, \nu_0) \equiv \mathcal{N}i\mathcal{G}_{a,r}(\overline{V}_0, \nu_0) \in \mathbb{R}^3 \times \mathbb{R}^+ \tag{5}$$

In the vector φ_{t_i}, the i-th observations of shifted log-activities x_{t_i} are stacked on the explanatory variables ψ_{t_i}, known as the *extended datum* (ref). The outer product of the extended datum, $\varphi_{t_i}\varphi_{t_i}'$, provides the prescribed memory-less data projection for inference of the normal linear regression parameters.

$$\varphi_{t_i} = \begin{pmatrix} x_{t_i} \\ \psi_{t_i} \end{pmatrix} \tag{6}$$

$\overline{V}_n$ and ν_n respectively serve as an accumulator and counter of these outer products of extended data, initialised with $\overline{V}_0$, ν_0. The conjugate *batch update* of these parameters is expressed in Equations (7) and (8) and the sequentially-processed *on-line update* is shown in Equations (9) and (10).

$$\overline{V}_n = \overline{V}_0 + \sum_{i=1}^{n} \varphi_{t_i}\varphi_{t_i}' \tag{7}$$

$$\nu_n = \nu_0 + n \tag{8}$$

$$\overline{V}_i = \overline{V}_{i-1} + \varphi_{t_i}\varphi_{t_i}' \tag{9}$$

$$\nu_i = \nu_{i-1} + 1 \tag{10}$$

A diffuse NiG prior is elicited using a small positive constant $\epsilon \approx 0.001$. This is because, for NiG propriety, $V_0 \in \mathbb{R}^{4\times4}$ must be symmetric and positive definite, and $\nu_0 > 0$.

$$\overline{V}_0 = \epsilon \cdot I_4 \tag{11}$$

$$\nu_0 = \epsilon \tag{12}$$

3.3 Physiological Hard Constraints Imposed on a

A number of hard constraints are imposed on the NiG prior in [4], based on known physiological behaviour of ^{131}I in the patient. The hard constraints confine the shaping parameters a to a convex domain $\mathbb{A}$, defined by a matrix of linear inequalities. In [4], the value chosen for the time-scale factor is $c = 1.3388$ days^{-1}, which in turn defines the coefficient matrix M and constant vector b in (14).

$$a \in \mathbb{A} \equiv \{a \mid Ma < b\} \tag{13}$$

$$M = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 4.8687 \\ 0 & -1 & 0 \end{pmatrix}, \; b = \begin{pmatrix} 0 \\ 0.2586 \\ -0.0144 \end{pmatrix} \tag{14}$$

Knowledge of the hard-constraints, $\mathcal{I}_c$, is introduced to the prior via an indicator function $\chi_{\mathbb{A}}(a) \in \{0, 1\}$. The resulting constrained posterior, following $D_n \equiv \{(t_i, d_{t_i})\}_{i=1}^n$ time-measurement pairs taken from a target patient, is

$$f(a, r|\mathcal{I}_c, D_n) \equiv \mathcal{N}i\mathcal{G}_{a,r}(\overline{V}_n, \nu_n)\chi_{\mathbb{A}}(a). \tag{15}$$

4 External Data Classification and Analysis

The data for this report is made available from the Clinic Nuclear Medicine (KNM), Motol Hospital, Prague. Each treatment record consists of a number ($2 < n \leq 9$) of serial time-activity measurement pairs. We denote this data as $D_n \equiv \{t_i, d_{t_i}\}_{i=1}^n$, where i is the discrete time index. As noted in [4], the measurement data across all 3876 treatment records within the KNM dataset is heterogeneous, indicating that transferring knowledge naively from the entire database may neglect potential covariates that would be informative to the target patient. Here we seek to nuance this previously "unselective" transfer.

4.1 Metadata Conditioning of External Data

In addition to the measurement data, D_n, a host of useful metadata is available for each record in the KNM dataset, primarily: a unique patient ID, allowing the same patient to be identified across multiple treatment records[1]; the type of administration given, diagnostic (Dx) or therapeutic (Rx); and the number of lesions the patient has (NL $\in \{1, 5\}$), which may be viewed as the disease-state of the organ. We focus here on the latter two metadata (administration type and NL) affecting patient measurements, D.

