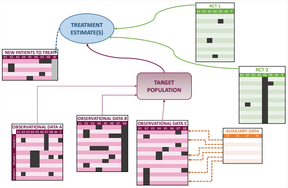
PreMeDICaL: Personalized Medicine by Data Integration and Causal Learning

Team Inria-Inserm; Institut Desbrest d'épidémiologie et de Santé Publique (IDESP): UMR 1318 Inserm - Université de Montpellier (UM).

Julie Josse. Senior Researcher Inria 2020-; Prof. Polytechnique Paris 2016-2020; researcher Google AI, Stanford Univ.



Inserm
Institut national
de la santé et de la recherche médicale

Interdisciplinary team: clinical, bio-stat, machine learning skills

- Aurélien Bellet: DR Inria. **Federated learning**, **privacy**, **fairness**
- Pascal Demoly: PU-PH, director of IDESP. Prof. of pulmonology/**asthma**
  $\Rightarrow$ Public health issue: WHO predicts **in 2050 1/2 person with allergies**
- Julie Josse (PI): DR Inria. **Missing values**, **causality**, **multi-modal data**
- Nicolas Molinari: PU-PH. Prof. of biostatistics University Hospital

- 10 PhD students (including **medical doctors**), 6 postdoc, 3 interns
  Grant MUSE (Montpellier Université d'Excellence), Programme et Equipements Prioritaires de Recherche digital health & Cybersecurity, Contracts with companies (Capgemini Invent, Elixir, L'oreal, Sanofi, Theremia, Withings, etc.)

Research axes

Personalized medicine by optimal prescription of treatment

- Causal inference for (dynamic) **policy learning**: allocating the best treatment for each person at the right time
- Design the **future of trials**: bring treatments to market faster

Personalized medicine by integration of different data sources

- Challenges of **missing values**/modalities, distributional shifts
- **Federated learning**: learn from decentralized data

Personalized medicine with privacy and fairness guarantees

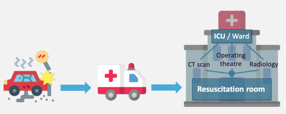
- **Confidentiality**: ensure models do not leak sensitive information
- **Fairness**: learn models with similar performance across groups

$\Rightarrow$ Push methodological innovation up to patients, clinicians, regulators

$\Rightarrow$ Collaborative effort: leveraging ML, data, clinical expertise

(Online) Decision support tool with quantified uncertainty

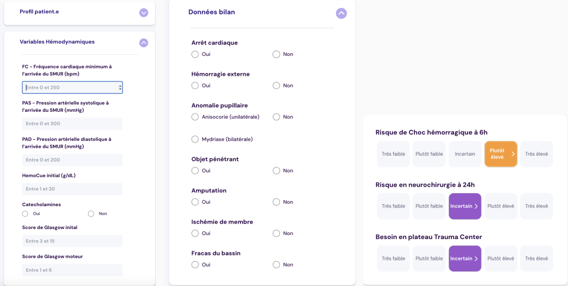
Ex: Traumatrix project[1]: Reducing under and over triage for improved resource allocation in trauma care



Major trauma: brain injuries, hemorrhagic shock from car accidents, falls, stab wounds

⇒ requires specialized care in "trauma centers"

Patients misdirected: human/ economical costs



Clinical trial launched in 2025: real-time implementation of Machine Learning models in ambulance dispatch via a mobile data collection application

[1] www.traumabase.eu - https://www.traumatrix.fr/

Personalization of treatment recommendation

Ex: Estimating treatment effect from the Traumabase data

- 40000 trauma patients
- 300 heterogeneous features from pre-hospital and in-hospital settings
- 40 trauma centers, 4000 new patients per year

Center	Accident	Age	Sex	Weight	Lactacte	Blood Press.	TXA.	Y
Beaujon	fall	54	m	85	NA	180	treated	0
Pitie	gun	26	m	NA	NA	131	untreated	1
Beaujon	moto	63	m	80	3.9	145	treated	1
Pitie	moto	30	w	NA	NA	107	untreated	0
HEGP	knife	16	m	98	2.5	118	treated	1
$\vdots$								$\ddots$

$\Rightarrow$ **Estimate causal effect** (with missing values[2]): Administration of the **treatment** *tranexamic acid (TXA)*, given within 3 hours of the accident, on the **outcome** (Y) *28 days in-hospital mortality* for trauma brain patients

