RESEARCH CENTRE
Inria Branch at the University of Montpellier

IN PARTNERSHIP WITH:

INSERM, Université Paul-Valéry Montpellier 3, Université de Montpellier

2023
ACTIVITY REPORT

Project-Team
PREMEDICAL

Precision Medicine by Data Integration and Causal Learning

IN COLLABORATION WITH: Institut Desbrest d'Épidémiologie et de Santé Publique (IDESP)

DOMAIN

Digital Health, Biology and Earth

THEME

Computational Neuroscience and Medicine

Inria

Contents

Project-Team PREMEDICAL

Creation of the Project-Team: 2022 June 01

Keywords

Computer sciences and digital sciences

A6.1. – Methods in mathematical modeling

A9.2. – Machine learning

A9.6. – Decision support

Other research topics and application domains

B2. – Health

B2.2. – Physiology and diseases

B2.3. – Epidemiology

1   Team members, visitors, external collaborators

Research Scientists

- Julie Josse [Team leader, INRIA, Advanced Research Position, HDR]

- Aurélien Bellet [INRIA, Senior Researcher, from Oct 2023, HDR]

- Aurélien Bellet [INRIA, Researcher, from Aug 2023 until Sep 2023, HDR]

Faculty Members

- Pascal Demoly [UNIV MONTPELLIER, Professor, Director of Idesp (UMR UM-INSERM)]

- Pierre Lafaye De Micheaux [UNIV MONTPELLIER III, Associate Professor, until Nov 2023]

- Nicolas Molinari [UNIV MONTPELLIER - PUPH, Professor, CHU Montpellier]

Post-Doctoral Fellow

- Jeffrey Naf [UNIV MONTPELLIER, from Feb 2023]

PhD Students

- Marie Felicia Beclin [UNIV MONTPELLIER]

- Ahmed Boughdiri [INRIA, from Oct 2023]

- Maxime Fosset [UNIV MONTPELLIER]

- Remi Khellaf [UNIV MONTPELLIER, from Oct 2023]

- Charlotte Voinot [SANOFI, CIFRE, from Apr 2023]

- Margaux Zaffran [INRIA, from Dec 2023]

- Margaux Zaffran [EDF, CIFRE, until Nov 2023]

- Pan Zhao [UNIV MONTPELLIER]

Technical Staff

- Ahmed Boughdiri [UNIV MONTPELLIER, Engineer, from Apr 2023 until Sep 2023]

Interns and Apprentices

- Pauline Bian [ELIXIR, Intern, from Oct 2023]

- Helene Bonneau–Chloup [ELIXIR, Intern, from Oct 2023]

- Remi Khellaf [Quinten Health, Intern, from Apr 2023 until Aug 2023]

Administrative Assistant

- Claire-Marine Parodi [INRIA, from Sep 2023]

External Collaborator

- Imke Mayer [CHARITE UNIV BERLIN, from Feb 2023 until Jul 2023]

2   Overall objectives

The objective of the PreMeDICaL team (**Pre**cision **Me**dicine by **D**ata Integration and **Ca**usal **L**earning) is to develop the next generation of methods/algorithms to extract knowledge from health data and improve the care of patients. More specifically, the aim is to develop learning tools for **personalized treatment effect prediction** and for **predicting outcome**, while **integrating different data sources** to **guide decisions** made by clinicians and authorities. PreMeDICaL has three research axes:

1. Personalized medicine by optimal prescription of treatment. We will develop causal inference techniques for **(dynamic) policy learning** (allocating the best treatment for each person at the right time), that handle missing values and leverage both RCTs and observational data. Using both data sources allow to **better design future RCTs** or to **launch a drug without running RCTs** and **in the longer term to rethink the evidence needed to bring treatments to the market** and to do so more quickly.

2. Personalized medicine by integration of different data sources. We will build predictive models for heterogeneous data: for instance given monitoring data in continuous time, images and clinical data what is the risk for an event to occur? Is it useful to have all the sources or do they provide the same information? We will additionally develop solutions to **learn from decentralized data (federated learning)**, to **handle missing values** in a supervised learning setting and **to improve the confidence of the outputs of the predictive models**.

3. Personalized medicine with privacy and fairness guarantees. We develop approaches to **ensure the confidentiality** of medical data and guarantee that models do not leak sensitive information. We additionally build methods to **handle fairness constraints** to ensure that models exhibit similar performance across different population groups.