Using this available metadata, we associate a target patient with a subpopulation of similar archive records, based on the equivalent administration type (Dx/Rx) and the number of lesions (1-5). This partition scheme instantiates 10 possible *classes* of archive subpopulations to which a target patient may be identified.

4.2 Domains for Modelling and Transfer of External Knowledge

For each treatment record within a class, the parameters of the associated biphasic model may be estimated via the parametric update proposed previously. Given estimates of model parameters may be obtained for all archive records, we propose modelling external class knowledge in the feature-space as this is ultimately the domain of interest, encapsulating all knowledge that is relevant to the learning task. This includes the benefit that this external knowledge may be pre-processed ahead of time.

For transfer to the target, a source (one-step-ahead) data-predictor is derived from the source parameter distribution. This gives access to the same calculus employed in the seminal work for this investigation [5, 9] for FPD-optimal transfer of source knowledge to the target. The distinction in this work is that the data-predictor is derived from *selected* source parameter knowledge, instead of a fixed source data set, and knowledge is transferred to complement the target's local data.

4.3 Representation of Source Parameter Knowledge

In modelling the source knowledge for each class, we propose here to neglect a_1 and transfer a distribution in the $\Theta^* \equiv (a_2, a_3)$ domain only. We propose this as: (i) (a_2, a_3) is the primary domain where non-homogeneity is identified in feature-space; and (ii) a_1 is a scaling term of a patient's biphasic activity model. This scaling term is intimately related to the administered activity, A_0, which differs by patient with a large variance - this is intrinsic to a given patient and thus we argue not worth transferring[2].

For each class, we choose to model the in-homogeneous external parameter space, Θ^*, as a Gaussian Mixture. The number of components, K, is chosen via the Rissanen MDL algorithm [10]. Thus, using available metadata, a target patient may be identified with a Gaussian Mixture Model (GMM) associated with one of 10 classes. Each GMM summarises the available archival knowledge, $\mathcal{I}_S$, within a class. This is represented as a source distribution, f_S, on the parameters of interest, Θ^*:

[1] This property wasn't utilised, but may be an interesting object of future investigation.

[2] Avoiding transfer of a_1 information additionally avoids the need for scaling of transferred data-predictor knowledge, which was encountered previously when data-predictions were not consistent with the target's observations.

$$f_S(\Theta^*|\mathcal{I}_S) = \sum_{k=1}^{K} \hat{\alpha}_k \mathcal{N}_{\Theta^*}(\hat{\mathbf{m}}_k, \hat{\Sigma}_k) \tag{16}$$

Here, $\hat{\alpha}_k$ represents the weight of the k-th component in the GMM. $\hat{\mathbf{m}}_k$ and $\hat{\Sigma}_k$ represent the k-th component mean and covariance matrix respectively, estimated using the Expectation Maximisation (EM) algorithm (easily done via the `fitgmdist` function in MATLAB).

5 Proposed External Parameter Update

The following section presents the proposed external parameter update of a novel selective external data-predictive distribution, within the FPD-optimal transfer framework [4, 9].

5.1 Target One-Step-Ahead Predictor

Given $\Theta \equiv a$, the target's likelihood estimation and one-step-ahead predictor are defined in the standard Bayesian Transfer Learning (BTL) format:

$$\Theta \sim f(\Theta) \tag{17}$$
$$x_i|\Theta \sim f(x_i|\Theta) \equiv L(\Theta|x_i) \tag{18}$$
$$x_i, \Theta \sim f(x_i, \Theta) \tag{19}$$
$$f(\Theta|\overbrace{x_1, \ldots, x_n}^{\mathbf{x_n}}) \propto f(\Theta) \cdot \prod_{i=1}^{n} L(\Theta|x_i) \tag{20}$$
$$f(x_{n+1}|\mathbf{x_n}) \propto \int f(x_{n+1}|\Theta) f(\Theta|\mathbf{x_n})\, d\Theta \tag{21}$$