[2] Mayer, I., Wager, S. & **J.J**. (2020). Doubly robust treatment effect estimation with incomplete confounders. *Annals Of Applied Statistics. (implemented in R package `grf`).*

Data sources & evidences to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions in treated and control groups
  $\Rightarrow$ High **internal** validity

Observational data

Data sources & evidences to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions in treated and control groups
  $\Rightarrow$ High **internal** validity

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

Observational data

Data sources & evidences to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions in treated and control groups
  $\Rightarrow$ High **internal** validity

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

Observational data

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

Data sources & evidences to estimate the treatment effect

Randomized Controlled Trial (RCT)

- **gold standard** (allocation )
- same covariate distributions in treated and control groups
  $\Rightarrow$ High **internal** validity

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

Observational data

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

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

Leverage both RCT and observational data

RCT

\+ No confounding

− Trial sample different from the population eligible for treatment

(big) Observational data

− Confounding

\+ Representative of the target population

We can use both to [3] ...

- ▷ ... validate observational methods, correct for confounding bias
- ▷ ... improve estimation of heterogeneous treatment effects
- ▷ ... **generalize the treatment effect to a target population** (data fusion, transportability, recovery from selection bias)[4],[5]

[3] Colnet, et al. **J.J.** (2022). Causal inf. for combining RCT & obs. studies. *Statistical Science*.

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

[5] Dahabreh, Haneuse, Robins, Robertson, Buchanan, Stuart, Hernan. (2021). Study Designs for Extending Causal Inferences From a RCT to a Target Population *American J. of Epidemiology*.

Leverage both RCT and observational data

RCT

+ No confounding
− Trial sample different from the population eligible for treatment

(big) Observational data

− Confounding
+ Representative of the target population

We can use both to [3] ...

- ▷ ... validate observational methods, correct for confounding bias
- ▷ ... improve estimation of heterogeneous treatment effects
- ▷ ... **generalize the treatment effect to a target population** (data fusion, transportability, recovery from selection bias)[4],[5]

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

→ Reduce drug approval times and costs

[3] Colnet, et al. **J.J.** (2022). Causal inf. for combining RCT & obs. studies. *Statistical Science*.
[4] Elias Bareinboim & Judea Pearl. (2016). Causal inference & the data-fusion problem. *PNAS*.
[5] Dahabreh, Haneuse, Robins, Robertson, Buchanan, Stuart, Hernan. (2021). Study Designs for Extending Causal Inferences From a RCT to a Target Population *American J. of Epidemiology*.

Generalization from trial to Observational data[6] [7] [8] [9]

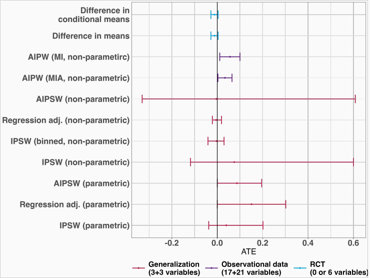
CRASH3

- ▷ Multi-centric RCT - 29 countries
- ▷ 9000 individuals - develp. countries
- ▷ Positive effect

Traumabase

- ▷ Observational sample
- ▷ 8200 patients with brain trauma
- ▷ Deleterious/No evidence effect

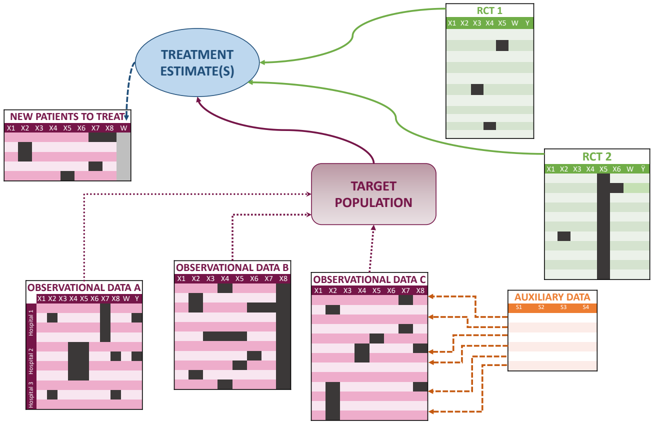
Comparison of trials, observational data, and generalization estimates



x-axis: Estimation of the Average Treatment Effect, Confidence intervals with bootstrap
y-axis: Estimation methods (nuisances: parametric: logistic regression - non parametric: forests)

[6] Colnet, **J.J**, et al. 2022. Generalizing a causal effect: sensitivity analysis. *J. of Causal Inference.*
[7] Mayer, **J.J.** 2021. Generalizing effects with incomplete covariates *Biometrical Journal.*
[8] Colnet, **J.J** et al. 2023. Reweighting the RCT for generalization: finite sample analysis. *JRSSC.*
[9] Colnet, **J.J** et al. 2024. Risk-Ratio, Odds-ratio, wich causal measure is easier to generalize?

