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

Sean Ernest Murray[a], Anthony Quinn[b,c]

[a]*Mathematical Institute*
University of Oxford
Oxford, OX2 6GG, UK
sean.murray@maths.ox.ac.uk

[b]*Department of EE Engineering*
Trinity College Dublin, the University of Dublin
Dublin, Ireland
aquinn@tcd.ie

[c]*UTIA*
Czech Academy of Sciences
Prague, Czech Republic
aquinn@utia.cas.cz

2021 32nd Irish Signals and Systems Conference (ISSC) | 978-1-6654-3429-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISSC52156.2021.9467862

***Abstract*—This paper 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 addresses some limitations of previous approaches which involved 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 and formally transferring a feature-space-conditioned, probabilistic data predictor from the sub-population to the specific patient. In addition, the transfer times are chosen to complement the patient's own data. Currently the proposed method yields positive transfer, with stable performance improvements up to 34%. Although this is found to be 9% below the performance of the current state-of-the-art, the proposed method is significant in that it can be applied to other transfer learning applications where inhomogeneous parameter knowledge is available in the source feature space.**

***Index Terms*—Bioinformatics, Decision support systems, Nuclear medicine, Bayesian Transfer learning**

I. Introduction

When a patient undergoes ^{131}I radioiodine (RAI) therapy for treatment of differentiated thyroid cancer (DTC), a key pharmacokinetic quantity of interest is the time-dependent activity of ^{131}I [1]. 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]. However, measurements of ^{131}I activity for a specific patient are typically of low quantity and quality, due to the economics and the nature of the measurement process respectively. A Bayesian approach is thus adopted, here and previously in [1], [3], due to the nature of the patient data and to enable incorporation of externally available knowledge.

In [1], Jirsa et al. introduce a biphasic (uptake-clearance) linear-regression model for ^{131}I activity in a specific patient, specified as the *target*. 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 given patient. This paper utilises the same biphasic model, but proposes a more nuanced transfer of external knowledge. This involves using available patient metadata to identify a sub-population of similar patients within the archive. From this an associated GMM in the sub-population *feature space* (i.e. the estimated regression parameters) is optimally processed as a GMM data-predictor and transferred to the target, supplementing its local parameter estimation. This is done based on the notion of complementary knowledge, in which knowledge is transferred to the target in regions where the target's data is sparse.

As in [1], the transfer is performed optimally via fully probabilistic design (FPD) [4], which outlines axioms on how we process the source knowledge while transferring to the target via a mean-field approach.

The layout of the paper as follows: in Section II we summarise the log-normal linear regression model for thyroid activity estimation first proposed in [1]. In Section III we propose to model the source knowledge in the feature space, conditioned on available metadata, and processed to complement the observed target data. Section IV outlines the optimal Bayesian transfer technique for processing this source knowledge, which is expressed as a parameter update with the virtue of not disrupting the recursive activity estimation. A performance evaluation is carried out in Section V.

II. Parametric Model for Thyroid Activity Estimation

A. Biphasic Model for Thyroid Gland Activity

A model for thyroid gland activity during RAI treatment for DTC is presented by Jirsa et al. in [1]. It is an uptake-clearance (biphasic) log-normal linear regression model for thyroid activity, A_t (MBq), at time t (days), given by

$$\begin{aligned}\ln(A_t) &= a_1 + a_2 \ln(ct) + a_3 (ct)^{2/3} \ln(ct) - \alpha t \\ &= \psi_t' a - \alpha t. \end{aligned} \quad (1)$$

The biphasic model is parameterised by three shaping parameters $a \in \mathbb{R}^3$ and one variance estimate, r, where $'$ denotes

The research has been supported by GAČR grant 18-15970S.