5.2 Source Predictor at Time t

For the one-step-ahead *source predictor*, we introduce the source parameter knowledge via the assertion $f(\Theta^*|\mathbf{x_n}) \equiv f_S(\Theta^*|\mathcal{I}_S)$ and marginalising over $\Theta^* \equiv (a_2, a_3)$:

$$f_S(x_{n+1}|\mathbf{x_n}) \propto \int f(x_{n+1}|\Theta^*) f(\Theta^*|\mathbf{x_n}) d\Theta^* \tag{22}$$
$$\propto \int L(\Theta^*|x_{n+1}) f_S(\Theta^*|\mathcal{I}_S)\, d\Theta^* \tag{23}$$
$$= \int \mathcal{N}_{x_{n+1}}(\psi'\mathbf{a}, r) \sum_{k=1}^{K} \hat{\alpha}_k \mathcal{N}_{(a_2,a_3)}(\hat{\mathbf{m}}_k, \hat{\Sigma}_k)\, d\Theta^* \tag{24}$$
$$= \sum_{k=1}^{K} \hat{\alpha}_k \int \mathcal{N}_{x_{n+1}}(\psi'\mathbf{a}, r) \mathcal{N}_{(a_2,a_3)}(\hat{\mathbf{m}}_k, \hat{\Sigma}_k)\, d\Theta^* \tag{25}$$

If we denote $x_{n+1} = \psi'\mathbf{a} + \epsilon$, where $\epsilon \sim \mathcal{N}_{x_{n+1}}(0, r)$, we may express the integral as:

$$\int \big(\psi'\mathbf{a} + \mathcal{N}_{x_{n+1}}(0, r)\big) \mathcal{N}_{(a_2,a_3)}(\hat{\mathbf{m}}_k, \hat{\Sigma}_k)\, d\Theta^* \tag{26}$$
$$= \int \psi'\mathbf{a} \cdot \mathcal{N}_{(a_2,a_3)}(\hat{\mathbf{m}}_k, \hat{\Sigma}_k)\, d\Theta^* + \mathcal{N}_{x_{n+1}}(0, r) \tag{27}$$
$$= \psi_1 a_1 + \int (\psi_2, \psi_3)'(a_2, a_3) \cdot \mathcal{N}_{(a_2,a_3)}(\hat{\mathbf{m}}_k, \hat{\Sigma}_k)\, d\Theta^* + \mathcal{N}_{x_{n+1}}(0, r) \tag{28}$$
$$= \psi_1 a_1 + (\psi_2, \psi_3)'\hat{\mathbf{m}}_k + \mathcal{N}_{x_{n+1}}(0, r) \tag{29}$$

Noting that for the component weights, $\sum_{k=1}^{K} \hat{\alpha}_k = 1$:

$$f_S(x_{n+1}|\mathbf{x}_n) \propto \sum_{k=1}^{K} \hat{\alpha}_k \left(\psi_1 a_1 + (\psi_2, \psi_3)' \hat{\mathbf{m}}_k + \mathcal{N}_{x_{n+1}}(0, r)\right) \tag{30}$$

$$= \psi_1 a_1 + \sum_{k=1}^{K} \hat{\alpha}_k (\psi_2, \psi_3)' \hat{\mathbf{m}}_k + \mathcal{N}_{x_{n+1}}(0, r) \tag{31}$$

We note that the $(n+1)$ measurement is a discrete, time-dependent index. In formulating externally-driven, "fictitious" predictions/interpolations, any positive time-value may be chosen by the modeller. Thus this discrete index is disregarded, being replaced with t. Additionally, we note that $\psi_1 = 1$ and (ψ_2, ψ_3) are functions of t, therefore we denote $\psi_t^* = (\psi_2, \psi_3)$. Thus, we define the source one-step-ahead predictor at a given time t as:

$$f_S(x_t|\mathbf{x}_n) \propto a_1 + \sum_{k=1}^{K} \hat{\alpha}_k \psi_t^* \hat{\mathbf{m}}_k + \mathcal{N}_{x_t}(0, r) \tag{32}$$

$$= \mathcal{N}_{x_t}(m_t^\dagger, r) \tag{33}$$

where,

$$m_t^\dagger = a_1 + \sum_{k=1}^{K} \hat{\alpha}_k \psi_t^* \hat{\mathbf{m}}_k \tag{34}$$