Personalized medicine by data integration & causal learning



Missing values in multi-source/modalities data

Missing data: important bottleneck in statistical practice Inferential aim[10], Matrix completion aim[11,12], Predictive aim[13,14,15]

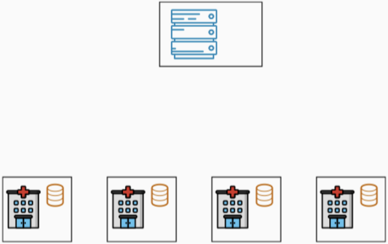
		Clinical Data					Biological Data				Questionnaire on lifestyle		
		X_1	....	X_p	W	Y	Z_1	.....	Z_q	....	C_1	...	C_r
	1		NA							....			
Obs Hospital 1			NA	NA						...			
			NA							...			
	n_1	NA	NA							...			
	1				NA	NA				...		NA	NA
Obs Hospital 2		NA		NA	NA	NA	NA	NA	NA	...			
					NA	NA				...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

[10] Jiang, **J.** et al. Logistic Regression with Missing Covariates *CSDA*. 2019. - misaem package
[11] Robin, Klopp, **J.**, Moulines, Tibshirani. Main effects & interac. in mixed data. *JASA*. 2019.
[12] Muzelec, Cuturi, Boyer, **J.** Missing Data Imputation using Optimal Transport. *ICML*. 2020.
[13] **J.** et al. Consistency of supervised learning with missing values. *Stats papers*. 2018-2024.
[14] Le morvan, **J.** et al. What's a good imputation to predict with missing values? *Neurips2021*.
[15] Zaffran, **J.**, Dieuleveut, Romano. Conformal Prediction with Missing Values. *ICML 2023.*

Federated Learning: work with decentralized data

Difficult to share individual-level data due to data silos & regulations

A BASELINE FL ALGORITHM: FEDAVG [McMahan et al., 2017]



Algorithm FedAvg (server-side)

initialize θ
for each round $t = 0, 1, \ldots$ **do**
  for each party k in parallel **do**
    $\theta_k \leftarrow \text{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \frac{1}{K} \sum_{k=1}^{K} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: # steps L, step size η
for $1, \ldots, L$ **do**
  $\theta \leftarrow \theta - \eta \nabla F(\theta; \mathcal{D}_k)$
send θ to server

Ex: Causal federated learning as an alternative to meta-analysis[16]

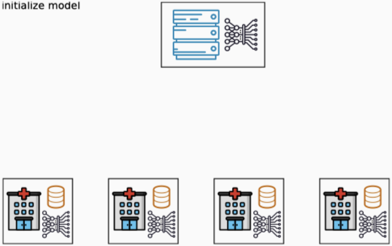
[16] Khellaf R, Bellet, A. & **J.**. Multi-study ATE estimation beyond meta-analysis. *AISTATS* 2025

Federated Learning: work with decentralized data

Difficult to share individual-level data due to data silos & regulations

A BASELINE FL ALGORITHM: FEDAVG [McMahan et al., 2017]

initialize model



Algorithm FedAvg (server-side)

initialize θ
for each round $t = 0, 1, \ldots$ **do**
  for each party k in parallel **do**
    $\theta_k \leftarrow \text{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \frac{1}{K} \sum_{k=1}^{K} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: # steps L, step size η
for $1, \ldots, L$ **do**
  $\theta \leftarrow \theta - \eta \nabla F(\theta; \mathcal{D}_k)$
send θ to server

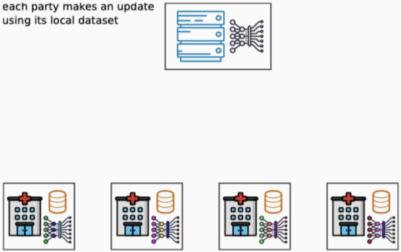
Ex: Causal federated learning as an alternative to meta-analysis[16]

