# Probabilistic estimation of radiation dose with few noisy measurements

Ferdinand Varga  Ladislav Jirsa  (with Anthony Quinn)

Charles University
Prague, Czech Republic

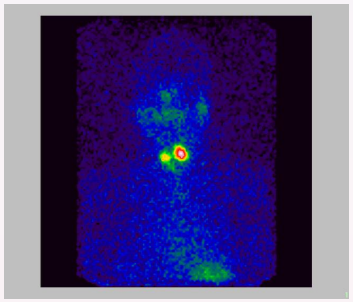
Academy of Sciences
Prague, Czech Republic

(Trinity College Dublin
Ireland)

6th March 2007
ÚTIA AV ČR

# Outline of Talk

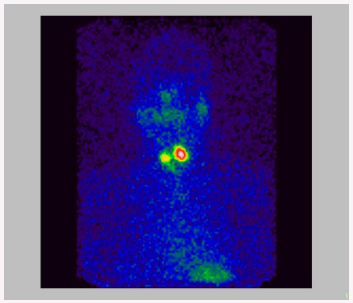
# Radiotherapy for Thyroid Cancer



**The Thyroid Gland:**

- Thyroid gland accumulates iodine from the blood;
- Iodine used in the production of thyroid hormones;
- Central to regulation of heart, metabolism, *etc.* by the *endocrine system*.

