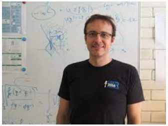
PreMeDICaL: Precision Medicine by Data Integration and Causal Learning

Inria Sophia Antipolis - Méditérannée, Antenne de Montpellier.

Institut Desbrest d'Épidémiologie et de Santé Publique (IDESP): UMR UA11 Inserm - Université de Montpellier (UM). Focus on epidemiology of chronic non communicable diseases.

Composition of the team

- Julie Josse (PI): Advanced researcher, Inria. Topics: Dimensionality reduction, matrix completion, causal inference, R statistical software
- Pascal Demoly: Director of Idesp. Respiratory physician, allergist, professor of pulmonology at the University Hospital, head of department
- Pierre Lafaye de Micheaux: Assistant professor (UPVM3). Topics: Measures of dependences, medical images, R statistical software
- Nicolas Molinari: Co-director of Idesp. Professor in biostatistics at the University Hospital, head of the statistics department
- 3 PhD students (co-supervised); 1 postdoc, 1 engineer (UM grant)
- Non permanent members: François Husson (Pr), 3 post-doc, 2 PhD



⇒ **Interdisciplinary** team with clinical, bio-stat & machine learning (ML) skills

Application context: respiratory allergy

- ▷ **Asthma**: chronic inflammatory disease of the bronchi which evolves by crisis, alters the respiratory system and may engage the vital prognosis
- ▷ **Large variability** in its manifestations:
  - interaction between the genetic background and the environment
  - association with other allergic diseases (like rhinitis, sleep issues)
- ▷ Due to environmental (air quality, temperature, biodiversity) & lifestyles (diets) changes, WHO in 2050, 1/2 person with allergies

- ▷ **Sources of information**: biological, clinical, environmental, etc.
- ▷ Underexploited. Today: **data collected, new tools for data fusion**

- ▷ Provide new knowledge (in terms of disease heterogeneity) that may change guidelines and practice
- ▷ New opportunities for new diagnostics and therapeutics, **design personalized solutions**, improving patient care and prevention

Locks

Data integration comes with methodological challenges

- ▷ *heterogeneous data*:
  - ◇ for a patient, different nature of data (clinical, images, bio)
  - ◇ for a pathology, data from different hospitals
  - ◇ experimental (trials) & non-experimental (observational) data
- ▷ *missing data:* different types (informative), patterns (systematic)

⇒ State-of-the-art ML/causal inference can not handle high dim. multi-sources data with distributional shifts & missing data

Locks

Data integration comes with methodological challenges

- ▷ *heterogeneous data*:
  - ◇ for a patient, different nature of data (clinical, images, bio)
  - ◇ for a pathology, data from different hospitals
  - ◇ experimental (trials) & non-experimental (observational) data
- ▷ *missing data:* different types (informative), patterns (systematic)

⇒ State-of-the-art ML/causal inference can not handle high dim. multi-sources data with distributional shifts & missing data

Gap between what is develop and what is used

- ▷ superiority of ML methods to parametric methods?
- ▷ lack of confidence: lack of uncertainty quantification, reproducibility & training
- ▷ lack of involvement of all stakeholders

⇒ Few research translated into clinically actionable solutions

PreMeDICaL research axes

Personalized medicine by optimal prescription of treatment

▷ causal inference techniques for (dynamic) policy learning

⇒ who to treat and when

▷ leverage both randomized control trials (RCTs) and observational data

⇒ launch a drug without running RCTs ?

⇒ rethink evidence needed to bring treatments to the market faster

Personalized medicine by integration of different data sources

▷ relevance of each data source from different scales

▷ solutions to handle missing values: complex structure of missing values, prediction with uncertainties

⇒ Push methodological innovation up to patients, clinicians, regulators
⇒ Collaborative effort: leveraging ML, data, clinical expertise and existing recommendations

Research axis 1:
Precision medecine by optimal prescription of treatment

Potential Outcome framework (Neyman, 1923, Rubin, 1974)

Causal effect for a binary treatment

▷ n i.i.d. obs $\Big(\underbrace{X_i}_{\text{covariates}} , \overbrace{W_i}^{\text{treatment}} , \underbrace{Y_i(1), Y_i(0)}_{\textit{potential outcomes}} \Big) \in \mathbb{R}^d \times \{0, 1\} \times \mathbb{R} \times \mathbb{R}$

▷ Individual causal effect of the treatment: $\Delta_i \triangleq Y_i(1) - Y_i(0)$