978-1-6654-3429-4/21/$31.00 ©2021 IEEE

transposition. 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 observation model of the log-scaled activity measurements, $\ln(d_t)$, at time t, is normally distributed [1]. Given the observations expressed as $x_t = \ln(d_t) + \alpha t$, from [1], a target patient's parametric observation model is as follows:

$$x_t = \psi_t' a + e_t, \quad e_t \overset{ciid}{\sim} \mathcal{N}(0, r), \tag{2}$$

$$f(x_t|a, r) \propto \mathcal{N}_{x_t}(\psi_t' a, r). \tag{3}$$

B. Normal-Inverse-Gamma Conjugate Update of a and r

The adopted conjugate form for estimation of a and r (4) is the standard-form multivariate normal-inverse-gamma (NiG) distribution [1], parameterised by the extended information matrix (EIM), $V_i \in \mathbb{R}^{4\times4}$, and the degree-of-freedom, $\nu_i \in \mathbb{R}^+$. These parameters respectively serve as an accumulator and counter of the outer products of the extended data (5), initialised with V_0, ν_0, which specify prior belief.

$$f(a, r|V_0, \nu_0) \equiv \mathcal{N}i\mathcal{G}_{a,r}(V_0, \nu_0) \in \mathbb{R}^3 \times \mathbb{R}^+ \tag{4}$$

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}, denoted as the *extended datum*. 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.

$$\varphi_{t_i} = \begin{pmatrix} x_{t_i} \\ \psi_{t_i} \end{pmatrix} \tag{5}$$

The conjugate *batch* update of these parameters is expressed in Equations (6) and (7) and the sequentially-processed *on-line* update is shown in Equations (8) and (9).

$$V_n = V_0 + \sum_{i=1}^{n} \varphi_{t_i}\varphi_{t_i}' \tag{6}$$

$$\nu_n = \nu_0 + n \tag{7}$$

$$V_i = V_{i-1} + \varphi_{t_i}\varphi_{t_i}' \tag{8}$$

$$\nu_i = \nu_{i-1} + 1 \tag{9}$$

A diffuse NiG prior is elicited using a small positive constant $\epsilon \approx 0.001$. This is because, for NiG propriety, $V \in \mathbb{R}^{4\times4}$ must be symmetric and positive definite, and $\nu > 9$. Following [1], as the minimum number of measurements of a patient within the database is $n = 2$, we therefore adopt the prior parameters

$$V_0 = \epsilon \cdot I_4, \tag{10}$$

$$\nu_0 = 7.05. \tag{11}$$

For the NiG posterior, the marginals of a and r are distributed as Student-t and inverse-gamma respectively, with respective first moments [5] given by

$$\mathbb{E}[a] = V_{xx}^{-1} v_{x1}, \qquad \mathbb{E}[r] = \frac{\lambda}{\nu_n - 7}, \tag{12}$$

where

$$V_n = \begin{pmatrix} v_{11} & v_{x1}' \\ v_{x1} & V_{xx} \end{pmatrix}, \quad \lambda = v_{11} - v_{x1}' V_{xx}^{-1} v_{x1}. \tag{13}$$

C. Physiological Hard Constraints Imposed on a

To encode the known metabolic behaviour of ^{131}I in the body, a number of hard constraints are imposed on the inference of a. The hard constraints confine the shaping parameters a to a convex domain $\mathbb{A}$, defined by a matrix of linear inequalities in [1]. Knowledge of the hard-constraints, $\mathcal{I}_H$, 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-activity measurement pairs taken from a target patient, is

$$f(a, r|\mathcal{I}_H, D_n) \equiv \mathcal{N}i\mathcal{G}_{a,r}(V_n, \nu_n)\chi_{\mathbb{A}}(a). \tag{14}$$

The marginal first moments (12) are therefore unavailable in closed form. In previous work, these values were estimated via stochastic sampling methods, whereas here we adopt a grid-based deterministic scheme for estimation.

III. External Data Classification and Analysis

The data for this research 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 [1], 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. In this paper, we seek to nuance this previously "unselective" transfer.

A. Metadata Conditioning of External Data

Using this available metadata, we associate a target patient with a sub-population of similar archive records, based on the equivalent administration type (diagnostic or therapeutic) and the number of lesions (1-5). This partition scheme instantiates 10 possible *classes* of archive sub-populations to which a target patient may be identified.

B. Modelling Domains 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 that estimates of model parameters may be obtained for all archive records, we propose modelling external class knowledge in the feature-space because 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.

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 inhomogeneity is identified in

the feature-space; and (ii) a_1 is a scaling term of a patient's biphasic activity model. This scaling term is dominantly influenced by the administered activity, A_0, which differs by patient with a large variance: it is intrinsic to a given patient and thus we argue that it is not appropriate to transfer it to the new patient.