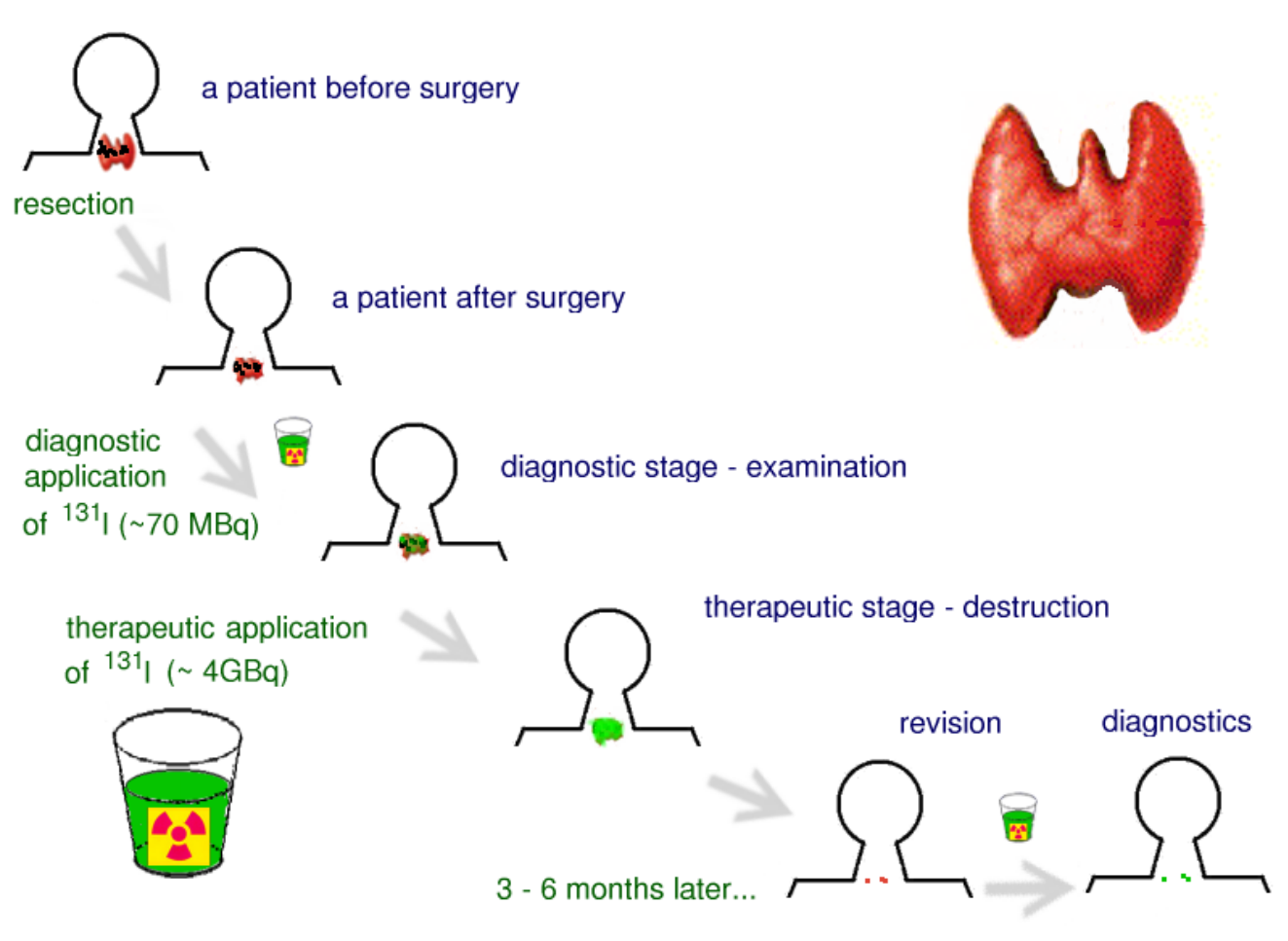
# Radiotherapy for Thyroid Cancer



**Thyroid Cancer:**

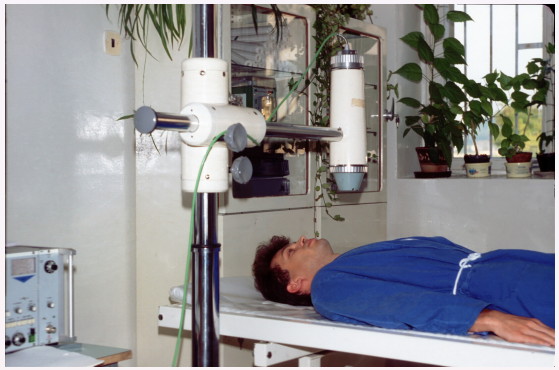
- Affects $\sim$5/100 000 people in Europe (60 % female);
- Nuclear Medicine (Radiotherapy): selective accumulation of radioactive iodine, $^{131}$I, by thyroid;
- tumour is destroyed by $\beta$-radiation (2-10 GBq);
- survival after treatment is $\sim$95 % for children, $\sim$55 % for the elderly.

# Treatment Schedule after Thyroid Cancer Diagnosis



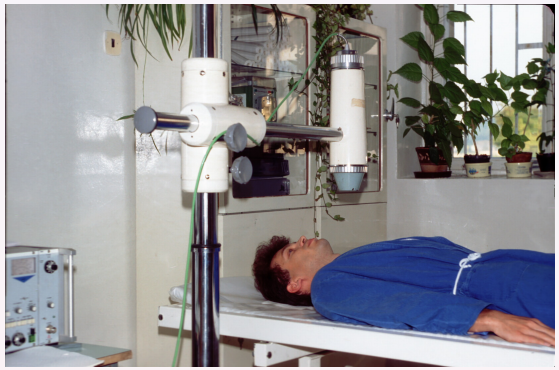
# Measurement, $d_t$, of Thyroid Activity, $A_t$

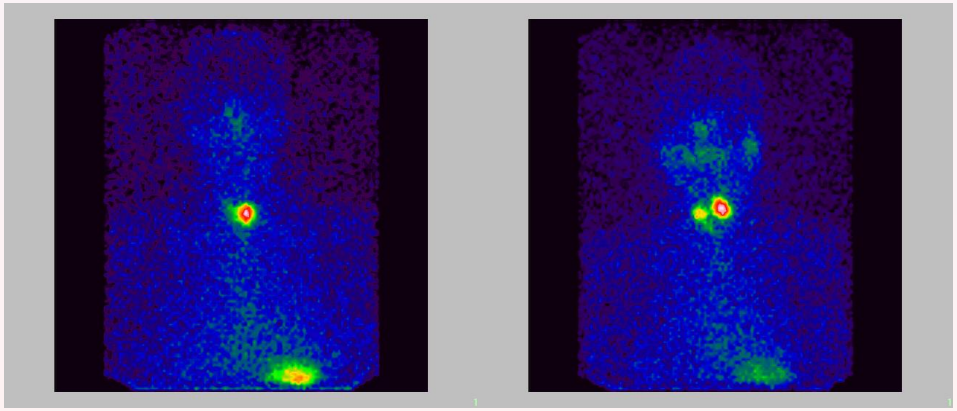
- Use of a scintillation probe or $\gamma$-camera;
- Detection of $\gamma$-photons.



# Measurement, $d_t$, of Thyroid Activity, $A_t$

- Use of a scintillation probe or $\gamma$-camera;
- Detection of $\gamma$-photons.