We note here that the quantity a_1 is left to be chosen by the modeller. Here we propose to use the target's own estimate of a_1, estimated before transfer. In Section 5.2.1 we propose an updated variance $\hat{r}_{t_c}$ based on the estimated covariances $\hat{\Sigma}_k$ of the components in the GMM, which are neglected in above calculations.

5.2.1 Definition of r

If the model

$$\int (\psi^{*'} \mathbf{a} + \mathcal{N}_{x_{n+1}}(0, r)) \mathcal{N}_{(a_2, a_3)}(\hat{\mathbf{m}}, \hat{\Sigma}) \, d\Theta^* \tag{35}$$

has expected value $\mathbb{E}[x_{n+1}] = \psi^{*'} \hat{\mathbf{m}}$ and variance $\text{VAR}[x_{n+1}] = \psi^{*'} \hat{\Sigma} \psi^* + r$, we define the variance of the GMM estimate $\hat{r}_{t_c}$ in terms of the variance of each component comprising of n_k data as $\hat{r}_{t_c,k}$. We additionally assume that n_k is large compared to the r.

$$\hat{r}_{t_c} = \sum_{k=1}^{K} \hat{\alpha} \, \hat{r}_{t_c,k} \tag{36}$$

$$= \sum_{k=1}^{K} \hat{\alpha} \left(\psi_{t_c}^{*'} \hat{\Sigma}_k \psi_{t_c}^* + \frac{r}{n_k}\right) \tag{37}$$

$$\approx \sum_{k=1}^{K} \hat{\alpha} \left(\psi_{t_c}^{*'} \hat{\Sigma}_k \psi_{t_c}^*\right) \tag{38}$$

As this term considers the covariance terms within the components of the GMM propose this quantity for use as the variance term in the source data predictor (48), instead of the biphasic model variance estimate r.

5.3 Complementary Prediction & FPD-Optimal Transfer

As it is useful to our application, the one-step-ahead source predictor need not predict future measurements only, but may also be used for interpolation between the target's observed measurements. We may thus choose to transfer knowledge at times where the target's data is scarce, denoted as *complementary times*. This avoids adding external information at times where the target has measurements. If we

choose to transfer information at target-complementary times, $t_c = t_1, \ldots, t_C$, we define for transfer a *complementary* source data-predictor distribution as

$$f_C(x_t, t|\mathbf{x}_n) = \frac{1}{C}\sum_{c=1}^{C} f_S(x_t|\mathbf{x}_n)\delta(t - t_c) \tag{39}$$

$$= \frac{1}{C}\sum_{c=1}^{C} \mathcal{N}_{x_t}(m_t^\dagger, r)\delta(t - t_c). \tag{40}$$

Now that the external data-predictive distribution has been derived, this may be substituted into the mean-field expression for the target's externally updated posterior [4, 9]. This is FPD-optimal transfer.

$$f(\Theta|D_n, \mathcal{I}_S) \propto f(\Theta|D_n) \exp\left[\nu_0 \int f_C(\varphi_t) \ln\left(f(\varphi_t|\Theta)\right)\, d\varphi_t\right] \tag{41}$$

$$\propto f(\Theta|D_n) \exp\left[\frac{\nu_0}{C}\sum_{c=1}^{C} \mathcal{N}_{x_t}(m_{t_c}^\dagger, r) \ln\left(f(x_{t_c}|\Theta)\right)\right] \tag{42}$$

$$= f(\Theta|D_n) \prod_{c=1}^{C} \exp\left[\frac{\nu_0}{C}\mathcal{N}_{x_t}(m_{t_c}^\dagger, r) \ln\left(f(x_{t_c}|\Theta)\right)\right] \tag{43}$$

$$= f(\Theta|D_n) \prod_{c=1}^{C} f(x_{t_c}|\Theta)^{\frac{\nu_0}{C}\mathcal{N}_{x_t}(m_{t_c}^\dagger, r)} \tag{44}$$

This is the resulting posterior for the source parameter estimate. The following section outlines the conjugate update for the NiG parameters .