In summary - for each class - each archived patient is represented by a length-2 feature vector in Θ^*-space, being the Bayesian (a_2, a_3)-estimate (12). The inhomogeneity of the distribution of these features is modelled in this paper via the GMM universal model. The number of components, K, is chosen via the Rissanen MDL algorithm [6]. Thus, using available metadata, a target patient may be identified with a GMM associated with one of 10 classes. Each GMM summarises the available archival knowledge, $\mathcal{I}_S$, within a particular class. It is represented as a source pdf, f_S, on the parameters of interest, Θ^*:

$$f_S(\Theta^*|\mathcal{I}_S) = \sum_{k=1}^{K} f_S(\Theta^*|L=k,\mathcal{I}_S)\Pr[L=k|\mathcal{I}_s] \quad (15)$$

$$= \sum_{k=1}^{K} \hat{\alpha}_k \mathcal{N}_{\Theta^*}(\hat{\mathbf{m}}_k, \hat{\Sigma}_k). \quad (16)$$

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[1] in MATLAB) [7]. Note that indexing into each of the 10 possible class GMMs is suppressed, for convenience, in (15), (16), and in the sequel. The augmented form (15) expresses $\hat{\alpha}_k$ as a probability mass function with discrete label variable $L = k$. The source GMM conditioned on the label $L = k$ is simply the k-th bivariate Gaussian component.

IV. 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 [1], [8].

A. 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 learning format [9],

$$\Theta \sim f(\Theta) \quad (17)$$

$$x_i|\Theta \sim f(x_i|\Theta) \equiv L(\Theta|x_i) \quad (18)$$

$$x_i, \Theta \sim f(x_i, \Theta) \quad (19)$$

$$f(\Theta|\overbrace{x_1,\ldots,x_n}^{\mathbf{x_n}}) \propto f(\Theta)\cdot\prod_{i=1}^{n} L(\Theta|x_i) \quad (20)$$

$$f(x_t|\mathbf{x_n}) \propto \int f(x_t|\Theta) f(\Theta|\mathbf{x_n})\, d\Theta. \quad (21)$$

[1]https://www.mathworks.com/help/stats/fitgmdist.html

In formulating externally-driven, "fictitious" data predictions, any positive time-value may be chosen by the modeller. Thus, we adopt the target data-predictor $X(t) = x(t),\ \forall\, t \in \mathbb{R}^+$, denoted as a static predictor $X_t = x_t$ for notational convenience.

B. Source Predictor at Time t

For the one-step-ahead *source predictor*, we condition on the target's isolated estimate of a_1 and r, introducing knowledge of the source parameter subvector $\Theta^* = (a_2, a_3)$ via the assertion $f(\Theta|\mathbf{x_n}) \equiv f_S(\Theta^*|\mathcal{I}_S)$. Additionally, we note that $\psi_1 = 1$ and (ψ_2, ψ_3) are functions of t, therefore we denote the associated time-dependent explanatory subvector as $\psi_t^* = (\psi_2, \psi_3)'$. In the $K = 1$ component case, the required data predictor is available as a standard result [9]. Taking a mixture of these predictive components, the full-form transferred data predictor (24) is, therefore, a K-component GMM.

$$f_S(x_t|\mathbf{x_n}, a_1, r)$$

$$\propto \int f(x_t|a_1, \Theta^*, r) f(\Theta^*|\mathbf{x_n}) d\Theta^* \quad (22)$$

$$= \sum_{k=1}^{K} \hat{\alpha}_k \int \mathcal{N}_{x_t}(\psi'\Theta, r)\, \mathcal{N}_{\Theta^*}(\hat{\mathbf{m}}_k, \hat{\Sigma}_k)\, d\Theta^* \quad (23)$$

$$= \sum_{k=1}^{K} \hat{\alpha}_k\, \mathcal{N}_{x_t}(m_{t,k}^{\dagger},\, r + \psi_t^{*'} \hat{\Sigma}_k \psi_t^*), \quad (24)$$

where $m_{t,k}^{\dagger} = a_1 + \psi_t^{*'} \hat{\mathbf{m}}_k$.