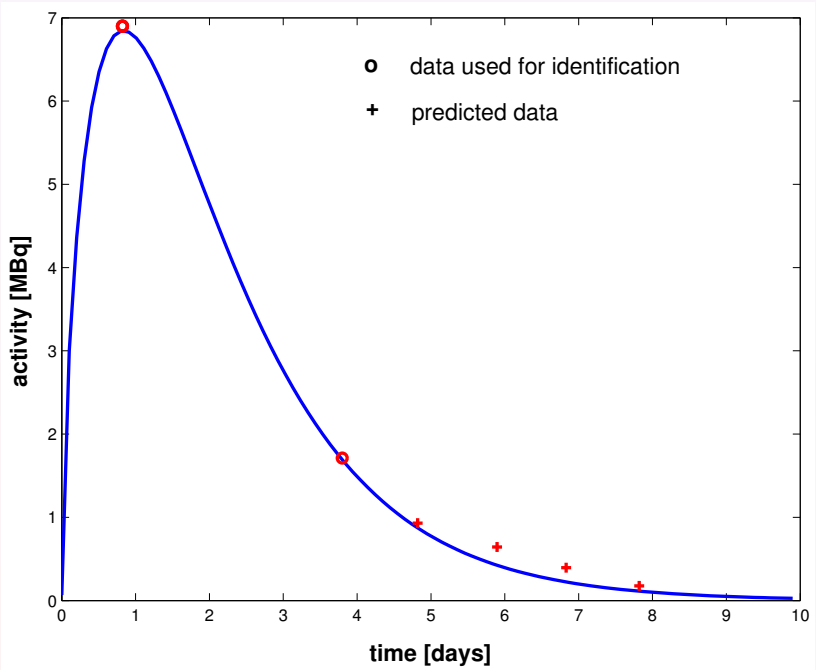
# Measurement, $d_t$, of Thyroid Activity, $A_t$

- The cumulative count in a prescribed Region-of-Interest at time $t$ is Poisson;
- Measured activity, $d_t$, is a scaled version via a calibration of the measurement equipment;
- $d_t$: log-normal (approximately, given some assumptions); $\ln(d_t)$: Normal (approximately);

# Measurement, $d_t$, of Thyroid Activity, $A_t$

- The cumulative count in a prescribed Region-of-Interest at time $t$ is Poisson;
- Measured activity, $d_t$, is a scaled version via a calibration of the measurement equipment;
- $d_t$: log-normal (approximately, given some assumptions); $\ln(d_t)$: Normal (approximately);
- BUT...high noise.

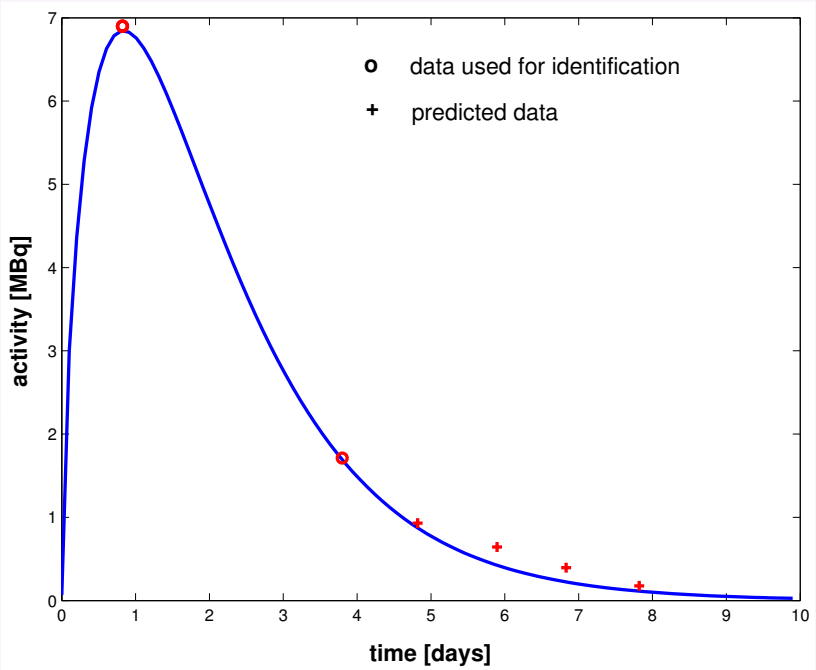
# Measurement, $d_t$, of Thyroid Activity, $A_t$




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

| $i$ | time $t_i$ [days] | meas. activity $d_i$ |
|---|---|---|
| 1 | 0.823 | 6.899 |
| 2 | 3.799 | 1.711 |

# Measurement, $d_t$, of Thyroid Activity, $A_t$




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

| $i$ | time $t_i$ [days] | meas. activity $d_i$ |
|---|---|---|
| 1 | 0.823 | 6.899 |
| 2 | 3.799 | 1.711 |

Typically, only 2–4 data pairs, $(t_i, d_i)$, are available, for economic and ethical reasons.

