Bayesian methods in Biomedicine

David Ríos Insua & Simón Rodríguez Santana
ICMAT-CSIC
ICMAT Datalab

About us…

- **ICMAT** - Severo Ochoa Excellence Award (3 times)
- **Datalab group** ([https://www.datalab.icmat.es](https://www.datalab.icmat.es)) AXA-ICMAT Chair since 2014
- Framework projects since 2014

Collaborations:

- ITEFI, IMF, CIB, Cajal, CNB, IFS, IIIA, I2SysBio,…

Resources + others:

- Open courses: “Intro to ML” and “Bayesian Data Science”
  → [https://datalab-icmat.github.io/courses_stats.html](https://datalab-icmat.github.io/courses_stats.html)
- HPC (Lovelace) + UAM computational resources (CCC)
Bayes... What else?

35+ years experience in Bayesian inference and decision analysis

Now, Bayesian ML or Bayesian Data Science

1. Better apportioning of uncertainty sources, including prior (and adversarial info)
2. Predictions based on predictive distributions
3. Coherent integration within a decision making framework
4. More robust inferences and decisions (even in hostile environments)

Complex applications (including biomed) motivating new methodology in Bayesian inference & DA
Vignette 1: Apportioning uncertainty (CVDs)
Vignette 1: CVDs (apportioning uncertainty)

Some context (in Europe):

- CVD are leading death cause
  - 3.9 millions deaths per year
  - 45% of all deaths
- Annual CVD treatment > 210 billion €

**IMPORTANT**: CV risk prediction for CVD management and control. 90% lifestyle induced
Approach and objectives

Bayesian network (BN) implementation

- Health assistance company annual checks (2012-2016)
- Mod. and non-mod. CVRFs and medical conditions + census information
- Large dataset (>200K) + expert knowledge to build underlying model

BN model provides:

- **Interpretable inference and prediction on CVRFs**
- **Decision-support tool** to suggest diagnosis, treatment, policy, and research actions
Bayesian network

Learned network including expert modifications

15 edges added and 7 reversed

Probability tables estimated from Multinomial-Dirichlet models

- Predictions: posterior means
- Hypothesis tests: complete distribution

\[
p(v_1, \ldots, v_{13}) = [p(v_1)p(v_2)p(v_3 | v_1, v_8)p(v_4 | v_1, v_2, v_3, v_5, v_6, v_8)] \times [p(v_5 | v_2, v_6, v_8)p(v_6 | v_1, v_2, v_7, v_8)p(v_7 | v_2) \times [p(v_8 | v_1, v_2)p(v_9 | v_1, v_7, v_{10}, v_{11})p(v_{10} | v_1, v_3)] \times [p(v_{11} | v_5, v_6, v_7)p(v_{12} | v_1, v_2, v_3, v_6, v_7, v_8) \times [p(v_{13} | v_1, v_2, v_6)]. \]
Individual:
Sex=Male, Age=(44,54], Edu=3, SeC=3, BMI=Obese, PA=Inactive, Smok=F, SD=Short, Anxiety=Yes, Depression=No

- Prob. of developing Hypertension = 45.63%

<table>
<thead>
<tr>
<th>MCVRF</th>
<th>Level</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI</td>
<td>Normal</td>
<td>11.30</td>
</tr>
<tr>
<td>Physical activity</td>
<td>Regularly active</td>
<td>34.57</td>
</tr>
<tr>
<td>Sleep</td>
<td>Normal</td>
<td>39.69</td>
</tr>
<tr>
<td>Anxiety</td>
<td>No</td>
<td>37.02</td>
</tr>
</tbody>
</table>

- Priority should be to **improve BMI**.
- If all the MCRF are improved, prob. decreases to **4.80%**
Software

**GeNie model** (Academic use) [https://datalab-icmat.github.io/software.html](https://datalab-icmat.github.io/software.html)
Vignette 2: DIGIVERT (predictive distributions)
Vignette 2: DIGIVERT (predictive distributions)

Molecular design problem

- Designing new molecules is **time** and **resource intensive**
- *Old approach*: Expert proposal + synthesis + measure candidates *in vitro*
- *Soon-to-be-old way*: High throughput virtual screening (HTVS)

**AI assisted de-novo design** → *Automatically proposing novel chemical structures that optimally satisfy desired properties*
**De-novo molecular design**

Many (*many*) models:
- $10^2$ (maybe $10^3$)
- Expertise important to navigate them

Partial picture:
http://www.vls3d.com/index.php

**Neural-network based:**
Super popular, can be hard to use
- **Bayesian optimization**
- New promising compounds
Vignette 3: AMLARA (robustness)
Vignette 2: AMLARA (robustness)

ML meets security:

Dermatoscopic image of a benign melanocytic nevus, along with the diagnostic probability computed by a deep neural network. Perturbation computed by a common adversarial attack technique. See (7) for details.

Combined image of nevus and attack perturbation and the diagnostic probabilities from the same deep neural network.

Source: Finlayson et al. (2019)

Gallego, DRI (2022), Rodríguez Santana et al. (2022)
Bayesian robust image classification

Adversarial Risk Analysis framework

![Graph showing accuracy over attacked test set against intensity of PGD attack (l_1) for different methods including NONE, AT, ALP, and ARA.](image-url)
Vignette 4: ONCOSCREEN CRC (coherent dec.)
Vignette 4: ONCOSCREEN CRC (coherent dec.)

ONCOSCREEN

- **EU Mission: Cancer** (2023-2026)
- 39 partners (Med devices, CRC specialists, AI-IT, Health Econ, Insur., Regulators, Patient Assoc,CROs,...) covering the whole CRC-value chain
- More info: [https://oncoscreen.health/](https://oncoscreen.health/)

CRC:

- **12.4% of cancer deaths**
- Only 14% in EU participate in screening programs (colonoscopy not so nice)
- 4 new screening devices promised and how to incentivise them

Currently on its first steps...
1- Predictive model
2- Decision model
3- Incentive model
Thank you!

Queries or questions, reach out!

Collabs welcome!

✉ david.rios@icmat.es
✉ simon.rodriguez@icmat.es

ICMAT Datalab https://www.datalab.icmat.es
Data and preprocessing

- Annual health (2012 - 2016) assessments from health assistance company
- Complemented with *census information*, inferring:
  - Socioeconomic status
  - Educational level
- Removal of outliers, duplicates, misrecorded and missing values
- Retain the most recent assessment of each individual

$\Rightarrow$ Final dataset contains **205,087 health assessments**
## Relevant variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_1$</td>
<td>Sex</td>
<td>{female, male}</td>
</tr>
<tr>
<td>$v_2$</td>
<td>Age</td>
<td>[24,34], [34,44], [44,54], [54,64], [64,74]</td>
</tr>
<tr>
<td>$v_3$</td>
<td>Education level</td>
<td>{1,2,3}</td>
</tr>
<tr>
<td>$v_4$</td>
<td>Socioeconomic status</td>
<td>{1,2,3}</td>
</tr>
<tr>
<td>$v_5$</td>
<td>Body mass index</td>
<td>{underw., normal, overw., obese}</td>
</tr>
<tr>
<td>$v_6$</td>
<td>Physical activity</td>
<td>{insufficiently active (1), regularly active (2)}</td>
</tr>
<tr>
<td>$v_7$</td>
<td>Sleep duration</td>
<td>{short, normal, excessive}</td>
</tr>
<tr>
<td>$v_8$</td>
<td>Smoker profile</td>
<td>{non-smoker, ex-smoker, smoker}</td>
</tr>
<tr>
<td>$v_9$</td>
<td>Anxiety</td>
<td>{yes, no}</td>
</tr>
<tr>
<td>$v_{10}$</td>
<td>Depression</td>
<td>{yes, no}</td>
</tr>
<tr>
<td>$v_{11}$</td>
<td>Hypertension</td>
<td>{yes, no}</td>
</tr>
<tr>
<td>$v_{12}$</td>
<td>Hypercholesterolemia</td>
<td>{yes, no}</td>
</tr>
<tr>
<td>$v_{13}$</td>
<td>Diabetes</td>
<td>{yes, no}</td>
</tr>
</tbody>
</table>

CVRFs = Cardiovascular risk factors
Diagnosis and evidence propagation

Example: Individual/ set of individuals with \( \text{Age} \geq 45, \text{BMI} = \text{Overweight}, \text{SD} \geq 6, \text{Anxiety} = \text{Yes} \)

\[
\Pr(v_{11} = y \mid v_1 = \text{male}, v_2 \geq 45, v_5 = \text{overw.}, v_6 = 1, v_7 = \leq 6h), v_9 = y)
\]

\[
= \frac{\Pr(v_1 = \text{male}, v_2 \geq 45, v_5 = \text{overw.}, v_6 = 1, v_7 = \leq 6h), v_9 = y, v_{11} = y)}{\Pr(v_1 = \text{male}, v_2 \geq 45, v_5 = \text{overw.}, v_6 = 1, v_7 = \leq 6h), v_9 = y)}
\]

\[
= 25.26 \% > 15.05\% \text{ (marginal probability)}
\]

- Individual should be **informed** of a high probability of having hypertension.

<table>
<thead>
<tr>
<th>BMI</th>
<th>Physical activity</th>
<th>Probability Male</th>
<th>Probability Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overw.</td>
<td>1</td>
<td>25.26</td>
<td>26.34</td>
</tr>
<tr>
<td>Overw.</td>
<td>2</td>
<td>19.79</td>
<td>20.70</td>
</tr>
<tr>
<td>Obese</td>
<td>1</td>
<td>45.54</td>
<td>46.95</td>
</tr>
<tr>
<td>Obese</td>
<td>2</td>
<td>34.49</td>
<td>35.78</td>
</tr>
<tr>
<td>Overw., obese</td>
<td>1</td>
<td>32.85</td>
<td>33.82</td>
</tr>
<tr>
<td>Overw., obese</td>
<td>2</td>
<td>22.90</td>
<td>23.85</td>
</tr>
</tbody>
</table>

Positive impact of PA
Limitations of the study

- The dataset has different structure to Spanish population *(healthy worker effect)*
- Some data were self-reported
- No explicit data concerning diet (except for alcohol)
- (...*in the end*) Only predictive claims, not causal
De-novo molecular design

Many (many) models:
- $10^2$ (maybe $10^3$)
- Expertise important to navigate them

Partial picture:
http://www.vls3d.com/index.php

Neural-network based:
Super popular, can be hard to use
- Bayesian optimization
- New promising compounds

Meaningful exploration + optimized property maximization search

Average distance between ZINC molecules latent space (19.66)