C. Complementary Prediction

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. Although the source knowledge is matched to the target, it is still reasonable to assume that data observed at the target is more relevant than the source knowledge we transfer to it. We thus choose to transfer knowledge at times where the target's data is scarce, termed *complementary times*. This avoids adding external information at times where the target already has data. If we choose to transfer information at target-complementary times, $t_c = t_1, \ldots, t_C$, over a window of $[1, W]$ days, we define for transfer the complementary data-predictive distribution, normalised by C, as

$$f_C(x_t|\mathbf{x}_n, a_1, r)$$

$$= \frac{1}{C}\sum_{c=1}^{C} f_S(x_t|\mathbf{x_n}, a_1, r)\, \delta(t - t_c) \quad (25)$$

$$= \frac{1}{C}\sum_{c=1}^{C}\sum_{k=1}^{K} \hat{\alpha}_k\, \mathcal{N}_{x_t}(m_{t,k}^{\dagger}, r)\, \delta(t - t_c) \quad (26)$$

This resulting distribution is the source knowledge object that is processed at the target via Bayesian transfer.

D. FPD-optimal transfer of the source predictor

The joint model for conditioning on the source knowledge, $\mathcal{I}_S$, is not available in our Bayesian transfer learning setting and so is designed optionally, using the axiomatic and optimal FPD framework, as done in [1] and other work [8], [10]. The resulting optimal conditional design is given by the mean-field operator (27). Thus we optimally process (26) via (27), denoting the observed data as D_n and the transfer weight as ν_S, i.e. the quantification of the target's trust in the source. The target distribution, $f(\Theta|D_n)$, is instantiated as the (constrained) NiG in (14), leading to the following expression for the posterior following FPD-optimal transfer.

$$
\begin{aligned}
&f(\Theta|D_n,\mathcal{I}_S)\\
&\quad\propto f(\Theta|D_n)\exp\left[\nu_S\int f_C(\varphi_t)\ln\left(f(\varphi_t|\Theta)\right)\,d\varphi_t\right] &(27)\\
&\quad\propto f(\Theta|D_n)\times\\
&\qquad\exp\left[\frac{\nu_S}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}\hat{\alpha}_k\,\mathcal{N}_{x_t}(m^{\dagger}_{t_c,k},r)\ln\left(f(x_{t_c}|\Theta)\right)\right] &(28)\\
&\quad\propto f(\Theta|D_n)\times\\
&\qquad\prod_{c=1}^{C}\prod_{k=1}^{K}\exp\left[\frac{\nu_S\hat{\alpha}_k}{C}\,\mathcal{N}_{x_t}(m^{\dagger}_{t_c,k},r)\ln\left(f(x_{t_c}|\Theta)\right)\right] &(29)\\
&\quad= f(\Theta|D_n)\prod_{c=1}^{C}\prod_{k=1}^{K}f(x_{t_c}|\Theta)^{\frac{\nu_S\hat{\alpha}_k}{C}\mathcal{N}_{x_t}(m^{\dagger}_{t_c,k},r)} &(30)
\end{aligned}
$$

This is the target (patient-specific) parameter inference, now benefiting from, i.e. optimally conditioned on, the transfer from the archive sub-population, as well as the target's own local observations.

E. NiG Parameter Update by External Predictor

In this paper, we take the same approach as in [1] by optimally processing an external predictor via the NiG parameters, but for the feature-space conditioned, complementary source data-predictor. Along with the source knowledge of Θ^*, we adopt the target's isolated marginal posterior estimates of mean a_1 and r from (12) as certainty equivalents for the transfer. The k-th component predicted extended datum, with uncertainty quantified by r, is

$$\varphi^{\dagger}_{t,k}=\begin{bmatrix}m^{\dagger}_{t,k}\\ \psi_t\end{bmatrix}. \tag{31}$$

We thus define the external batch update of the NiG parameter V_n by addition of V_C as

$$V_C=\nu_S\left(\frac{1}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}\hat{\alpha}_k\varphi^{\dagger}_{t_c,k}\varphi^{\dagger'}_{t_c,k}+r\begin{bmatrix}1\\0\\0\\0\end{bmatrix}\begin{bmatrix}1&0&0&0\end{bmatrix}\right) \tag{32}$$

This involves $c=1,2,\ldots,C<W$ complementary-time outer products, each derived as a certainty equivalent from a K-component predictor, with normalising constant C. The transferred mean $m^{\dagger}_{t,k}$ acts like an observation, but with its uncertainty reflected in the automatic adjustment via the transferred variance. The degree-of-freedom NiG parameter ν_n is incremented by ν_S in an approach similar to [1], where it is also noted that the value ν_S is data-driven. Thus, in this paper ν_S is left as a hyperparameter for testing. Defining $V_P=V_n+V_C$ and $\nu_P=\nu_n+\nu_S$, the posterior conditioned on source knowledge $\mathcal{I}_S$ is defined in (33) and the entire update is summarised in Algorithm 1.