# The Two Key Inference Tasks

- PREDICT $A_t$: important for radio-hygienic management of the patient;

# The Two Key Inference Tasks

- PREDICT $A_t$: important for radio-hygienic management of the patient;
- INFERENCE of dose $S\xi$, particularly its uncertainty.

$$\xi = \int_0^\infty A_t \, \mathrm{d}t$$

$S$: an organ-specific constant (MIRD)

# The Two Key Inference Tasks

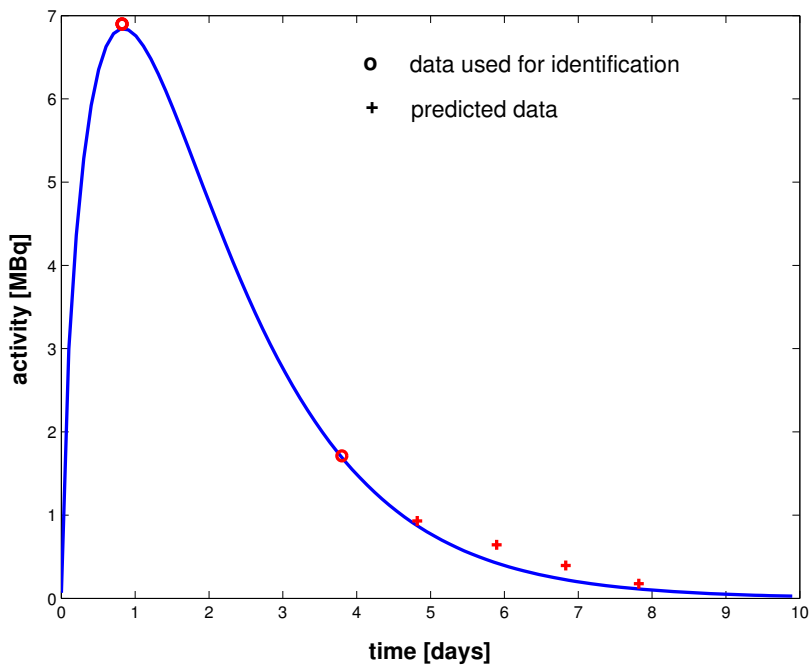
- PREDICT $A_t$: important for radio-hygienic management of the patient;
- INFERENCE of dose $S\xi$, particularly its uncertainty.

$$\xi = \int_0^\infty A_t \, \mathrm{d}t$$

$S$: an organ-specific constant (MIRD)

A stressful regime of inference is implied.

# Traditional Approaches to these Tasks




- CLEARANCE only is modelled;
- UPTAKE neglected, or treated heuristically;
- Informal treatment of uncertainty in $\xi$.

# The Bayesian Approach

Two aspects of the problem recommend a Bayesian attack:

- A parametric observation model;
- Two sources of prior knowledge:
  - side-information from a population of patients (database);
  - hard parameter constraints.

# The Bi-Phasic Model for $A_t$

A model has been proposed (Hermanska *et al.*, 2001):

$$\ln A_t = a_1 + a_2 \ln(c\,t) + a_3 (c\,t)^{\frac{2}{3}} \ln(c\,t) - \frac{t}{T_\mathrm{p}} \ln 2$$

| | |
|---|---|
| $A_t$ | activity in time $t$ |
| $T_\mathrm{p}$ | physical half-life of $^{131}$I (8.04 days) |
| $a \equiv (a_1, a_2, a_3)'$ | model parameters |
| $c$ | time scale hyperparameter (see later) |
| $\psi_t \equiv \left(1,\ \ln(c\,t),\ (c\,t)^{2/3} \ln(c\,t)\right)'$ | regressor |
| $x_t \equiv \ln d_t + \frac{t}{T_\mathrm{p}} \ln 2 = \psi_t' a + e_t$ | measurement process $d_t$ |
| $e_t \sim \mathcal{N}(0, r)$ | uncorrelated white noise |
| $\Psi_t \equiv (x_t, \psi_t')'$ | extended regressor |

The observation model is Normal:

$$f(x_t | a, r) = \mathcal{N}_{x_t}(\psi_t' a, r) = \frac{1}{\sqrt{2\pi r}} \exp\left\{ -\frac{(x_t - \psi_t' a)^2}{2r} \right\}$$

# The Conjugate Distribution

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

The conjugate distribution is *Gauss-inverse-Wishart*:
(or *Normal-inverse-Gamma*)

$$f(a,r|D) = \mathcal{GiW}(V_n, \nu_n)$$

$$\begin{aligned} V_n &= V_0 + \sum_{i=1}^{n} \Psi_{t_i} \Psi_{t_i}' \quad \text{(extended information matrix)} \\ \nu_n &= \nu_0 + n \quad \text{(degrees of freedom)} \end{aligned}$$

# The Conjugate Distribution

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

The conjugate distribution is *Gauss-inverse-Wishart*:
(or *Normal-inverse-Gamma*)

$$f(a,r|D) = \mathcal{GiW}(V_n, \nu_n)$$

$$\begin{aligned} V_n &= V_0 + \sum_{i=1}^{n} \Psi_{t_i} \Psi_{t_i}' \quad \text{(extended information matrix)} \\ \nu_n &= \nu_0 + n \quad \text{(degrees of freedom)} \end{aligned}$$

We seek to build the statistics of the prior from available external information.

# The External Information: A Patient Database

A large database is available at Motol Hospital, Prague:

- 3 876 data sequences of thyroid activity measurements;
- Each sequence contains $2 \leq n \leq 7$ data pairs $\{(t_i, d_i)\}_{i=1}^{n}$.

# The External Information: A Patient Database

A large database is available at Motol Hospital, Prague:

- 3 876 data sequences of thyroid activity measurements;
- Each sequence contains $2 \leq n \leq 7$ data pairs $\{(t_i, d_i)\}_{i=1}^{n}$.

Pre-processing:

- Measured activities, $d_{t_i}$, in each sequence were normalized by the maximum activity in the sequence;
- For each $t_i$, normalized activities $\{d_{t_i}\}$ were averaged across patients;
- average normalized activities for $t = 1, 2$ and 10 days were chosen to represent the population-based $d_t$;
- The corresponding 3 extended regressors, $\{\Psi_{0,i}\}_{i=1}^{3}$, were constructed as sufficient statistics for the database.

# The Data-Informed Prior

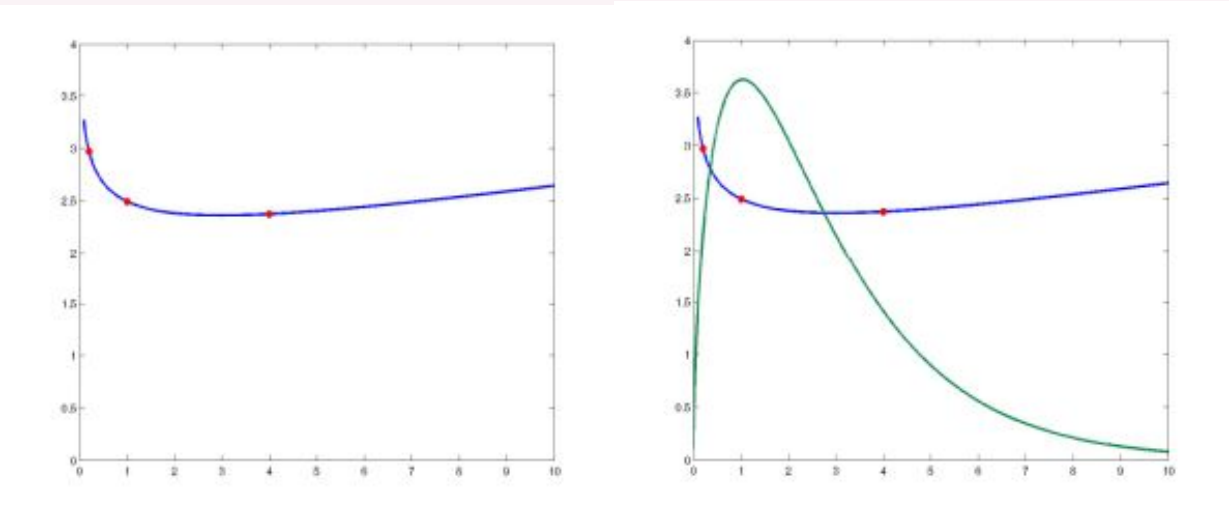
The data-informed prior:

$$f(a, r|\mathcal{I}_0) = \mathcal{GiW}(V_0, \nu_0)$$

- $\nu_0 = 7.05 \quad \Rightarrow \quad E[r]$ exists;
- $V_0 = \sum_{i=1}^{3} \left(\pi_i \pi_i' + \rho\rho'\right)$, where [Kracik, 2005]
  - $\pi_i = \alpha\Psi_{0,i}$, and $\alpha = 0.001$ weights the side information;
  - $\rho = (\sqrt{\hat{r}_0}, 0, 0, 0)'$;
  - $\hat{r}_0 = 0.019$ is the variance of $e_t$ estimated from the patient database.