Missing problem: Δ_i never observed (only observe one outcome/indiv)

Covariates			Treatment	Outcome(s)	
X_1	X_2	X_3	W	Y(0)	Y(1)
1.1	20	F	1	**?**	200
-6	45	F	0	10	**?**
0	15	M	1	**?**	150
	. . .		. . .	. . .	. . .
-2	52	M	0	100	**?**

Cov.			Treat.	Out.
X_1	X_2	X_3	W	Y
1.1	20	F	1	200
-6	45	F	0	10
0	15	M	1	150
	. . .		. . .	. . .
-2	52	M	0	100

Potential Outcome framework (Neyman, 1923, Rubin, 1974)

Causal effect for a binary treatment

- n i.i.d. obs $(\underbrace{X_i}_{\text{covariates}} , \overbrace{W_i}^{\text{treatment}} , \underbrace{Y_i(1), Y_i(0)}_{\textit{potential outcomes}}) \in \mathbb{R}^d \times \{0, 1\} \times \mathbb{R} \times \mathbb{R}$
- Individual causal effect of the treatment: $\Delta_i \triangleq Y_i(1) - Y_i(0)$

Missing problem: Δ_i never observed (only observe one outcome/indiv)

Covariates			Treatment	Outcome(s)	
X_1	X_2	X_3	W	Y(0)	Y(1)
1.1	20	F	1	**?**	200
-6	45	F	0	10	**?**
0	15	M	1	**?**	150
	...		...	...	...
-2	52	M	0	100	**?**

Cov.			Treat.	Out.
X_1	X_2	X_3	W	Y
1.1	20	F	1	200
-6	45	F	0	10
0	15	M	1	150
	...		...	...
-2	52	M	0	100

Average Treatment Effect (ATE): $\tau \triangleq \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$

The ATE is the difference of the average outcome had everyone gotten treated and the average outcome had nobody gotten treatment

Data to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation)
- covariate distributions of treated and control groups are balanced
  ⇒ High **internal** validity

- expensive, long, ethical limitations
- small sample size: restrictive inclusion criteria
  ⇒ No personalized medicine
- trial sample different from the population eligible for treatment
  ⇒ Low **external** validity

Data to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation 🪙)
- covariate distributions of treated and control groups are balanced
  ⇒ High **internal** validity
- expensive, long, ethical limitations
- small sample size: restrictive inclusion criteria
  ⇒ No personalized medicine
- trial sample different from the population eligible for treatment
  ⇒ Low **external** validity

Observational data

- low cost
- large amounts of data (registries, biobanks, EHR, claims)
  ⇒ patient's heterogeneity
- representative of the target populations
  ⇒ High **external** validity

Data to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation 🪙)
- covariate distributions of treated and control groups are balanced
  ⇒ High **internal** validity

- expensive, long, ethical limitations
- small sample size: restrictive inclusion criteria
  ⇒ No personalized medicine
- trial sample different from the population eligible for treatment
  ⇒ Low **external** validity

Observational data

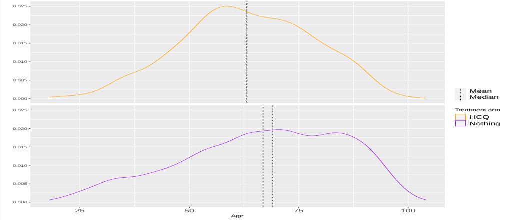
- "big data": low quality
- lack of a controlled design opens the door to **confounding bias**
  ⇒ Low **internal** validity

- low cost
- large amounts of data (registries, biobanks, EHR, claims)
  ⇒ patient's heterogeneity
- representative of the target populations
  ⇒ High **external** validity

Observational data: non random assignment

	survived	deceased	Proportion(survived $\mid$ treatment)	Pr(deceased $\mid$ treatment)
HCQ	497 (11.4%)	111 (2.6%)	0.817	0.183
HCQ+AZI	158 (3.6%)	54 (1.2%)	0.745	0.255
none	2699 (62.1%)	830 (19.1%)	0.765	0.235

Mortality rate 22.9% - for HCQ 18.3% - non treated 23.5%: treatment helps?



Comparison of the distribution of Age between HCQ and non treated.

Younger patients (with lower risk of death) are more likely to be treated.
If control group does not look like treatment group, difference in response may be **confounded** by differences between the groups.

⇒ **Unconfoundness** identifiability assumption: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i$.