$$f(a,r|\mathcal{I}_S,\mathcal{I}_H,D_n)\equiv\mathcal{NiG}_{a,r}(V_P,\nu_P)\chi_{\mathbb{A}}(a). \tag{33}$$

Algorithm 1: Outline of patient-specific update

Result: Posterior conditioned on source knowledge

1 initialise prior with V_0,ν_0 and hard constraints $\mathcal{I}_H$, $f(a,r|\mathcal{I}_H)$;
2 process local target data D_n to obtain the pre-posterior $f(a,r|\mathcal{I}_H,D_n)$;
3 select a GMM in (a_2,a_3) conditioned on patient metadata as the knowledge source $\mathcal{I}_S$;
4 optimally transfer a data-predictor at complementary times $c=1,\ldots,C$ to obtain the posterior, $f(a,r|\mathcal{I}_S,\mathcal{I}_H,D_n)$;

V. 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).

A. Choice of Performance Metric

For comparability between transfer methods, the performance metric adopted is the root mean square prediction error (RMSPE) of H/O observations (35). 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. We take a modified approach to that in [1], which simply evaluates the predicted error from a H/O of the fourth measurement. 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 also noted that all H/Os contain additive noise. The RMSPE is taken across all available H/Os, such that the predictive performance of the model is evaluated over the entire activity curve.

B. Test Hyperparameters

One of the hyperparameters used for testing is the size of the window, W (days), over which complementary times may be selected and hence transfer is permitted to occur. Here, the range of transfer windows tested is $W \in [6, 10]$ days. Note that in previous work in [1], statistics are transferred at fixed days $p = 1, 2, 10$, independent of W. Additionally, ν_S is data-driven in previous work, therefore tests are run for a range of values of ν_S here.

Algorithm 2: Performance evaluation

Result: RMSPE evaluation

1 Evaluate the parameter update of the regression 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 [1]):

$$\varepsilon_j = \mathbb{E}_{f(d_t|V_m,\nu_m)}\left[\ln(d_t)\right] - \ln(d_{t_j}) \qquad (34)$$

3 For the RMSPE, take the average of Euclidean norm error per sample j, between the extrapolated activity model and the h hold-out points:

$$RMSPE = \sqrt{\frac{\sum_{j=1}^{h} \varepsilon_j^2}{h}} = \frac{|\varepsilon|}{\sqrt{h}} \qquad (35)$$

C. Test Cases

1) Selective transfer: This is the proposed approach, involving a belief-weighted complementary merging of external statistics, obtained from a selected sub-population source modelled in the (a_2, a_3) feature space, processed as a GMM data-predictor and optimally transferred at target-complementary times;
2) Legacy transfer: The current state-of-the-art (SoTA) method employed in the work that formed the basis for this paper [1], involving a non-selective source modelled in the data-space over the entire population, transferred as a GMM data-predictor at fixed times of $k = 1, 2, 10$ days;
3) Control case: No transfer is performed and performances are taken based on estimates from the target's isolated regression model.

D. Results

The results of Algorithm 2 are shown in Figure 1 for a sample of class-1 patients. Note that some archive classes do not contain enough patients with $n \geq 4$ observations for which H/Os may be taken. As there are 387 class-1 patients who fulfil this criterion, class-1 is chosen for the following performance evaluation.

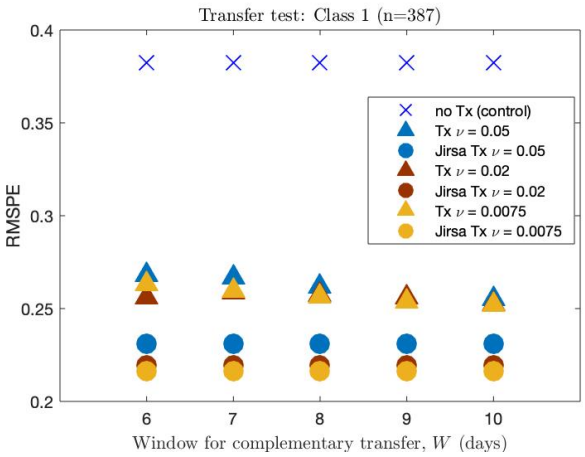

Fig. 1. Performance for a sample of patients in class-1 of the proposed transfer (triangles) versus the isolated patient (crosses) and SoTA transfer of Jirsa et al. [1] (discs), plotted for a range of W and transfer weight ν_S.