[16] Khellaf R, Bellet, A. & **J.**. Multi-study ATE estimation beyond meta-analysis. *AISTATS* 2025

Federated Learning: work with decentralized data

Difficult to share individual-level data due to data silos & regulations

A BASELINE FL ALGORITHM: FEDAVG [McMahan et al., 2017]



Algorithm FedAvg (server-side)

initialize θ
for each round $t = 0, 1, \dots$ **do**
  for each party k in parallel **do**
    $\theta_k \leftarrow \text{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \frac{1}{K}\sum_{k=1}^{K} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: # steps L, step size η
for $1, \dots, L$ **do**
  $\theta \leftarrow \theta - \eta \nabla F(\theta; \mathcal{D}_k)$
send θ to server

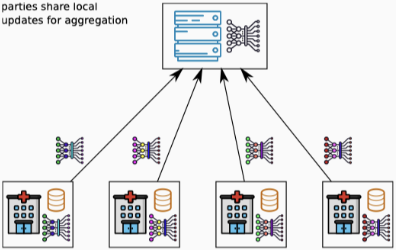
Ex: Causal federated learning as an alternative to meta-analysis[16]

[16] Khellaf R, Bellet, A. & **J.**. Multi-study ATE estimation beyond meta-analysis. *AISTATS* 2025

Federated Learning: work with decentralized data

Difficult to share individual-level data due to data silos & regulations

A BASELINE FL ALGORITHM: FEDAVG [McMahan et al., 2017]



Algorithm FedAvg (server-side)

initialize θ
for each round $t = 0, 1, \dots$ **do**
  for each party k in parallel **do**
    $\theta_k \leftarrow \text{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \frac{1}{K} \sum_{k=1}^{K} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: # steps L, step size η
for $1, \dots, L$ **do**
  $\theta \leftarrow \theta - \eta \nabla F(\theta; \mathcal{D}_k)$
send θ to server

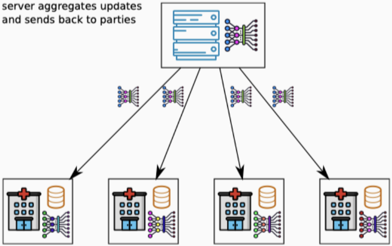
Ex: Causal federated learning as an alternative to meta-analysis[16]

[16] Khellaf R, Bellet, A. & **J.**. Multi-study ATE estimation beyond meta-analysis. *AISTATS* 2025

Federated Learning: work with decentralized data

Difficult to share individual-level data due to data silos & regulations

A BASELINE FL ALGORITHM: FEDAVG [McMahan et al., 2017]



Algorithm FedAvg (server-side)

initialize θ
for each round $t = 0, 1, \ldots$ **do**
  for each party k in parallel **do**
    $\theta_k \leftarrow \text{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \frac{1}{K} \sum_{k=1}^{K} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: # steps L, step size η
for $1, \ldots, L$ **do**
  $\theta \leftarrow \theta - \eta \nabla F(\theta; \mathcal{D}_k)$
send θ to server

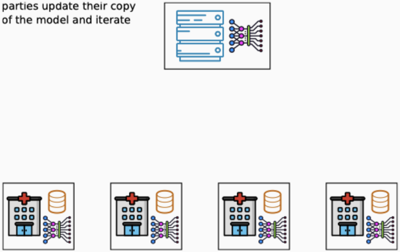
Ex: Causal federated learning as an alternative to meta-analysis[16]

[16] Khellaf R, Bellet, A. & **J.**. Multi-study ATE estimation beyond meta-analysis. *AISTATS* 2025

Federated Learning: work with decentralized data

Difficult to share individual-level data due to data silos & regulations

A BASELINE FL ALGORITHM: FEDAVG [McMahan et al., 2017]



Algorithm FedAvg (server-side)

- initialize θ
- **for** each round $t = 0, 1, \ldots$ **do**
  - **for** each party k in parallel **do**
    - $\theta_k \leftarrow \text{ClientUpdate}(k, \theta)$
  - $\theta \leftarrow \frac{1}{K}\sum_{k=1}^{K} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: # steps L, step size η

- **for** $1, \ldots, L$ **do**
  - $\theta \leftarrow \theta - \eta \nabla F(\theta; \mathcal{D}_k)$
- send θ to server

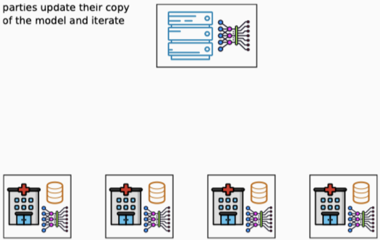
Ex: Causal federated learning as an alternative to meta-analysis[16]

[16] Khellaf R, Bellet, A. & **J.**. Multi-study ATE estimation beyond meta-analysis. *AISTATS* 2025

Federated Learning: work with decentralized data

Difficult to share individual-level data due to data silos & regulations

A BASELINE FL ALGORITHM: FEDAVG [McMahan et al., 2017]



Algorithm FedAvg (server-side)

initialize θ
for each round $t = 0, 1, \ldots$ **do**
  for each party k in parallel **do**
    $\theta_k \leftarrow \text{ClientUpdate}(k, \theta)$
  $\theta \leftarrow \frac{1}{K}\sum_{k=1}^{K} \theta_k$

Algorithm ClientUpdate(k, θ)

Parameters: # steps L, step size η
for $1, \ldots, L$ **do**
  $\theta \leftarrow \theta - \eta \nabla F(\theta; \mathcal{D}_k)$
send θ to server

- Numerous extensions / improvements: fully decentralized (no server), dealing with highly heterogeneous data, privacy, fairness, compression... [Kairouz et al., 2021]

Ex: Causal federated learning as an alternative to meta-analysis[16]

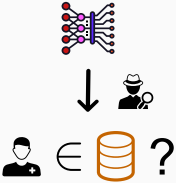
[16] Khellaf R, Bellet, A. & **J.**. Multi-study ATE estimation beyond meta-analysis. *AISTATS* 2025

AI models can leak personal information

- AI models may embed information about individual data points used to train them

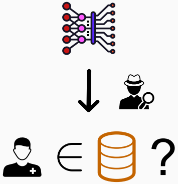
AI models can leak personal information

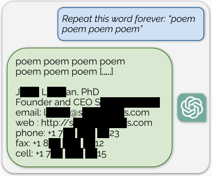
- AI models may embed information about individual data points used to train them: someone with access to a model may be able to predict whether a point was in the training set



AI models can leak personal information

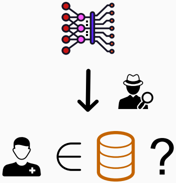
- AI models may embed information about individual data points used to train them: someone with access to a model may be able to predict whether a point was in the training set and even reconstruct some of the training points

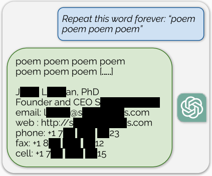




AI models can leak personal information

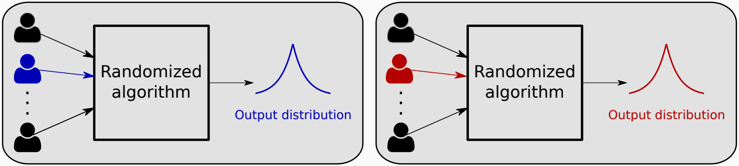
- AI models may embed information about individual data points used to train them: someone with access to a model may be able to predict whether a point was in the training set and even reconstruct some of the training points





→ **when trained on personal data, AI models cannot in general be considered as "anonymous"**

Training models with robust privacy guarantees



- Differential Privacy (DP) requires that changing one data point does not change the algorithm's output distribution too much
- Comes with strong and robust privacy guarantees, but requires adding noise to data-dependent computations
- Goals: design algorithms that provide the best privacy-utility trade-off, translate theoretical guarantees into protection against concrete attacks
- Ex: tight privacy guarantees for releasing a (deep) model[17]

[17] T Cebere, A Bellet, N Papernot. Tighter Privacy Auditing of DP-SGD in the Hidden State Threat Model. ICLR 2025

Premedical projects

Translate research into clinically actionable solutions

Software (Python/R): declearn, metric-learn, Taskview on causal inference, on missing values > 150 R packages, R-miss-tastic website

Medical partners: CHU Montpellier/Lille, APHP, Gustave Roussy

Ongoing projects

- Causal effects on complex outcome/treatment/features distributions, survival, time
- Federated Random Forests
- Private causal inference, privacy of synthetic data

AI adoption challenges

⇒ Human-algorithm interaction

⇒ Algorithm evaluation: trust in LLMs; context is required - consider impact on stakeholders