5.4 NiG Parameter Update by External Predictor

Previous work [4] involved transferring a fixed, non-selective, GMM data-predictor, where each component p is obtained from the mean and standard deviation of archive activity measurements within ± 0.2 days of of the corresponding integer days $p = 1, 2, 10$. The associated times of each measurement around integer day p are quantised ("censored") to equal p, thus giving the following model.

$$M(x_t, t) = \frac{1}{3}\sum_{p=1,2,10} \mathcal{N}(\hat{x}_p, \hat{\sigma}_p^2)\delta(t - p) \tag{45}$$

Given the expression (47) for updating the NiG parameter V_0 using an external data-predictive distribution, the resulting update employed by Jirsa et al. [4] is:

$$V_0 = \nu_0 \int M(\varphi_t)\varphi_t'\varphi_t\, d\varphi_t \tag{46}$$

$$= \frac{\nu_0}{3}\sum_{p=1,2,10} \hat{\varphi}_p'\hat{\varphi}_p + \hat{\sigma}_p^2 \begin{bmatrix}1\\0\\0\\0\end{bmatrix}\begin{bmatrix}1 & 0 & 0 & 0\end{bmatrix} \tag{47}$$

In this paper, we take the same approach of processing an external predictor using the NiG term V_0 but for the newly derived, complementary, selective source data-predictor:

$$f_C(x_t, t|\mathbf{x}_n) \propto \sum_{c=1}^{C} \mathcal{N}_{x_t}(m_t^\dagger, r)\delta(t - t_c) \tag{48}$$

Defining the source extended datum as,

$$\varphi_t^\dagger = \begin{bmatrix} m_t^\dagger \\ \psi_t \end{bmatrix} \tag{49}$$

we define the one-shot transfer as an update of the NiG parameter V_n, by addition of V_C, with normalising constant C:

$$V_C = \frac{\nu_0}{C} \sum_{c=1}^{C} \varphi_{t_c}^{\dagger} \varphi_{t_c}^{\dagger\prime} + r \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \tag{50}$$

In [4], ν_0 is data-driven, the optimum found to be $\nu_0 = 0.21995$. Thus, in this work ν_0 is left as a hyperparameter for testing.

6 Framework for Performance Evaluation

For the benefit of investigation, many of the patient records within the archive supplied by the KNM are enriched with more data points than is typically obtained during RAI treatment. In real practice, generally patients receive ($n = 3$) measurement-pairs following administration. Therefore, additional measurements that are available from the data archive (some patients have up to $n = 9$) can be redacted from the observation model and used for testing instead. Such measurement pairs will be referred to as 'hold-outs' (H/Os).

6.1 Performance Metric

For comparability between transfer methods, the performance metric adopted is the mean prediction error (MPE) of hold-out observations. The general approach adopted here and by Jirsa et al. [4] considers only patients with $n > 3$ measurement pairs. For a given target, the first three measurement pairs are used to inform the NLR model. Following this, the log of the measured activity of the fourth measurement time t_4 is taken as an extrapolation (or expected value) of the estimated log-activity at t_4. Given the first $n = 3$ measurements, this error, ε, is identified in [4] as:

$$\varepsilon = \mathbb{E}_{f(d_4|\overline{V}_n,\nu_n)} \left[\ln(d_{t_4})\right] - \ln(d_{t_4}) \tag{51}$$

In comparison to [4], a slightly modified performance metric is adopted here for the following reasons:

- The performance evaluation in [4] is shown to work well. However, quite typically, more than one additional hold-out is available for a given target record. As measurement-pairs are serial, the later hold-outs that are neglected in this approach typically lie on the right-skirt of the activity curve. The above metric therefore frequently ignores the performance of the model in predicting the later clearance stages of activity.
- The experimental setup should emulate real clinical practice. For patients with large numbers of measurements taken within the study, the first three measurements can be very near one another temporally, and do not reflect the distribution of measurements over the treatment period that would be taken in real practice. Adequate sampling over a sufficient period is required for accurate model fitting [2], particularly sampling on both sides of t_m we argue in this case.