The aim is to **push methodological innovation up to the stakeholders** (patients, clinicians, regulators, etc.). Consequently, beyond these methodological developments, innovative responses to the public health challenge posed by respiratory allergies are targeted. In addition to **leveraging machine learning algorithms and leveraging appropriate data, combining them with clinical expertise and existing recommendations** is necessary. Long- term aims are to have both a strong scientific and societal impact with a substantial impact on the quality of care for patients and major consequences for the medical profession by **providing a much earlier access to innovative solutions and more efficient treatment and care.** With a successful proof of concept in the domain of allergies, by having clear **reproducible pipelines, methodologies, software** (by providing clinical decision making system tools) we could thereafter consider other pathologies (such as traumatology and oncology studied at IDESP). Hence, a joint team between Inria and Inserm provides a unique opportunity for trans-disciplinary research and collaboration bringing together mathematical, methodological, technological and medical expertise. The PreMeDICaL team contributes to precision medicine (where the treatment/device is adapted on a patient basis) and to translational medicine which aims at bridging the gap between fundamental research and its practical use.

3   Research program

3.1   Research Axis 1: Personalized medicine by optimal prescription of treatment

In machine learning (ML)/artificial intelligence (AI) progress has yielded powerful predictive models, yet they rely on correlations and lack an understanding of underlying mechanisms or intervention strategies. Causality is crucial for actionable insights, recommendations, and addressing "what if" scenarios, with applications in health, public policies, econometrics, and advertising. Causal inference gains prominence for addressing AI challenges like interpretability and robustness offering solutions akin to "AI-like human" approaches in novel settings. This axis aims to innovate causal machine learning at the AI-personalized medicine intersection, optimizing treatment allocation and enabling drug launches without randomized control trials.

Randomized controlled trials are considered the gold standard approach for assessing the causal effect (i.e., the treatment effect) of an intervention or a treatment on an outcome of interest. Indeed, the allocation of the treatment is under control, which implies that there is no confounding factors (the distribution of covariates for treated and control patients is asymptotically balanced) that could interfere with the treatment and simple estimators (such as the difference in mean effect between the treated and controls) can be used to consistently estimate the average treatment effect (ATE). However, RCTs can come with drawbacks. They can be expensive, take a long time to set up, and be compromised by insufficient sample size due to either recruitment difficulties or restrictive inclusion/exclusion criteria. These criteria can lead to a narrowly defined trial sample that differs markedly from the population potentially eligible for the treatment (distributional shift). Therefore, the findings from RCTs can lack generalizability (or *external validity*). This has been largely published in the field of respiratory and allergic diseases, see for instance [32] which highlights that the population from RCTs represents less than 10% of the population that will receive treatments.

In contrast, there is an abundance of observational data, collected without systematically designed interventions. Such data can come from different sources: they can be collected from research sources (such as disease registries, cohorts, biobanks, epidemiological studies), or they can be routinely collected (through electronic health records, insurance claims, administrative databases, patients' App, etc). In that sense, observational data can be readily available, can include large samples representative of the target populations, and can be less costly than RCTs. To leverage observational data for treatment effect estimation in health domains, several laws built on studies by the USA Food and Drug Administration (FDA) encourage the use of "real world data" (RWD), defined as data "derived from sources other than randomized clinical trials", for regulatory decision making. Clinical evidence regarding the usage and potential benefits or risks of a medical product derived from the analysis of RWD is named Real World Evidence (RWE). The European Medicines Agency (EMA) is also a very active regulatory authority working with RWD to facilitate development and access to medicines. However, despite the large number of methods available to estimate the causal treatment effect from observational data such as matching, inverse probability weighting (IPW) or more recent doubly robust methods based on machine learning there are often concerns about the quality of these "big data" and causal claims. Indeed, building on observational data is still not consensual due to the lack of controlled experimental interventions, which opens the door to confounding biases (lack of *internal validity*).

Observational data and clinical trial data can provide different perspectives when evaluating an intervention or a medical treatment. Combining the information gathered from experimental and observational data is a promising avenue for medical research, because the knowledge acquired from integrative analyses could not be gathered from a single-source analysis alone. Three potential high impact applications of observational and clinical data are:

1. Predicting the effect of a treatment estimated on a RCT, on a new target population (generalization);

2. Comparing RCTs and RWE to validate observational methods;

3. Better estimation of heterogeneous treatment effects.

There is an abundant literature on bridging the findings from an RCT to a target population and combining both sources of information. Similar problems have been termed as *transportability*, and *data fusion* and have connections to the *covariate shift/domain generalization* problem in ML. [21] reviewed the methods to (a) generalize the treatment effect while integrating the distributional shift (IPSW, g-formula, AIPSW, calibration weighting, etc.), or (b) improve the estimate of the conditional average treatment effect (CATE, i.e. heterogeneous effect) while correcting for confounding factors not measured in the observational study. However, these methods have many shortcomings and there are still many challenges to address. We provide below examples of methodological locks we will overcome.