Leverage both RCT and observational data

RCT

– Narrowly defined population

+ High **internal** validity

Observational data

– **Confounding**

+ High **external** validity

We could use both to [1] . . .

- ▷ . . . validate observational methods
- ▷ . . . correct confounding bias
- ▷ . . . improve estimation of heterogeneous treatment effects
- ▷ . . . **generalize the Average Treatment Effect to a (broader) target population** (data fusion, transportability, data integration)[2]

[1] Colnet, J.J. et al. (2020). Causal inference methods for combining RCT and observational studies: a review. *Statistical Science*.

[2] Elias Bareinboim & Judea Pearl. (2016). Causal inference & the data-fusion problem. *PNAS*.

Leverage both RCT and observational data

RCT

– Narrowly defined population

+ High **internal** validity

Observational data

– **Confounding**

+ High **external** validity

We could use both to [1] . . .

▷ . . . validate observational methods

▷ . . . correct confounding bias

▷ . . . improve estimation of heterogeneous treatment effects

▷ . . . **generalize the Average Treatment Effect to a (broader) target population** (data fusion, transportability, data integration)[2]

The FDA has greenlighted the usage of the drug palbociclib to men with breast cancer, though clinical trials were performed only on women

→ Reduce drug approval times and costs for patients who could benefit

[1] Colnet, J.J. et al. (2020). Causal inference methods for combining RCT and observational studies: a review. *Statistical Science*.

[2] Elias Bareinboim & Judea Pearl. (2016). Causal inference & the data-fusion problem. *PNAS*.

Generalization task

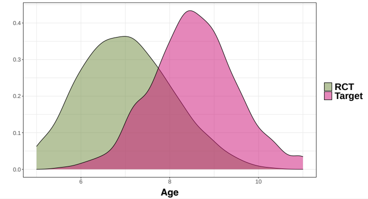
	S	X_1	X_2	X_3	W	Y
1	1	1.1	20	5.4	1	24.1
. . .	1		. . .		. . .	. . .
$n-1$	1	-6	45	8.3	0	26.3
n	1	0	15	6.2	1	23.5
$n+1$	0	-2	52	7.1	NA	NA
$n+2$	1	-1	35	2.4	NA	NA
. . .	0		. . .		NA	NA
$n+m$	1	-2	22	3.4	NA	NA

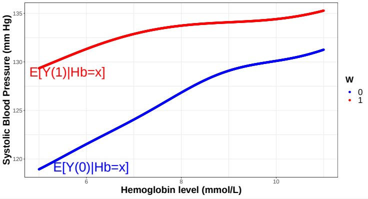
Available data with observed treatment and outcome only in the RCT.

- ▷ S indicator of eligibility for the trial
- ▷ covariates distribution not the same in the in the RCT & target pop:

$$f_{X|S=1} \neq f_X$$

$$\Rightarrow \quad \underbrace{\tau_1 = \mathbb{E}[Y(1) - Y(0)|S=1]}_{\text{ATE in the RCT}} \neq \underbrace{\mathbb{E}[Y(1) - Y(0)] = \tau}_{\text{Target ATE}}.$$





Estimators of the average treatment effect by generalization

	S	X_1	X_2	X_3	W	Y
1	1	1.1	20	5.4	1	24.1
...	1		...		...	...
$n-1$	1	-6	45	8.3	0	26.3
n	1	0	15	6.2	1	23.5
$n+1$	0	-2	52	7.1	NA	NA
$n+2$	1	-1	35	2.4	NA	NA
...	0		...		NA	NA
$n+m$	1	-2	22	3.4	NA	NA

Available data with observed treatment and outcome only in the RCT.

- ▷ **weight the RCT sample** so that it ressembles the target pop (IPSW)
- ▷ model the conditional outcomes & **extrapolate** to the target pop (gformula)
- ▷ combining the previous two ideas (**doubly robust approaches**, AIPSW)[3]

$$\hat{\tau}_{AIPW} \triangleq \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + W_i\frac{Y_i - \hat{\mu}_{(1)}(X_i)}{\hat{e}(X_i)} - (1-W_i)\frac{Y_i - \hat{\mu}_{(0)}(X_i)}{1-\hat{e}(X_i)}\right)$$

with $\mu_{(w)}(x) \triangleq \mathbb{E}[Y_i(w) \mid X_i = x]$ and $e(x) \triangleq P(W_i = 1 \mid X_i = x), \quad \forall\, x \in \mathcal{X}$.