As the performance of the transfer method is a function of which hold-outs are taken during testing, the aim is to show that this dependence is insignificant or unsystematic. Therefore, the adopted approach is to partition between measurements (typical of clinical practice) and H/O (research) data. This is done heuristically. As the global maximum activity must be in the range $t_m \in (4, 72)$ hours and the model captures both uptake and clearance of ^{131}I, the observed measurements are thus chosen (where available) as: (i) the first measurement pair; (ii) the first measurement pair in the window of 2-6 days; and (iii) the first measurement pair taken at greater than 6 days. For the third measurement, while the inclination may be to maximise the range over which a patient is measured, the issue of additive noise becomes more significant measurement times further away from t_m. It's noted that all H/Os contain additive noise. The mean predictive error is taken across all available H/Os, such that the predictive performance of the model is evaluated over the entire activity curve.

6.2 Testing Hyperparameters

One of the hyperparameters used for testing is the size of the window, W (days), over which transfer is permitted to occur. Here, the range of transfer windows tested is $W \in [4, 10]$ days. Note that in previous work in [4], statistics are transferred at fixed days $k = 1, 2, 10$. Additionally testing is run for a range of values of ν_0.

6.3 Error Evaluation

1. Evaluate the parameter update of the NLG model, as prescribed by the NiG distribution, for $i = 1, \ldots, m$ observations;

2. For each hold-out $j = 1, \ldots, h$ find the error between the H/O observation and that of the expected value of the estimated log-activity, from m real observations (as done for a single observation in [4]):
$$\varepsilon_j = \mathbb{E}_{f(d_t|\overline{V}_n, \nu_n)} \left[\ln(d_{t_j})\right] - \ln(d_{t_j}) \tag{52}$$

3. For the MPE, take the average of Euclidean norm error per sample j, between the extrapolated activity model and the h hold-out points:
$$MPE = \sqrt{\frac{\sum_{j=1}^{h} \varepsilon_j^2}{h}} = \frac{|\varepsilon|}{\sqrt{h}} \tag{53}$$

6.4 Test Cases

1. Fully selective transfer: This is the proposed approach, involving a belief-weighted complementary merging of external statistics, obtained from a selected subpopulation GMM source;

2. Control case: No transfer is performed and performances are taken based on estimates from the target's local parametric model.

A topic of future work is to include a robust comparison with the method proposed in [4]. Proven advancements on this method will be necessary for the publication of this work.

7 Results and Conclusion

This report outlined a novel model for selective, complementary update of a patient-specific model for thyroid activity. The algorithm proposed for updating the target's model is based on the work performed in [4], with some alterations to accommodate the new selective framework. Additionally, a more nuanced performance metric is proposed, in order to achieve estimates from data that might be more typically obtained in clinical practice.

Quantitative results for the proposed algorithm are currently limited. From Figure 1 it is evident that the average MPE is minimised using the proposed transfer approach for $\nu_0 = 0.01$ and a transfer window size of $W = 8$. While this is an early report of results, it is nevertheless a positive test case for selective transfer as an improved method for inference of ^{131}I activity. This indicates that further investigation and testing of the proposed selective method may lead to concrete evidence of improvement on methods used in [4].