As expected the results for the control case and SoTA transfer are invariant to the window size W, this parameter being relevant only for complementary transfer. We note that both the proposed transfer and the SoTA transfer depend on the transfer weight ν_S. Figure 1 relates to a sample of class-1 patients, and performance might also be tested for other classes which have an adequate amount of data. In the performance study above, the average RMSPE for the proposed transfer approach is minimised at a transfer window size of $W = 10$ and a transfer weight of $\nu_S = 0.0075$. The performance additivity achieved for the proposed transfer is 34% versus an improvement of 43% for the SoTA.

VI. Discussion

This paper presents a methodology for Bayesian transfer of external knowledge to a target regression model, based on feature-space conditioning of source knowledge, partitioned and selected with respect to available metadata, and invoking the notion of complementary data transfer.

For transfer to the target, the source (one-step-ahead) probabilistic data-predictor is inferred from the source parameter distribution. This gives access to the same calculus employed in the seminal work for this investigation [1], [4] for FPD-optimal transfer of source knowledge to the target. The distinction in this work is that the probabilistic 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.

While the performance does not rival the SoTA [1] yet, the proposed approach is nevertheless a positive transfer algorithm, reducing the predictive error in the local target patient by 34%, on average, in the class-1 population. This performance improvement is sustained across values of the

test parameter W. A key benefit over [1] is that we model in the feature space (i.e. in the parameter estimate space), and thus the proposed algorithm is less application-specific: it may be generalised and used in Bayesian transfer learning contexts where source knowledge of the feature space exists, rather than direct observational knowledge. Also novel is the notion of selectivity, where features or metadata of the target learning task may be conditioned on to identify relevant source knowledge. These additions can be achieved without disrupting the recursive data processing at the target.

In order to achieve performance additivity beyond [1], we note the following for future work:

- There is significant sensitivity to the transfer weight ν_S, for which positive transfer is yielded for low values compared to the pre-posterior $\nu_n \geq 9.05$. This value was data-driven previously [1] but left as a hyperparameter here. It would benefit from formal specification, for example based on quantified uncertainty in the source knowledge domain.
- There is a large variance in the RMSPE, indicating heterogeneity is still present in the transfer. No additivity in performance was found for $k = 1$ versus $k = K$ components in the source GMMs. When the source distribution is processed through the FPD mean-field operator (27), second-order moments fail to transfer. This intrinsic limitation of FPD in the Gaussian context has recently been overcome by proposing a valid reversal of the underlying KLD objective [11]. It will be interesting to introduce this reversal here, since it has the potential to robustify the transfer from the source GMMs.
- Data are sparse in some classes. It would be beneficial for clinics to have access to more of these class-conditioned data or to harvest them. There are also alternative ways to partition metadata, for example since each patient is assigned a unique ID, a patient's regression model may be conditioned on their administration history, which reflects how therapy is designed in practice [1].

REFERENCES

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

[2] International Commission on Radiological Protection. 2007 recommendations of the international commission on radiological protection publication 103. *Ann ICRP*, 37(2.4):2, 2007.

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

[4] Miroslav Kárný and Tomáš Kroupa. Axiomatisation of fully probabilistic design. *Information Sciences*, 186(1):105–113, 2012.

[5] Václav Šmídl and Anthony Quinn. Mixture-based extension of the AR model and its recursive Bayesian identification. *IEEE Transactions on Signal Processing*, 53(9):3530–3542, 2005.

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

[7] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley, 2007.

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

[9] Christopher M. Bishop. *Pattern recognition and machine learning*. Information Science and Statistics. Springer, New York, 2006.

[10] Anthony Quinn, Miroslav Kárný, and Tatiana V. Guy. Optimal design of priors constrained by external predictors. *International Journal of Approximate Reasoning*, 84:150–158, 2017.

[11] Milan Papež and Anthony Quinn. Transferring model structure in bayesian transfer learning for gaussian process regression. *CoRR*, 2101.06884, 2021.