$\Rightarrow$ $\hat{\tau}_{AIPW}$ consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent.
$\Rightarrow$ possibility to use any (machine learning) procedure such as **random forests**, neural nets, etc. to estimate $\hat{e}(x)$ and $\hat{\mu}_{(w)}(x)$.

[3] Chernozukov, Duflot, et al (2018), *Double/debiased machine learning for treatment and structural parameters. Econometrics journal.*

Exemple of projects in research axis 1

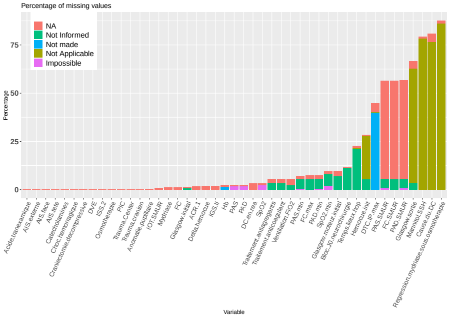
- Violation of the identifiability assumptions (sensitivity analysis)
- Missing values in causal inference
- Survival causal inference
- Policy learning (off-line reinforcement learning)

- CRO-AIT project with ALK (pharmaceutical company specializing in development of drugs for severe respiratory allergies).

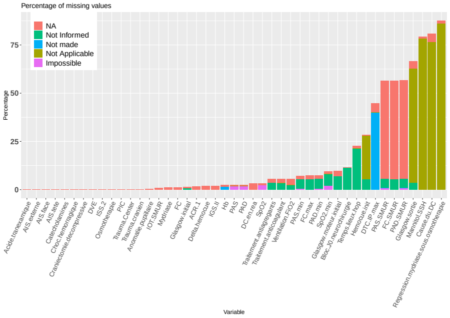
  • replacement of Inhaled Corticosteroid Therapy by 'acarizax' for dust mite allergic asthma (2 trials 600/800 patients, 1 obs data 6000 patients followed for 12-18 months, questionnaires at 3 times).
  • benefit of grazax® in prevention of asthma in children

Research axis 2:
Precision medecine by data integration

Missing data: important bottleneck in data science



Missing data: important bottleneck in data science



"One of the ironies of Big Data is that missing data play an ever more significant role" (R. Samworth, 2019)

Complete case analysis (deletion):

- Loss of information: An $n \times p$ matrix, each entry is missing with probability 0.01. $p = 5 \implies \approx 95\%$ of rows kept; $p = 300 \implies \approx 5\%$ of rows kept
- Bias: Resulting sample not representative of the target population

Due to the pandemic, a lot of patients did not perform some examination

Prediction with missing values

$\tilde{X} = X \odot (1 - M) + \mathtt{NA} \odot M$. New feature space is $\widetilde{\mathbb{R}}^d = (\mathbb{R} \cup \{\mathtt{NA}\})^d$.

$$Y = \begin{pmatrix} 4.6 \\ 7.9 \\ 8.3 \\ 4.6 \end{pmatrix} \quad \tilde{X} = \begin{pmatrix} 9.1 & \mathtt{NA} & 1 \\ 2.1 & \mathtt{NA} & 3 \\ \mathtt{NA} & 9.6 & 2 \\ \mathtt{NA} & 5.5 & 6 \end{pmatrix} \quad X = \begin{pmatrix} 9.1 & 8.5 & 1 \\ 2.1 & 3.5 & 3 \\ 6.7 & 9.6 & 2 \\ 4.2 & 5.5 & 6 \end{pmatrix} \quad M = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Find a prediction function that minimizes the risk.

$$\text{Bayes rule: } f^* \in \underset{f\colon \widetilde{\mathbb{R}}^d \to \mathbb{R}}{\arg\min}\ \mathbb{E}\left[\left(Y - f(\tilde{X})\right)^2\right].$$

$$f^*(\tilde{X}) = \mathbb{E}\left[Y \mid \tilde{X}\right] = \mathbb{E}\left[Y \mid X_{obs(M)}, M\right]$$
$$= \sum_{m \in \{0,1\}^d} \mathbb{E}\left[Y | X_{obs(m)}, M = m\right] \mathbb{1}_{M=m}$$

$\Rightarrow$ One model per pattern (2^d)

Le Morvan, J. J, E. Scornet. & G. Varoquaux. Neurips 2021, Neurips 2020 (Oral), Aistat 2020.

Prediction with missing values

$\tilde{X} = X \odot (1 - M) + \mathtt{NA} \odot M$. New feature space is $\widetilde{\mathbb{R}}^d = (\mathbb{R} \cup \{\mathtt{NA}\})^d$.

$$Y = \begin{pmatrix} 4.6 \\ 7.9 \\ 8.3 \\ 4.6 \end{pmatrix} \quad \tilde{X} = \begin{pmatrix} 9.1 & \mathtt{NA} & 1 \\ 2.1 & \mathtt{NA} & 3 \\ \mathtt{NA} & 9.6 & 2 \\ \mathtt{NA} & 5.5 & 6 \end{pmatrix} \quad X = \begin{pmatrix} 9.1 & 8.5 & 1 \\ 2.1 & 3.5 & 3 \\ 6.7 & 9.6 & 2 \\ 4.2 & 5.5 & 6 \end{pmatrix} \quad M = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Find a prediction function that minimizes the risk.

$$\text{Bayes rule: } f^* \in \underset{f\colon \widetilde{\mathbb{R}}^d \to \mathbb{R}}{\arg\min}\ \mathbb{E}\left[\left(Y - f(\tilde{X})\right)^2\right].$$

$$f^*(\tilde{X}) = \mathbb{E}\left[Y \mid \tilde{X}\right] = \mathbb{E}\left[Y \mid X_{obs(M)}, M\right]$$

$$= \sum_{m \in \{0,1\}^d} \mathbb{E}\left[Y | X_{obs(m)}, M = m\right] \ \mathbb{1}_{M=m}$$

$\Rightarrow$ One model per pattern (2^d)

Le Morvan, J. J, E. Scornet. & G. Varoquaux. Neurips 2021, Neurips 2020 (Oral), Aistat 2020.

Estimation with missing values using imputation

⇒ Incomplete data

X_1	X_2	X_3	...	Y
NA	20	10	...	shock
-6	45	NA	...	shock
0	NA	30	...	no shock
NA	32	35	...	shock
-2	NA	12	...	no shock
1	63	40	...	shock

Estimation with missing values using imputation

⇒ Incomplete data

X_1	X_2	X_3	...	Y
NA	20	10	...	shock
-6	45	NA	...	shock
0	NA	30	...	no shock
NA	32	35	...	shock
-2	NA	12	...	no shock
1	63	40	...	shock

⇒ Completed data

X_1	X_2	X_3	...	Y
3	20	10	...	shock
-6	45	6	...	shock
0	4	30	...	no shock
-4	32	35	...	shock
-2	75	12	...	no shock
1	63	40	...	shock

Estimation with missing values using imputation

⇒ Incomplete data

X_1	X_2	X_3	...	Y
NA	20	10	...	shock
-6	45	NA	...	shock
0	NA	30	...	no shock
NA	32	35	...	shock
-2	NA	12	...	no shock
1	63	40	...	shock

⇒ Completed data

X_1	X_2	X_3	...	Y
3	20	10	...	shock
-6	45	6	...	shock
0	4	30	...	no shock
-4	32	35	...	shock
-2	75	12	...	no shock
1	63	40	...	shock

A single value can't reflect the uncertainty of prediction

Multiple impute 1) Generate M plausible values for each missing value

X_1	X_2	X_3	Y
3	20	10	s
-6	45	6	s
0	4	30	no s
-4	32	35	s
-2	75	12	no s
1	63	40	s

X_1	X_2	X_3	Y
-7	20	10	s
-6	45	9	s
0	12	30	no s
13	32	35	s
-2	10	12	no s
1	63	40	s

X_1	X_2	X_3	Y
7	20	10	s
-6	45	12	s
0	-5	30	no s
2	32	35	s
-2	20	12	no s
1	63	40	s

Missing values in multi-source, multi-scale data

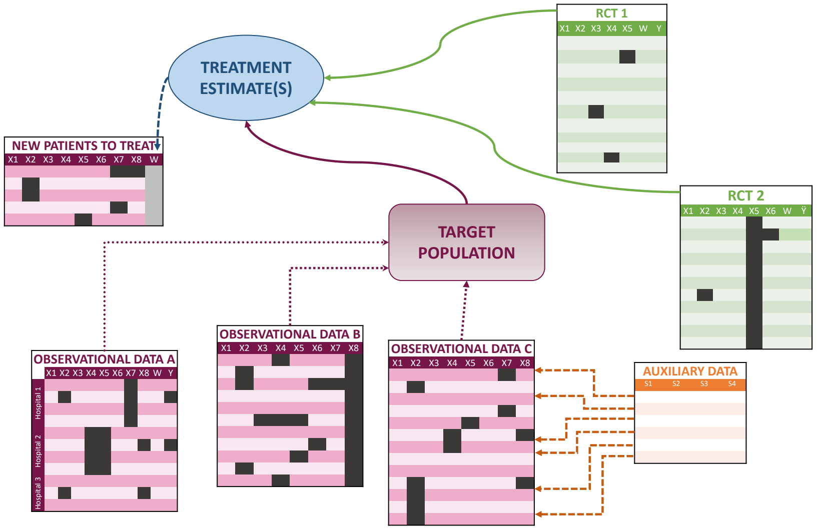
			Clinical Data					Biological Data				Questionnaire on lifestyle	
		X_1	....	X_p	W	Y	Z_1	.....	Z_q	....	C_1	...	C_r
	1		NA							....			
Obs Hospital 1			NA	NA						...			
			NA							...			
	n_1	NA	NA							...			
	1				NA	NA				...		NA	NA
Obs Hospital 2		NA		NA	NA	NA	NA	NA	NA	...			
					NA	NA				...	NA	NA	NA
	n_2				NA	NA				...			
...	...	...	...	...	...	...	...	...	...	...	...	...	...
	1	NA	NA	NA						...		NA	
Obs Hospital K		NA								...		NA	
		NA								....		NA	
	n_K	NA								....		NA	

Classical methodologies are not designed to handle high-dimensional data with selection biais and informative missing data.

Exemple of projects in research axis 2

- relationship between different sources (measure of dependencies)
- (informative) missing values in time series and structured by blocks (low rank matrix approximation)
- confidence in machine learning algorithms with missing values (conformal prediction)
- distributed computing with missing values (low rank matrix approximation+ optimal transport)

- Benralitrap project. CT air-trapping characterization for the early identification of Benralisumab responders among eosinophilic asthma patients.



TREATMENT ESTIMATE(S)

RCT 1

X1 X2 X3 X4 X5 W Y

RCT 2

X1 X2 X3 X4 X5 X6 W Ÿ

NEW PATIENTS TO TREAT

X1 X2 X3 X4 X5 X6 X7 X8 W

TARGET POPULATION

OBSERVATIONAL DATA A

X1 X2 X3 X4 X5 X6 X7 X8 W Y

Hospital 1

Hospital 2

Hospital 3

OBSERVATIONAL DATA B

X1 X2 X3 X4 X5 X6 X7 X8

OBSERVATIONAL DATA C

X1 X2 X3 X4 X5 X6 X7 X8

AUXILIARY DATA

S1 S2 S3 S4

Interdisciplinary aspects

Clinicians:

- ▷ decide relevant scientific questions
- ▷ access to patient databases (hospital, academic and industrial)
- ▷ know which methods will be accepted by the community and can lead to clinically actionable solutions
- ▷ make the links with patient associations and with state agencies
- ▷ interprete the results generated

⇒ Practice inspires theory, guide the development of methods and theory may guide the practice
⇒ Bridge two-way translation between model output and real-life data

Work in progress:

- ▷ N.M & P.L.M, 1 intern and 1 phD: the Benralitrap project.
- ▷ JJ & P.D, 1 phD: AIT-CRO project.

Organization

Local ecosystem

- ISDM: Institute of Data Science of Montpellier
- IMAG: Institut Montpelliérain Alexander Grothendieck
- Joint group meeting with the research group in ML
- Montpellier Université d'Excellence, MUSE

Location

- Inria: 860 rue Saint Priest
- Idesp: Campus Santé, IURC, 641 avenue du doyen Gaston Giraud

Short - mid terms objectives

From a methodological point of view

- innovative methods to handle missing values
- new developpement in causal inference
- provide easy-to-use tools (such as R package) and reproducible pipelines to allow for direct deployment by stakeholders

From a patient/medical point of view

- personalized benefit of treatment (over time)
- identify subgroup of patients
- adoption by the medical community of advanced techniques

PreMeDICaL: bio-statistics and ML, methodological specificities, a rapid transfer through software development and focus on allergy

Long terms objectives

From a methodological point of view

- new area for multiple imputation with non random missing values - inclusion of new data collected by medical devices
- designing future clinical trials supported by authorities (run trials to test assumptions). Organization of a défi inria
- software as decisions tools

From a patient/medical point of view

- give patient better care and early access to innovation
- guide decisions made by investigators, sponsors and authorities
- better management of resources

PreMeDICaL: bio-statistics and ML, methodological specificities, a rapid transfer through software development and focus on allergy