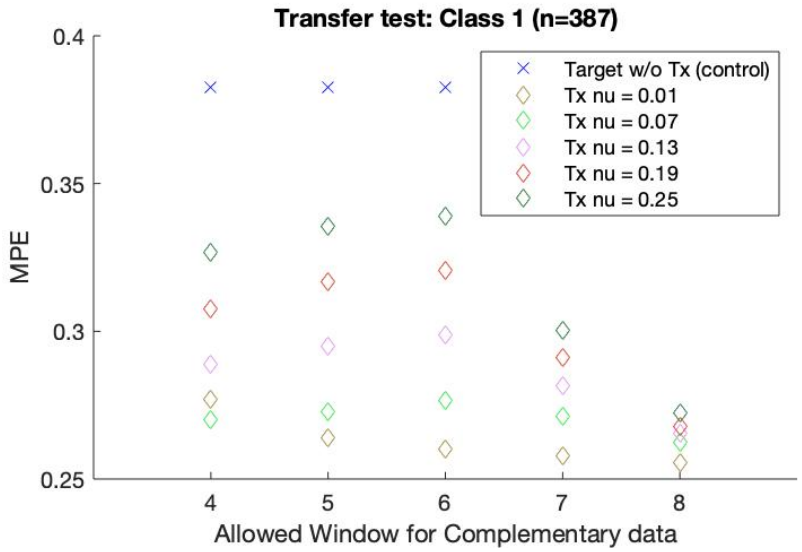

Figure 1: Initial results of MPE for Class 1 data for a range of transfer weight nu_0 values and allowed window for transfer W

References

[1] José M. Bernardo and Adrian F.M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley Blackwell, Hoboken, NJ, USA, May 2008.

[2] J Hermanská, M Kárný, J Zimák, L Jirsa, M Sámal, and P Vlcek. Improved prediction of therapeutic absorbed doses of radioiodine in the treatment of thyroid carcinoma. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, 42(7):1084–90, Jul 2001.

[3] Ladislav Jirsa. *Advanced Bayesian Processing of Clinical Data in Nuclear Medicine*. PhD thesis, FJFI ČVUT, Prague, 1999.

[4] Ladislav Jirsa, Ferdinand Varga, and Anthony Quinn. Identification of thyroid gland activity in radioiodine therapy. *Informatics in Medicine Unlocked*, 7:23–33, Jan 2017.

[5] Jan Kracík and Miroslav Kárný. Merging of data knowledge in bayesian estimation. In *Proceedings of the Second International Conference on Informatics in Control, Automation and Robotics*, pages 229–232. Barcelona, 2005.

[6] Mazzaglia, Stella, Tonghi, Tuvé, Politi, Pellegriti, and Gueli. Absorbed dose evaluation in radioiodine therapy with different approaches. *Instruments*, 3(3):39, Aug 2019.

[7] Luc GT Morris, R Michael Tuttle, and Louise Davies. Changing trends in the incidence of thyroid cancer in the united states. *JAMA Otolaryngology–Head & Neck Surgery*, 142(7):709–711, 2016.

[8] Radiological Protection. ICRP publication 103. *Ann ICRP*, 37(2.4):2, 2007.

[9] Anthony Quinn, Miroslav Kárný, and Tatiana V. Guy. Fully probabilistic design of hierarchical bayesian models. *Information Sciences*, 369:532–547, Nov 2016.

[10] J. Rissanen. Hypothesis selection and testing by the MDL principle. *The Computer Journal*, 42(4):260–269, 1999.

[11] Daniel W. Schafer and Ethel S. Gilbert. Some statistical implications of dose uncertainty in radiation dose–response analyses. *Radiation Research*, 166(1):303–312, Jul 2006.

[12] F. Sun, G. E. Gerrard, J. K. Roberts, T. Telford, S. Namini, M. Waller, G. Flux, and V. M. Gill. Ten year experience of radioiodine dosimetry: is it useful in the management of metastatic differentiated thyroid cancer? *Clinical Oncology*, 29(5):310–315, May 2017.

[13] Chai Hong Yeong, Mu hua Cheng, and Kwan Hoong Ng. Therapeutic radionuclides in nuclear medicine: Current and future prospects. *Journal of Zhejiang University: Science B*, 15(10):845–863, Oct 2014.