- Handling missing values and unmeasured covariates with multi-source data;

- Transfer Learning of optimal individualized treatment regimes with right-censored survival data;

- Policy learning and dynamic treatment policy with missing values;

- Generalization of different causal measures: Risk Ratio, Survival Ratio, etc;

- Providing finite sample guarantees;

- Study of causal effects in metric spaces

- Guide variable selection and provide importance variables measures and tests in treatment effects setting

Such development will have significant societal impact in patient care and cost reduction, ultimately guiding future RCT designs.

3.2   Research axis 2: Personalized medicine by integration of different data sources

In this axis we focus both on integrating heterogeneous data/multiview/multimodal (time series, images, text, numerical or categorical data) potentially from different centers to establish predictive, as well as quantifying the uncertainty associated to predictive models. For the former, we will focus on handling missing values and on federated learning strategies, while for the latter we will consider uncertainty quantification approaches.

Federated learning [30] is a recent paradigm which enables model training across decentralized devices or servers holding local data samples, without exchanging them. Only the model updates, not the raw data, are sent to a central server, where they are aggregated to improve the global model. In the medical domain, federated learning helps to address privacy concerns by allowing models to be trained on data distributed across various healthcare institutions and/or companies without centrally aggregating sensitive patient information. This facilitates collaborative inference without compromising data security, making it particularly valuable for developing robust and generalizable medical AI models across diverse datasets while respecting privacy regulations.

Most statistical learning and artificial intelligence methodologies provide point predictions, without any indication of the degree of confidence that can be given to these predictions (i.e. without predictive intervals). This lack of uncertainty quantification of predictive models is a **major barrier to the adoption of powerful machine learning methods by society**. Probabilistic forecasts, i.e. predicting the entire distribution probability and not only the conditional expectation, could partially tackle this issue but they are only valid asymptotically, require strong assumptions on the data (e.g. normality) or/and are model-dependent. The emergent field of conformal prediction (CP) [38, 34, 31] is a promising framework for **distribution-free uncertainty quantification**. It is a general procedure to build predictive intervals for any predictive model (including black-box methods such as deep learning), which are valid (i.e. achieve nominal marginal coverage), in finite sample, and without assumption on the data generation process except the exchangeability. This is extremely promising for decision support tools in critical applications: healthcare, autonomous driving, etc. An extension of CP (Conformalized Quantile Regression, [35]) was used to predict the U.S. presidential elections (2020) by the Washington Post.

We provide below examples of methodological challenges we will overcome.

- Relationship between the different sources;

- (Informative) missing values in time series and structured by blocks;

- Conformal prediction with missing values [9]; Relationship between predictive intervals and confidence intervals

- Federated learning with missing values;

- Federated causal inference.

3.3   Research Axis 3: Personalized medicine with privacy and fairness guarantees

In this axis, we aim to address privacy and fairness concerns in machine learning, with a focus on the challenges raised by medical applications. By integrating privacy and fairness into the design of the algorithms, we can enhance the trustworthiness of machine learning applications, promote ethical practices, and facilitate the responsible deployment of personalized medicine technologies for the benefit of diverse patient populations.

While training ML models on personal or otherwise confidential data can be beneficial in many applications such as healthcare, this can also lead to undesirable disclosure of sensitive information. Take for instance patient records, which often contain highly personal and identifiable information such as medical histories, diagnostic results, and genetic data. If a machine learning model trained on this data is not appropriately designed and secured, it may be possible for an attacker to deduce private information about individuals by analyzing the output of the model. Indeed, concrete attacks have been designed to predict whether a particular individual was part of the training set [37], and even to reconstruct some of the training data points [33]. Privacy-preserving machine learning aims to mitigate these concerns by incorporating techniques that safeguard sensitive information during the training and deployment of models. We focus on Differential Privacy (DP), a framework that provides a mathematical definition of privacy guarantees. In a nutshell, DP ensures that the inclusion or exclusion of any single data point does not significantly impact the output distribution of the training algorithm, thereby bounding the amount of information that can be inferred from the trained model about any individual in the dataset. DP requires to incorporate a certain amount of randomness into the algorithms, and thus yields a necessary trade-off between privacy and utility (e.g., accuracy of the resulting model). A key challenge is then to design methods that achieve the best possible trade-offs. We consider both centralized training by a trusted curator, and federated/decentralized training by participants who do not trust each other. We seek to characterize the achievable trade-offs, and to design algorithms with optimal privacy-utility trade-offs for a variety of machine learning and statistical inference tasks. Finally, we will also consider the relationship between missing values imputation methods and the generation of synthetic data which is often used to tackle privacy constraints.