# Hard Parameter Constraints

Unconstrained inference of $a$ implies non-physical $A_t$ for some patient cases.



# Hard Parameter Constraints

- $A_t \to 0^+$ as $t \to 0^+$ and as $t \to +\infty$;
- $A_t$ achieves a unique global maximum at $t_\mathrm{m} > 0$;
- from medical experience, $t_\mathrm{m} \in \langle t_\mathrm{l}, t_\mathrm{u} \rangle$, where $t_\mathrm{l} = 4$ hours (0.167 days) and $t_\mathrm{u} = 72$ hours (3 days);
- for $t > t_\mathrm{h} > t_\mathrm{m}$, $A_t$ decreases faster than the decrease caused by the physical decay of $^{131}$I, i.e. by the term $-\frac{t}{T_\mathrm{p}} \ln 2$.

# Hard Parameter Constraints

- $A_t \to 0^+$ as $t \to 0^+$ and as $t \to +\infty$;
- $A_t$ achieves a unique global maximum at $t_\mathrm{m} > 0$;
- from medical experience, $t_\mathrm{m} \in \langle t_\mathrm{l}, t_\mathrm{u} \rangle$, where $t_\mathrm{l} = 4$ hours (0.167 days) and $t_\mathrm{u} = 72$ hours (3 days);
- for $t > t_\mathrm{h} > t_\mathrm{m}$, $A_t$ decreases faster than the decrease caused by the physical decay of $^{131}$I, i.e. by the term $-\frac{t}{T_\mathrm{p}} \ln 2$.

This implies

- $c = 1.3388$, $\mathbb{A} = \left\{ a \left| \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 4.8687 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} < \begin{bmatrix} 0 \\ 0.2586 \\ -0.0144 \end{bmatrix} \right. \right\}$
- $\Rightarrow \chi_{\mathbb{A}}(a)$

# Posterior Inference of the Bi-Phasic Parameters

Bayes' theorem with data-informed prior and hard constraints:

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

$$\psi_{t_i} = \begin{bmatrix} 1 \\ \ln(c\,t) \\ (c\,t)^{\frac{2}{3}} \ln(c\,t) \end{bmatrix} \qquad \begin{aligned} V_n &= V_0 + \sum_{i=1}^{n} \Psi_{t_i} \Psi'_{t_i} \\ \nu_n &= \nu_0 + n \end{aligned}$$

# Posterior Inference of the Bi-Phasic Parameters

Bayes' theorem with data-informed prior and hard constraints:

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

$$\psi_{t_i} = \begin{bmatrix} 1 \\ \ln(c\,t) \\ (c\,t)^{\frac{2}{3}} \ln(c\,t) \end{bmatrix} \quad \begin{aligned} V_n &= V_0 + \sum_{i=1}^{n} \Psi_{t_i} \Psi'_{t_i} \\ \nu_n &= \nu_0 + n \end{aligned}$$

Conjugate update preserved. In particular:

$$f(a | D, \mathcal{I}_0, \mathcal{I}_c) = \mathcal{S}t_a(V_n, \nu_n) \; \chi_{\mathbb{A}}(a).$$

# Posterior Inference of the Bi-Phasic Parameters

$$f(a|D,\mathcal{I}_0,\mathcal{I}_c) = \mathcal{S}t_a(V_n,\nu_n)\ \chi_{\mathbb{A}}(a).$$

HOWEVER:

- moments of $f(a|D,\mathcal{I}_0,\mathcal{I}_c)$ are unavailable in closed form: problem of $\chi_{\mathbb{A}}(a)$;
- $f(\xi|D,\mathcal{I}_0,\mathcal{I}_c)$ is unavailable in closed form:

  mapping $a \to \xi(a) = \int_0^\infty A_t\, dt$ is unavailable in closed form.

# Posterior Inference of the Bi-Phasic Parameters

$$f(a|D,\mathcal{I}_0,\mathcal{I}_c) = \mathcal{S}t_a(V_n,\nu_n)\ \chi_{\mathbb{A}}(a).$$

HOWEVER:

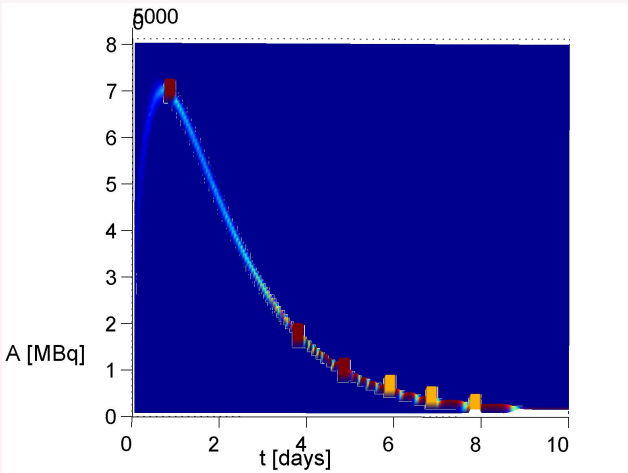
- moments of $f(a|D,\mathcal{I}_0,\mathcal{I}_c)$ are unavailable in closed form: problem of $\chi_{\mathbb{A}}(a)$;
- $f(\xi|D,\mathcal{I}_0,\mathcal{I}_c)$ is unavailable in closed form:
  mapping $a \to \xi(a) = \int_0^\infty A_t\, dt$ is unavailable in closed form.

$\Rightarrow$ Empirical approximation of $f(a|D,\mathcal{I}_0,\mathcal{I}_c)$ and $f(\xi|D,\mathcal{I}_0,\mathcal{I}_c)$.

# MCMC by Langevin Diffusion Algorithm

**Langevin Diffusion Algorithm** (Roberts *et al.*, 1996)

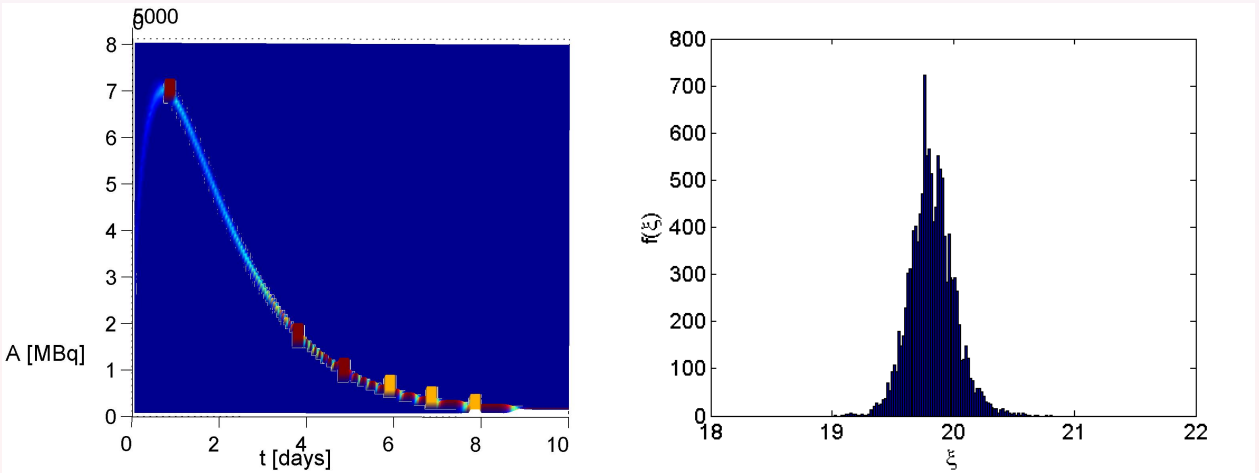
- Used to sample from $f(a|D, \mathcal{I}_0, \mathcal{I}_c)$;
- Higher acceptance rate, and, therefore, faster convergence than Metropolis-Hastings.



# MCMC by Langevin Diffusion Algorithm

**Langevin Diffusion Algorithm** (Roberts *et al.*, 1996)

- Used to sample from $f(a|D, \mathcal{I}_0, \mathcal{I}_c)$;
- Higher acceptance rate, and, therefore, faster convergence than Metropolis-Hastings.




$\xi(a) = \int_0^\infty A_t \, \mathrm{d}t$ evaluated by numerical integration.

# MCMC by Langevin Diffusion Algorithm

- **Initialization** of the Markov Chain by the MAP estimate;
- **Step-size** tuned empirically;
- **Stopping rule** via relaxation to a Dirichlet measure (Quinn+Karny, 2006):

$$f(a|D,\mathcal{I}_0) \quad \rightarrow \quad \mathcal{D}(\hat{F}_i,\nu_i') \equiv \mathcal{D}_i, \qquad \text{after } i \text{ i.i.d. samples}$$

$$N \quad = \quad \min\left\{i \ : \ \mathrm{KLD}\left[\mathcal{D}_i||\mathcal{D}_{i-1};\mathbb{P}_{K_i}\right] < \epsilon\right\}.$$

# MCMC by Langevin Diffusion Algorithm

- Setting $\epsilon = 0.002$, $E[N] = 4\,529$, estimated on 700 patients;
- Variability of $\xi$-inference less than 3 % compared to using 50 000 samples;
- Practicable Bayesian implementation: one run of MC takes less than 1 s on PC (1 GHz).

# Prediction of Activity, $A_t$

**Patient-Specific Prediction of $A_t$:**

- 2 355 patients from the database; 4 measurements each;
- For each patient, $n = 3$ measurements used for inference;
- $4^{\text{th}}$ measurement used to evaluate the prediction, $\hat{A}_{t_4}$;
- Distribution of prediction error for three prior knowledge structures ($\nu_0 = 7.05$):

# Prediction of Activity, $A_t$

**Patient-Specific Prediction of $A_t$:**

- 2 355 patients from the database; 4 measurements each;
- For each patient, $n = 3$ measurements used for inference;
- $4^{\text{th}}$ measurement used to evaluate the prediction, $\hat{A}_{t_4}$;
- Distribution of prediction error for three prior knowledge structures ($\nu_0 = 7.05$):

| $V_0$ | $\chi_{\mathbb{A}}(a)$ | **mean** | **median** | **st.dev.** | **data** | **outliers** |
|---|---|---|---|---|---|---|
| $V_0^{(1)}$ | no | 0.0576 | −0.0066 | 0.475 | 1 403 | 2.28 % |
| $V_0^{(1)}$ | yes | −0.0968 | −0.1456 | 0.431 | 2 355 | 0.85 % |
| $V_0^{(2)}$ | yes | −0.0004 | −0.0544 | 0.416 | 2 355 | 0.81 % |

$$V_0^{(1)} = 10^{-6}\, I_4, \qquad V_0^{(2)} = \sum_{i=1}^{3} \left(\pi_i \pi_i' + \rho\rho'\right)$$

# Inference of Dose, $\xi$

Empirical approximation of $f(\xi|D,\mathcal{I}_0)$ studied for skewness:

| $f(\xi\|D,\mathcal{I}_0)$ | $\hat{\xi}$ | **median** | $\sigma_\xi$ |
|---|---|---|---|
| $\xi$ | 1.66 | 0.84 | 3.53 |
| ln $\xi$ | 0.29 | 0.24 | 0.61 |

- Supports hypothesis that $f(\xi|D,\mathcal{I}_0,\mathcal{I}_c)$ is log-normal;
- Bayesian test normal vs. log-normal $\Rightarrow f(\xi|\cdot)$ is log-normal
- A parametric approximation for $f(\xi|D,\mathcal{I}_0,\mathcal{I}_c)$ might be sought, however, log-normal is sufficient for practice.

# The Clinical Prospects for this Work

- To compare thyroid responses to low and high administations of $^{131}$I (the *stunning* effect);
- To use $f(\xi|D, \mathcal{I}_0, \mathcal{I}_c)$ in the design of a Bayesian medical decision support system;
- To use $f(\xi|D, \mathcal{I}_0, \mathcal{I}_c)$ to infer dose in other organs (MIRD methodology).

# Conclusions

- Inference of $A_t$ and $\xi$ via the biphasic model, in a *stressful* regime;

# Conclusions

- Inference of $A_t$ and $\xi$ via the biphasic model, in a *stressful* regime;
- The Bayesian prior is vital:
  - hard constraints on $a$;
  - data-informed conjugate prior via side-information (a patient database);
- Data-informed prior decreased mean of prediction error of $A_t$ by 9 %; variance also reduced;

# Conclusions

- Inference of $A_t$ and $\xi$ via the biphasic model, in a *stressful* regime;
- The Bayesian prior is vital:
  - hard constraints on $a$;
  - data-informed conjugate prior via side-information (a patient database);
- Data-informed prior decreased mean of prediction error of $A_t$ by 9 %; variance also reduced;
- MCMC via Langevin diffusion algorithm: patient-specific tuning of sample size via stopping rule ($E[N] = 4\,529$);

# Conclusions

- Inference of $A_t$ and $\xi$ via the biphasic model, in a *stressful* regime;
- The Bayesian prior is vital:
  - hard constraints on $a$;
  - data-informed conjugate prior via side-information (a patient database);
- Data-informed prior decreased mean of prediction error of $A_t$ by 9 %; variance also reduced;
- MCMC via Langevin diffusion algorithm: patient-specific tuning of sample size via stopping rule ($E[N] = 4\,529$);
- Bayesian elicitation of $f(\xi|D, \mathcal{I}_0, \mathcal{I}_c)$ provides doctors with valuable precision quantification for absorbed dose, $\xi$.