A Bayesian hierarchical model to integrate dietary exposure and biomarker measurements for the risk of cancer

Marta Pittavino, PhD

Senior Research Associate, Research Center for Statistics UNIGE, Geneva - Switzerland
Summary

- Outline on measurement error
  - Combine biomarkers with self-reports
- A Bayesian hierarchical model with three components
  - An application from the EPIC study
- Concluding remarks
Outline on measurement error

- Difference between the actual true value and the measured value
Outline on measurement error

- Difference between the actual true value and the measured value

- Self-reported assessments (questionnaire: Q and 24-HDR: R) of dietary exposure prone to random and systematic measurement errors
Outline on measurement error

- Difference between the actual true value and the measured value

- Self-reported assessments (questionnaire: Q and 24-HDR: R) of dietary exposure prone to random and systematic measurement errors

- Estimates of the association between dietary factors and risk of disease can be biased
Outline on measurement error

- Difference between the actual true value and the measured value

- Self-reported assessments (questionnaire: Q and 24-HDR: R) of dietary exposure prone to random and systematic measurement errors

- Estimates of the association between dietary factors and risk of disease can be biased

It has been suggested to complement self-reports with objective measurements, as dietary biomarkers (M)
Why combine self-reports and biomarkers?

- Objective biomarkers have long integrated self-reported dietary (or physical activity) measurements

  - Several validation studies with recovery (OPEN, NHS, WHI, EPIC) and concentration biomarker measurements (Ferrari et al., AJE, 2007)

  - Biomarkers also integrated in diet-disease association studies (Freedman et al., EPI, 2011; Tasevska et al., AJE, 2018; Prentice et al., AJE, 2017)
Motivating scenario

- Plasma levels of Vit-B6 were inversely associated with kidney cancer risk in EPIC (Johansson, JNCI, 2014)
- Serum levels of Vit-B6, unlike Q measurements, were inversely related to lung cancer (Johansson, JAMA, 2010)
- Dietary folate inversely related to breast cancer (de Battle, JNCI, 2014), unlike plasma levels (Matejcic, IJC, 2017)
Application

- Data from two EPIC nested case-control studies on kidney (n=1,108) and lung (n=1,764) cancers

  - Q, R and M measurements on vit-B6 and folate

  R measurements, available on ~10% of the sample, were imputed

  Data were log-transformed to approximate normality
The EPIC Study

- Prospective cohort with 500,000 participants from 23 centres
  - Dietary and lifestyle exposures assessed at baseline
  - Biological samples collected at baseline from 80% disease-free participants
The setting

- Develop a Bayesian latent factor hierarchical model to:
  
  - Integrate self-reported (Q) and (R) with concentration biomarker (M) measurements
  
  - Evaluate the measurement error structure
  
  - Estimate the unknown association between unknown true dietary intakes (X) and risk of disease (Y)
How did we get here?

- Long time ago, Ray Caroll and Pietro Ferrari discussed the idea of a Bayesian model for measurement errors

  - DQs and 24-HDRs of energy and fat intakes were related to the risk of breast cancer in EPIC

Pietro Ferrari and Martyn Plummer submitted an application to WCRF in 2013.

First results were produced last year.
A Bayesian multilevel model for estimating the diet/disease relationship in a multicenter study with exposures measured with error: The EPIC study

Pietro Ferrari\textsuperscript{1, *}, †, Raymond J. Carroll\textsuperscript{2}, Paul Gustafson\textsuperscript{3} and Elio Riboli\textsuperscript{4}

\textsuperscript{1}International Agency for Research on Cancer, Lyon, France
\textsuperscript{2}Texas A&M University, College Station, U.S.A.
\textsuperscript{3}Department of Statistics, University of British Columbia, Vancouver, Canada
\textsuperscript{4}St Mary’s Campus, Imperial College, London, U.K.
How did we get here?

- Long time ago, Ray Caroll and Pietro Ferrari discussed the idea of a Bayesian model for measurement errors
  - Combine DQs and 24-HDRs of energy and fat intakes, and relate them to risk of breast cancer in EPIC

- Pietro Ferrari and Martyn Plummer submitted an application to WCRF in 2013
  - First results were produced in 2017
The Bayesian hierarchical model

1) Exposure model:  \( p(X|\mu, \Sigma) \)

2) Measurement model:  \( p(Q, R, M|X) \)

3) Disease model:  \( p(Y|X) \)
The Bayesian hierarchical model

1) Exposure model: \( p(X|\mu, \Sigma) \)

2) Measurement model: \( p(Q, R, M|X) \)

3) Disease model: \( p(Y|X) \)
The Bayesian hierarchical model

1) Exposure model:  
\[ p(X|\mu, \Sigma) \]

2) Measurement model:  
\[ p(Q, R, M|X) \]

3) Disease model:  
\[ p(Y|X) \]
1. The exposure model

- $X_{ik}$: vector of unknown true (latent factor) dietary intake of vit-B6 and folate ($k = 1, 2$), with $i = 1, \ldots, n$:

  $$X_{ik} \sim MVN(\mu_k, \Sigma_X) = MVN(0, \Sigma_X)$$

  $$\Sigma_X^{-1} \sim Wishart(D_x, r_x)$$
1. The exposure model

- $X_{ik}$: vector of unknown true (latent factor) dietary intake of vit-B6 and folate ($k = 1, 2$), with $i = 1, \ldots, n$:

  $$X_{ik} \sim MVN(\mu_k, \Sigma_X) = MVN(0, \Sigma_X)$$

  $$\Sigma_X^{-1} \sim \text{Wishart}(D_x, r_x)$$

- $D_x$: scale matrix, $(k \times k)$

- $r_x$: rank of the Wishart distribution
2. The measurement model

for \( i = 1, \ldots, n \) and \( k = 1, 2 \):

\[
Q_{ik} = \alpha_{Q_k} + \beta_{Q_k} \cdot X_{ik} + \epsilon_{Q_k}
\]

\[
R_{ik} = X_{ik} + \epsilon_{R_k}
\]

\[
M_{ik} = \alpha_{M_k} + \beta_{M_k} \cdot X_{ik} + \epsilon_{M_k}
\]
2. The measurement model

for $i = 1, \ldots, n$ and $k = 1, 2$:

\[
Q_{ik} = \alpha_{Q_k} + \beta_{Q_k} \cdot X_{ik} + \epsilon_{Q_k}
\]
\[
R_{ik} = X_{ik} + \epsilon_{R_k}
\]
\[
M_{ik} = \alpha_{M_k} + \beta_{M_k} \cdot X_{ik} + \epsilon_{M_k}
\]

- Assumptions:

\[
\text{cov} (\epsilon_{Q_k}, \epsilon_{R_k}) \neq 0, \text{cov} (\epsilon_{Q_1}, \epsilon_{Q_2}) \neq 0, \text{cov} (\epsilon_{R_1}, \epsilon_{R_2}) \neq 0,
\]
2. The measurement model

for $i = 1, \ldots, n$ and $k = 1, 2$:

\[
\begin{align*}
Q_{ik} &= \alpha Q_k + \beta Q_k \cdot X_{ik} + \epsilon Q_k \\
R_{ik} &= X_{ik} + \epsilon R_k \\
M_{ik} &= \alpha M_k + \beta M_k \cdot X_{ik} + \epsilon M_k
\end{align*}
\]

- Assumptions:

\[
\begin{align*}
\text{cov}(\epsilon Q_k, \epsilon R_k) &\neq 0, \quad \text{cov}(\epsilon Q_1, \epsilon Q_2) \neq 0, \quad \text{cov}(\epsilon R_1, \epsilon R_2) \neq 0, \\
\text{cov}(\epsilon Q_k, \epsilon M_k) &= 0, \quad \text{cov}(\epsilon R_k, \epsilon M_k) = 0, \quad \text{cov}(\epsilon M_1, \epsilon M_2) = 0
\end{align*}
\]
2. The measurement model (ii)

for \( i = 1, \ldots, n \) and \( k = 1, 2 \):

\[
Q_{ik} = \alpha Q_k + \beta Q_k \cdot X_{ik} + \epsilon Q_k \\
R_{ik} = X_{ik} + \epsilon R_k \\
M_{ik} = \alpha M_k + \beta M_k \cdot X_{ik} + \epsilon M_k
\]
2. The measurement model (ii)

for \( i = 1, \ldots, n \) and \( k = 1, 2 \):

\[
Q_{ik} = \alpha_{Q_k} + \beta_{Q_k} \cdot X_{ik} + \epsilon_{Q_k}
\]
\[
R_{ik} = X_{ik} + \epsilon_{R_k}
\]
\[
M_{ik} = \alpha_{M_k} + \beta_{M_k} \cdot X_{ik} + \epsilon_{M_k}
\]

- Prior distributions:

\[
\alpha_{Q_k} \sim N(0, \sigma_{\alpha Q_k}^2), \quad \beta_{Q_k} \sim N(0, \sigma_{\beta Q_k}^2)
\]
\[
\epsilon_{QR} \sim MVN(0, \Sigma_{\epsilon QR}), \quad \Sigma_{\epsilon QR}^{-1} \sim \text{Wishart}(D_{\epsilon QR}, r_{\epsilon QR})
\]
2. The measurement model (ii)

for \( i = 1, \ldots, n \) and \( k = 1, 2 \):

\[
\begin{align*}
Q_{ik} & = \alpha_{Q_k} + \beta_{Q_k} \cdot X_{ik} + \epsilon_{Q_k} \\
R_{ik} & = X_{ik} + \epsilon_{R_k} \\
M_{ik} & = \alpha_{M_k} + \beta_{M_k} \cdot X_{ik} + \epsilon_{M_k}
\end{align*}
\]

- Prior distributions:

\[
\begin{align*}
\alpha_{Q_k} & \sim N(0, \sigma_{\alpha_{Q_k}}^2), & \beta_{Q_k} & \sim N(0, \sigma_{\beta_{Q_k}}^2) \\
\epsilon_{QR} & \sim \text{MVN}(0, \Sigma_{\epsilon_{QR}}), & \Sigma_{\epsilon_{QR}}^{-1} & \sim \text{Wishart}(D_{\epsilon_{QR}}, r_{\epsilon_{QR}}) \\
\alpha_{M_k} & \sim N(0, \sigma_{\alpha_{M_k}}^2), & \beta_{M_k} & \sim N(0, \sigma_{\beta_{M_k}}^2) \\
\epsilon_{M} & \sim \text{MVN}(0, \Sigma_{\epsilon_{M}}), & \Sigma_{\epsilon_{M}}^{-1} & \sim \text{Wishart}(D_{\epsilon_{M}}, r_{\epsilon_{M}})
\end{align*}
\]
3. The disease model

- $Y_i \in (0, 1)$: disease indicator for the $i^{th}$ study subject

- $Y_i \sim Bin(1, \pi_i)$, with $\pi_i$ the probability of developing the disease

- A conditional logistic model is assumed as:

$$P(Y_i|X_k, Z_p) = H(\gamma_1 \cdot X_{i1} + \gamma_2 \cdot X_{i2} + Z_{ip}^T \gamma_3)$$
3. The disease model

- $Y_i \in (0, 1)$: disease indicator for the $i^{th}$ study subject
  
  $Y_i \sim Bin(1, \pi_i)$, with $\pi_i$ the probability of developing the disease

- A conditional logistic model is assumed as:
  
  \[
P(Y_i|X_k, Z_p) = H(\gamma_1 \cdot X_{i1} + \gamma_2 \cdot X_{i2} + Z_{ip}^T \gamma_3)
  \]

  $\Gamma = (\gamma_1, \gamma_2, \gamma_3) \sim N(0, \sigma_\Gamma^2_{(k+p)})$
Analytical steps

- R measurements, available on ~10% of the sample, were imputed as $E(R|Q)$
- Data were log-transformed to approximate normality
- Residuals were computed by country, study, smoking, batch, age and sex (M), country, age and sex (Q), age and sex (R)
- Disease models were run separately by study
- Only results for kidney cancer disease model are shown
- Analyses were run in R with JAGS (Martyn Plummer)
3. **Results: disease model**

Table 3. Relative risk estimates ($\hat{R}R_k = e^{\hat{\gamma}_k}$).

<table>
<thead>
<tr>
<th></th>
<th>Vit-B6</th>
<th></th>
<th>Folate</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{R}R_1$</td>
<td>(95% CI)</td>
<td>$\hat{R}R_2$</td>
<td>(95% CI)</td>
</tr>
<tr>
<td>Q</td>
<td>1.00</td>
<td>(0.88, 1.14)</td>
<td>0.96</td>
<td>(0.84, 1.10)</td>
</tr>
<tr>
<td>R</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>M</td>
<td>0.79</td>
<td>(0.69, 0.89)</td>
<td>0.95</td>
<td>(0.84, 1.07)</td>
</tr>
</tbody>
</table>
3. Results: disease model

Table 3. Relative risk estimates \( \hat{R}_k = e^{\hat{\gamma}_k} \).

<table>
<thead>
<tr>
<th></th>
<th>Vit-B6</th>
<th>Folate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{R}_1 ) (95% CI(^1))</td>
<td>( \hat{R}_2 ) (95% CI(^1))</td>
</tr>
<tr>
<td>Q</td>
<td>1.00 (0.88, 1.14)</td>
<td>0.96 (0.84, 1.10)</td>
</tr>
<tr>
<td>R</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>M</td>
<td>0.79 (0.69, 0.89)</td>
<td>0.95 (0.84, 1.07)</td>
</tr>
<tr>
<td>X</td>
<td>0.71 (0.58, 0.88)</td>
<td>0.85 (0.72, 1.02)</td>
</tr>
</tbody>
</table>

\(^1\)Credible Intervals
Concluding remarks

- Challenging to tackle the complexity of dietary and biomarker measurements
  
  - After measurement error correction Vit-B6 and folate intakes were inversely associated with the risk of kidney cancer

- Aspects still to tune up
  
  - Great potential to use data from studies with available *concentration* biomarker data
Concluding remarks (ii)

- Are Bayesian modelling worth the trouble in nutritional epidemiology?
  - They are flexible and can be very informative

- Need of informative data, possibly replicates of R (and M) measurements
  - Need of more biomarker measurements of dietary exposure
Acknowledgments

- Pietro Ferrari, Hannah Lennon, Martyn Plummer
- WCRF/AICR for funding the project
- Mattias Johansson, Veronique Chajes
- Ray Carroll, Paul Gustafson, Victor Kipnis, Heather Bowles
- EPIC collaborators