Fairness considerations are also vital in machine learning to avoid bias in algorithms. Indeed, biased models could lead to unequal treatment of individuals based on factors like ethnicity or gender [36], potentially exacerbating healthcare disparities. For instance, if a machine learning model is trained predominantly on data from a specific demographic group, it may not generalize well to other groups, leading to inaccurate predictions for underrepresented populations. This can result in suboptimal health-care outcomes, with certain individuals receiving inadequate attention or misdiagnoses. Additionally, historical biases present in healthcare data may be learned by machine learning models and perpetuated in their predictions. We aim to address these fairness challenges by incorporating fairness considerations into the machine learning pipeline, i.e., during data collection and preprocessing, model training and/or evaluation. An approach of particular interest is the introduction of group fairness constraints during the training phase [39]. Such constraints explicitly define the desired level of fairness and prevent the model from making predictions that disproportionately favor or disfavor specific population groups. As for privacy, we seek to study fairness in centralized training, but also in the context of federated learning which raises specific challenges as fairness on decentralized data becomes difficult to measure globally.

In addition to considering privacy and fairness in machine learning separately, we also aim to understand the interplay and potential tension between these two requirements, as well as to design algorithms that can provide optimal and tunable trade-offs.

4　Application domains

The first application domain of PreMeDICaL is **respiratory diseases** and in particular **Asthma**. For more than 30 years, there has been an increase in a number of chronic non-communicable diseases (NCD), such as asthma and allergies, respiratory diseases. Allergies are the fourth most common chronic disease in the world. The World Health Organization (WHO) predicts that by 2050, one in two people in the world will suffer from allergies. In France, the number of people suffering from allergies has doubled in 20 years, particularly among children and young people. Although the expression of these diseases results from the interaction between the genetic background and the environment, especially through epigenetic mechanisms, their sudden increase is solely due to the environmental changes that occurred in the last decades because of the Western lifestyle, the genetic heritage requiring centuries to change. A full understanding of the complexity of chronic NCD prompts researchers to analyze large data utilizing proper markers and tools (e.g., biological, clinical, behavioral, economic, social, demographic, environmental data, patient experience, patient social networks) in an etiological and evaluative way to determine phenotypical patients' pathways, explain their impacts, their causes, their influences, prevent

them and improve their prognosis. Integrating these different sources of information, collected by several actors (healthcare professionals, public authorities or patients themselves), thus offer new opportunities to design personalized solutions by adapting treatment to the patient and the organizational context, leading to improved patient care and prevention policies.

With a successful proof of concept in the domain of allergies, by having clear **reproducible pipelines, methodologies, software**, we will thereafter consider other pathologies (such as traumatology and oncology studied at IDESP).

5   Social and environmental responsibility

5.1   Impact of research results

From a methodological point of view, the aim is to improve and develop new statistical and ML methods for establishing evidence on the efficiency of treatment by data enrichment (data fusion) and for predicting outcomes quantifying the uncertainty. An important output of this research is that these methodological works have a concrete impact on designing future clinical trials and that the new methodology will be supported by regulatory authorities. Indeed, exploiting both RCTs and observational data serve different purposes such as prediction of the treatment effect on new populations, increasing the generalization of clinical trials (so that they are more representative of the patient population who may benefit from the treatment) and also defining new inclusion criteria (because we identify subgroups who can benefit from treatment). This research is part of the PEPR project "Next methodological challenges in clinical trials in the era of digital health". Through axis 3 of our research program, we also aim to design methods that can incorporate some societal requirements related to fairness and privacy.
From a technological point of view, the aim is to provide software (starting with open access) for these methods to be applied in practice by studies stakeholders, clinicians and the clinical trial community.
From the clinical and patients point of view, the different projects aim to quantify the clinical benefit of intervention (over time), taking into account all patient characteristics, and to provide useful clinical prognosis tools allowing clinicians to optimally treat every patient, while also guaranteeing some level of fairness and privacy. The aim is to give patients better care and early access to innovation. In addition, these works can lead to a better adoption by the medical community of certain (advanced) techniques used to estimate the effects of treatment on patients (by comparing the results obtained in an RCT with the RWE).
From a public-health point of view, the aim is to guide decisions made by investigators, sponsors and authorities. Better trials' designs may also have an important impact in terms of cost reduction. Finally, we aim at having a significant impact in the field of allergy treatments providing new knowledge that may change guidelines and practice.

6   Highlights of the year

6.1   Awards

Margaux Zaffran received the For Women In Science Fondation L'Oréal – UNESCO French Young Talents Award.

6.2   Other

- Aurélien Bellet (Senior Researcher) has joined PreMeDICaL.

- PreMeDICaL is involved in 3 projects of PEPR Digital Health that have started in 2023.

- **Traumatrix project**:

  The objective of Traumatrix is to support SAMU regulation in the prioritization of severe trauma patients by providing targeted and individualized predictions of patient needs: risk of hemorrhagic shock, need for neurosurgery, need for Trauma therapies. These predictions will: Improve the

orientation of patients with severe trauma and reduce their under-triage; Graduate the severity of the patient to better prepare for reception and provision of resources within the care center.

– Following the PREPS - