Bayesian methods in Biomedicine

David Ríos Insua & Simón Rodríguez Santana
ICMAT-CSIC

ICMAT
INSTITUTO DE CIENCIAS MATEMÁTICAS

ICMAT Datalab

About us…

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

Collaborations:

- ITEFI, IMF, CIB, Cajal, CNB, IFS, IIIA, I2SysBio,…
- AIHUB (https://aihub.csic.es/)

Resources + others:

- Open courses: “*Intro to ML*” and “*Bayesian Data Science*”

  ⟶https://datalab-icmat.github.io/**courses_stats**.html

- HPC (Lovelace) + UAM computational resources (CCC)

ICMAT
DataLab

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

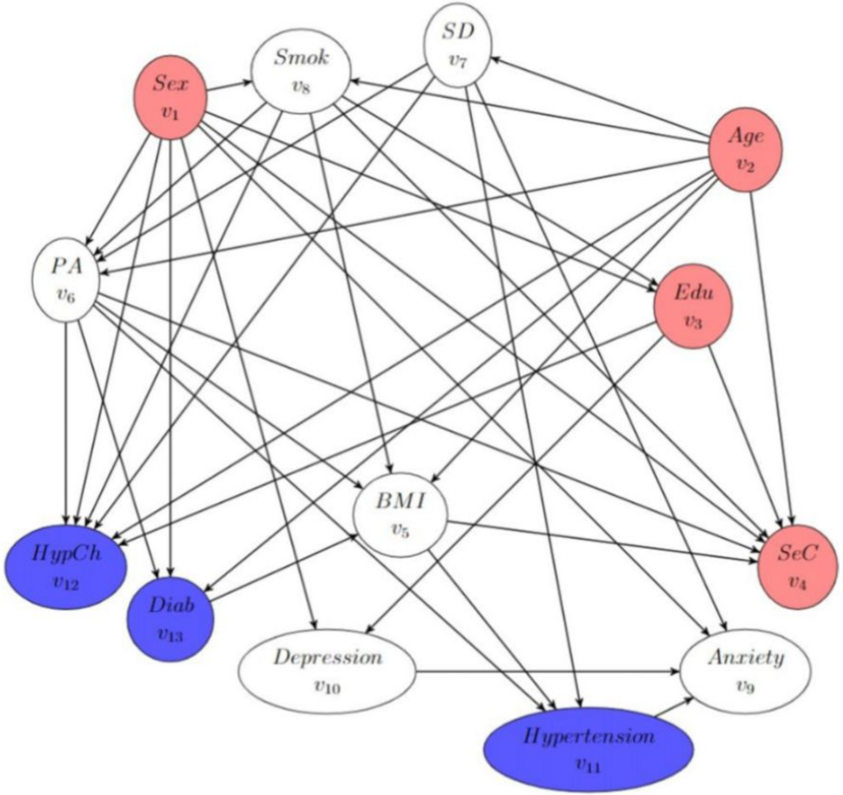
Bayesian network (BN) implementation

- 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

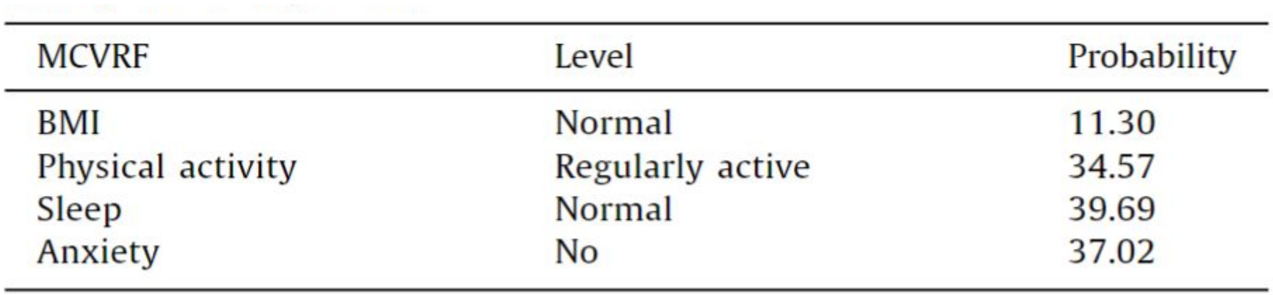
- Predictions: posterior means
- Hypothesis tests: complete distribution

$$\begin{aligned} p(v_1, \ldots, v_{13}) = {} & [p(v_1)p(v_2)p(v_3 \mid v_1, v_8)p \\ & (v_4 \mid v_1, v_2, v_3, v_5, v_6, v_8)] \\ & \times[p(v_5 \mid v_2, v_6, v_8)p(v_6 \mid v_1, v_2, v_7, v_8)p(v_7 \mid v_2) \\ & p(v_8 \mid v_1, v_2)p(v_9 \mid v_1, v_7, v_{10}, v_{11})p(v_{10} \mid v_1, v_3)] \\ & \times[p(v_{11} \mid v_5, v_6, v_7)p(v_{12} \mid v_1, v_2, v_3, v_6, v_7, v_8) \\ & p(v_{13} \mid v_1, v_2, v_6)], \end{aligned}$$

Therapies through influential findings

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%**

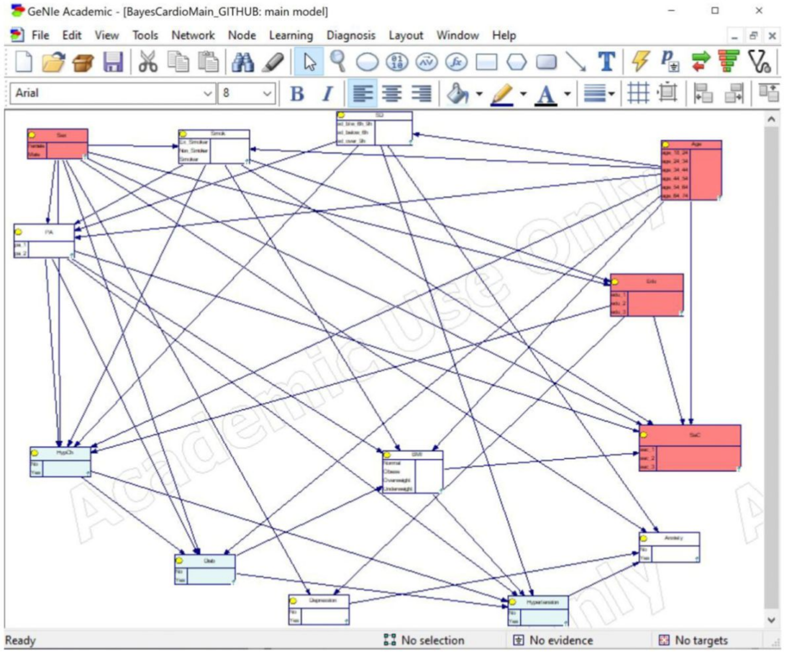


MCVRF	Level	Probability
BMI	Normal	11.30
Physical activity	Regularly active	34.57
Sleep	Normal	39.69
Anxiety	No	37.02

- 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



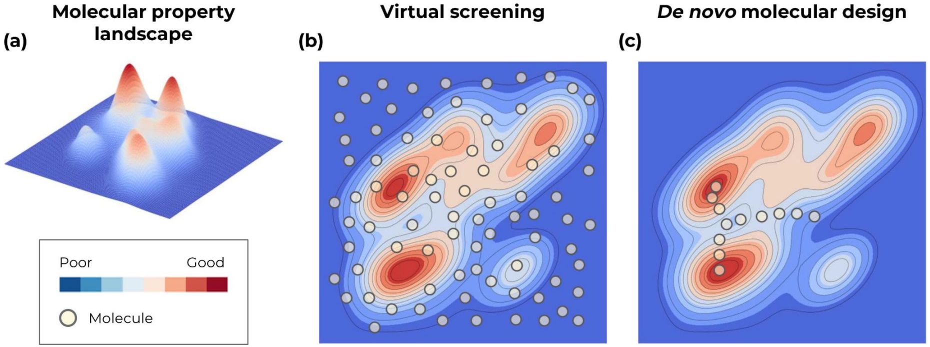
Vignette 2: DIGIVERT (predictive distributions)

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

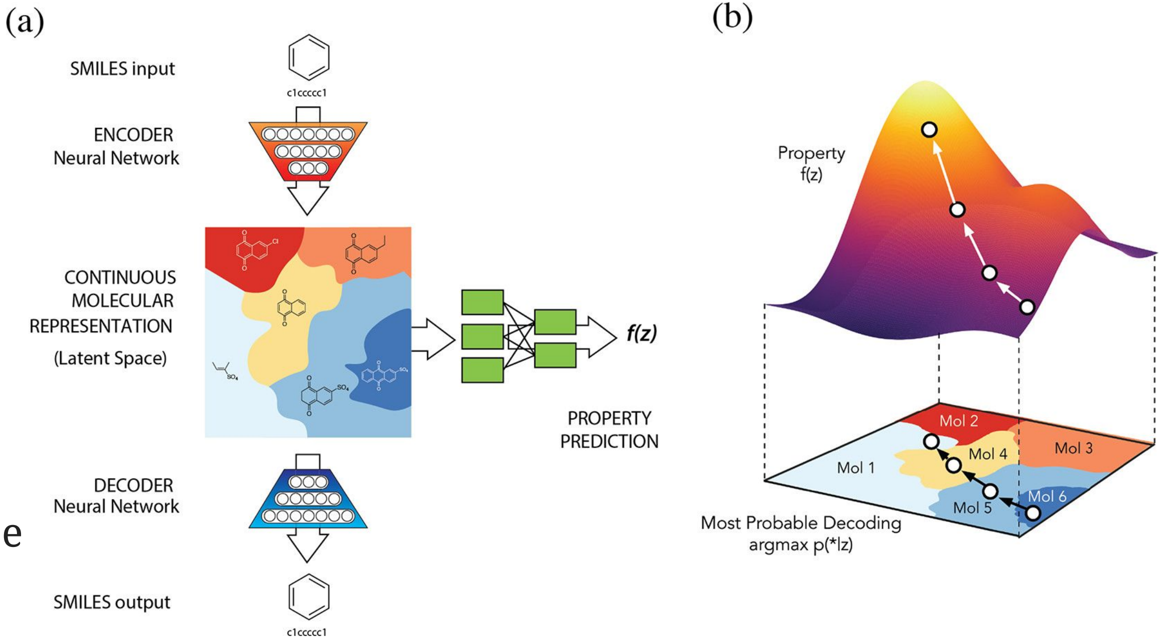
Partial picture:

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

Neural-network based:

Super popular, can be hard to use

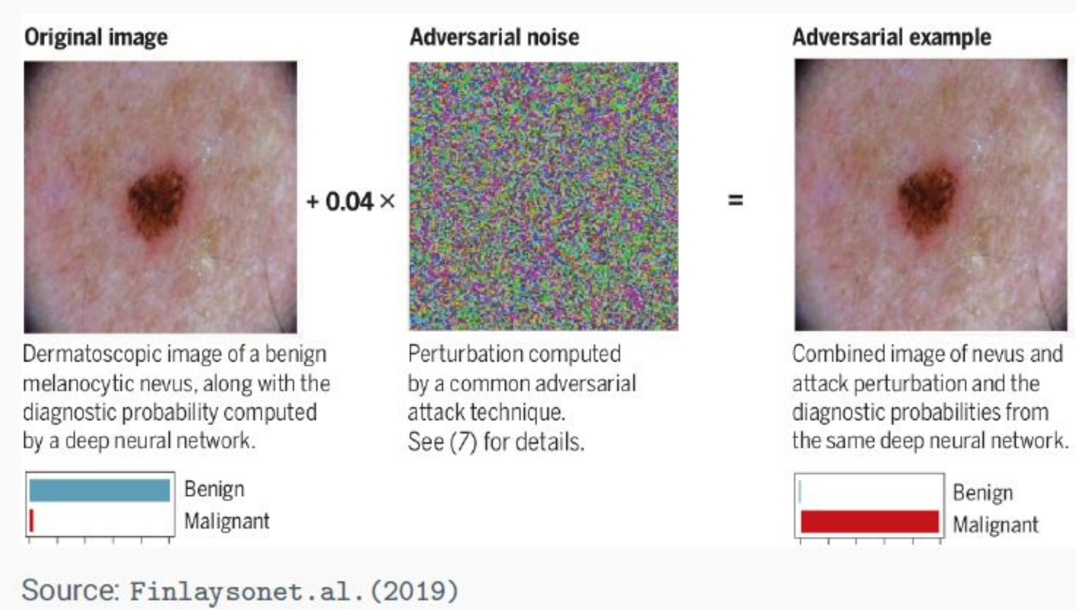
- **Bayesian optimization**
- New promising compounds



Vignette 3: AMLARA (robustness)

Vignette 2: AMLARA (robustness)

ML meets security:



Original image

Dermatoscopic image of a benign melanocytic nevus, along with the diagnostic probability computed by a deep neural network.

Adversarial noise

Perturbation computed by a common adversarial attack technique. See (7) for details.

Adversarial example

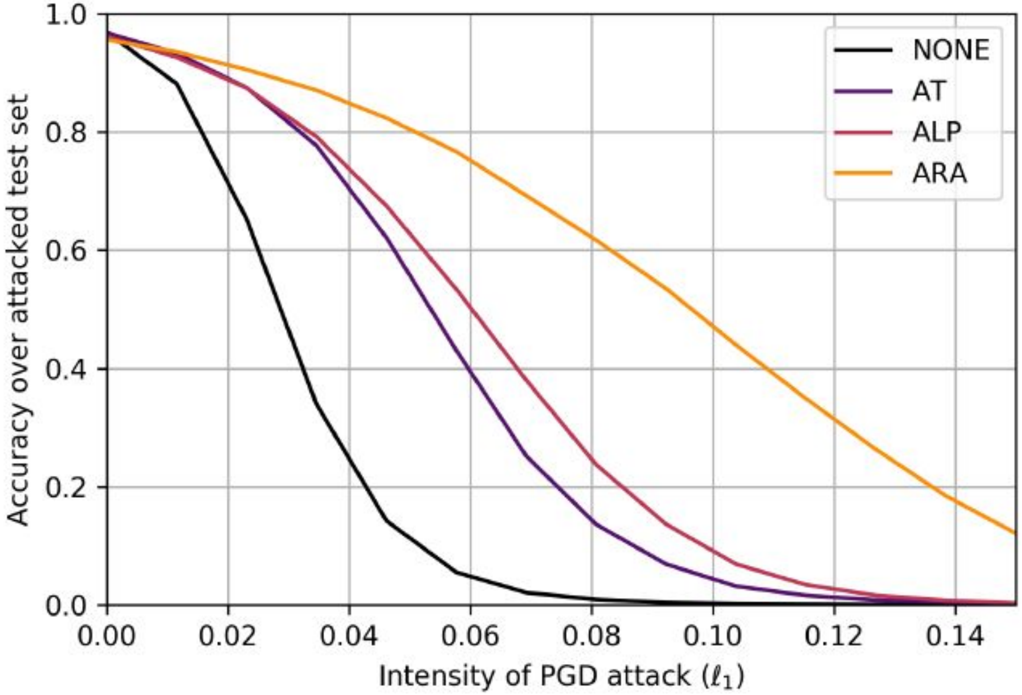
Combined image of nevus and attack perturbation and the diagnostic probabilities from the same deep neural network.

Source: Finlaysonet.al.(2019)

Gallego, DRI (2022), Rodríguez Santana et al. (2022)

Bayesian robust image classification

Adversarial Risk Analysis framework



Vignette 4: ONCOSCREEN CRC (coherent dec.)

Vignette 4: ONCOSCREEN CRC (coherent dec.)

ONCOSCREEN

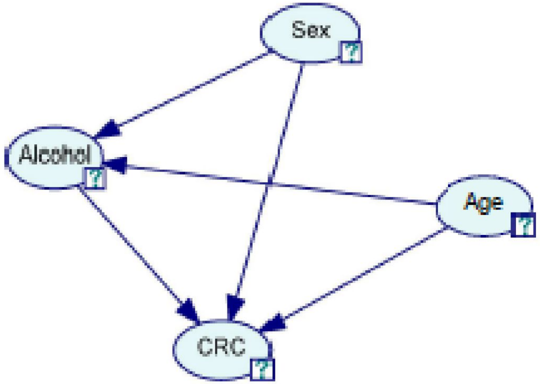
- **EU Mission: Cancer** (2023-2026)
- 39 partners (Med devices, CRC specialists, AI-IT, Health Econ, Insur., Regulators, Patient Assoc,CROs,…) covering the whole CRC-value chain
- More info: https://oncoscreen.health/

CRC:

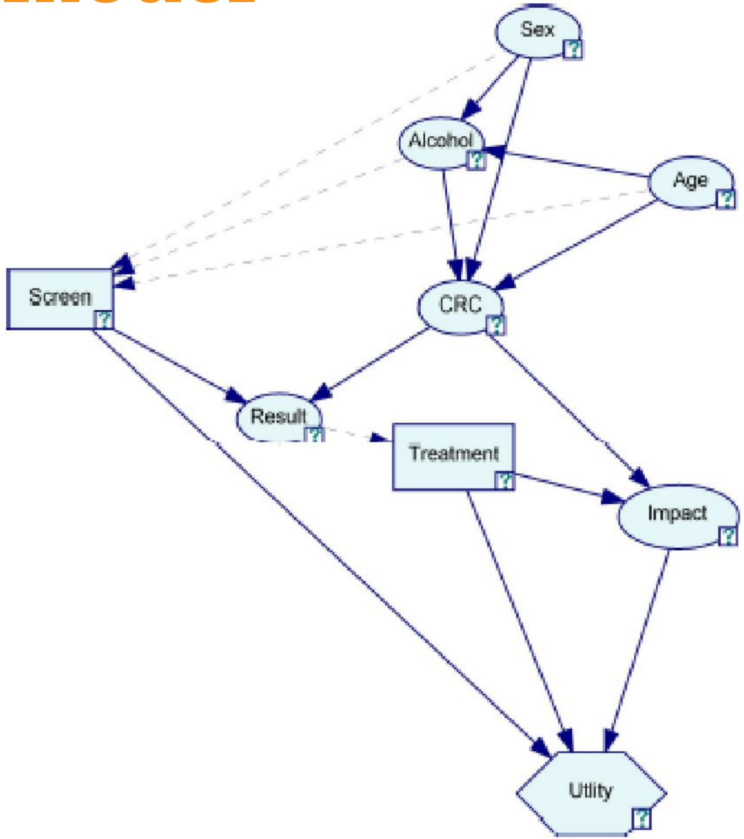
- **12.4% of cancer deaths**
- Only 14% in EU participate in screening programs (colonoscopy not so nice)
- 4 new screening devices promised and how to incentivise them

Currently on its first steps…

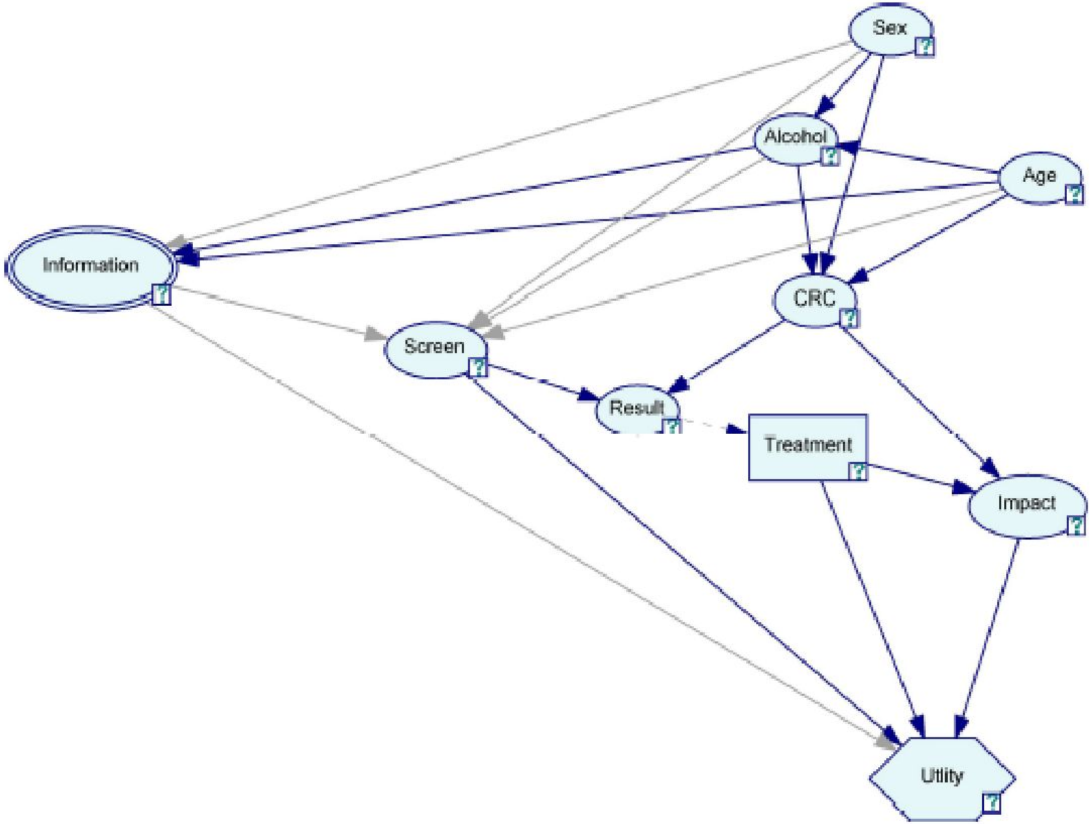
1- Predictive model



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

Extra

- Annual health (2012 - 2016) assessments from health assistance company

- Complemented with *census information*, inferring:
  - Socioeconomic status
  - Educational level

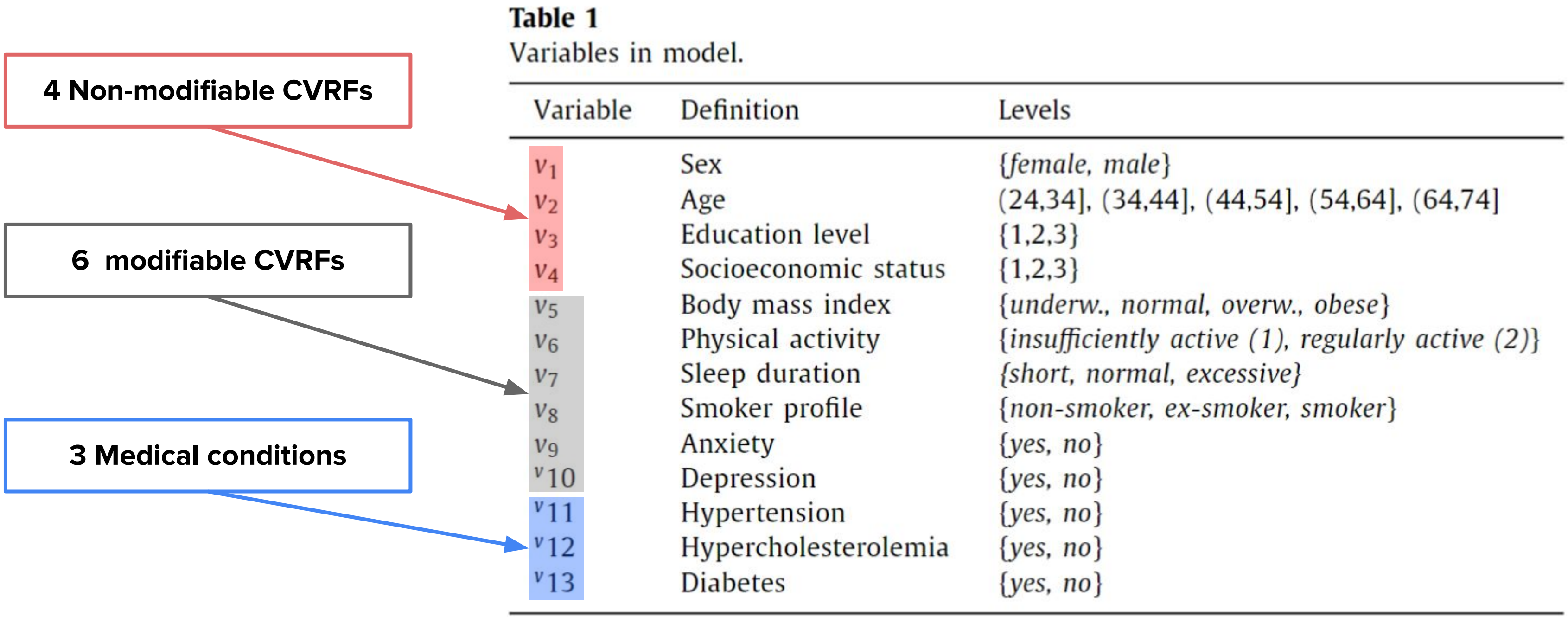
- Removal of outliers, duplicates, misrecorded and missing values

- Retain the most recent assessment of each individual

⇨ Final dataset contains **205,087 health assessments**

Relevant variables

Extra



4 Non-modifiable CVRFs

6 modifiable CVRFs

3 Medical conditions

Table 1
Variables in model.

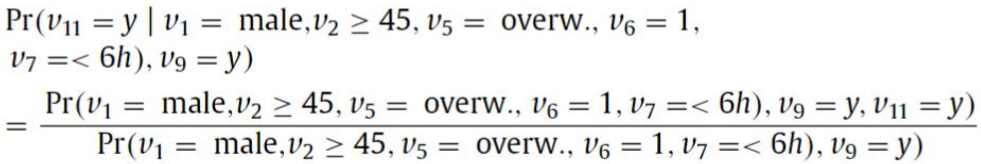
Variable	Definition	Levels
v_1	Sex	{*female, male*}
v_2	Age	(24,34], (34,44], (44,54], (54,64], (64,74]
v_3	Education level	{1,2,3}
v_4	Socioeconomic status	{1,2,3}
v_5	Body mass index	{*underw., normal, overw., obese*}
v_6	Physical activity	{*insufficiently active (1), regularly active (2)*}
v_7	Sleep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

CVRFs = Cardiovascular risk factors

Diagnosis and evidence propagation

Extra

Example: Individual/ set of individuals with `Age≥45,BMI=Overweight,SD≥6,Anxiety=Yes`



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

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

= 25.26 % > 15.05% (marginal probability)

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

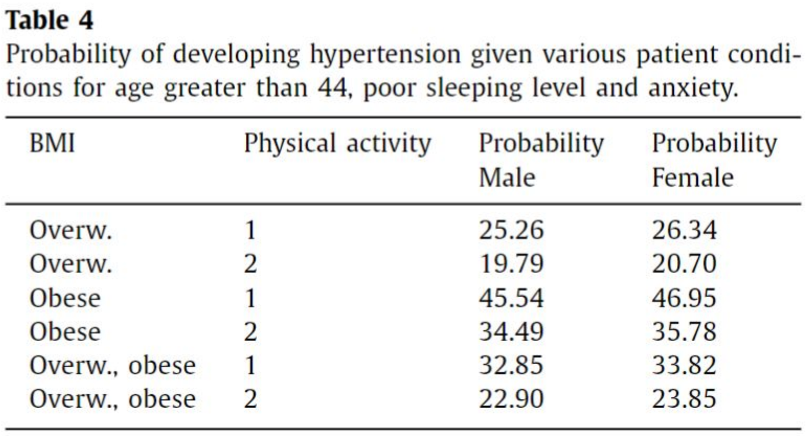


Table 4
Probability of developing hypertension given various patient conditions for age greater than 44, poor sleeping level and anxiety.

BMI	Physical activity	Probability Male	Probability Female
Overw.	1	25.26	26.34
Overw.	2	19.79	20.70
Obese	1	45.54	46.95
Obese	2	34.49	35.78
Overw., obese	1	32.85	33.82
Overw., obese	2	22.90	23.85

Positive impact of PA

Limitations of the study

Extra

- 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:

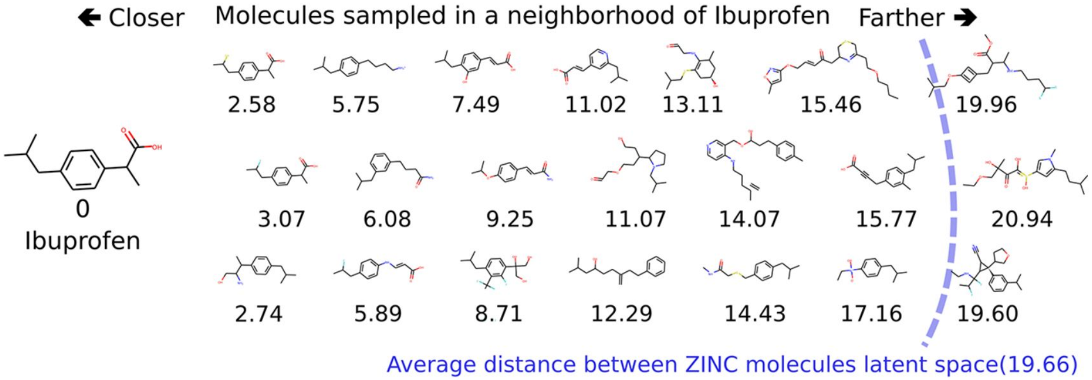
Super popular, can be hard to use

- **Bayesian optimization**
- New promising compounds

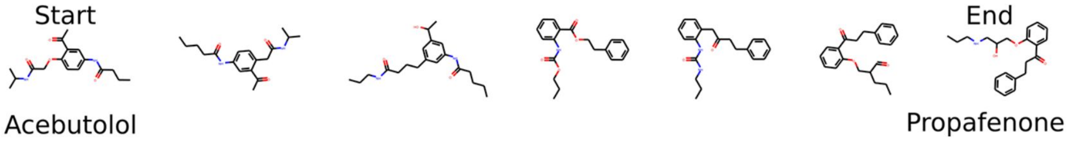
Meaningful exploration

+

optimized property maximization search



← Closer
Molecules sampled in a neighborhood of Ibuprofen
Farther →
0
Ibuprofen
2.58
5.75
7.49
11.02
13.11
15.46
19.96
3.07
6.08
9.25
11.07
14.07
15.77
20.94
2.74
5.89
8.71
12.29
14.43
17.16
19.60
Average distance between ZINC molecules latent space(19.66)



Start
Acebutolol
End
Propafenone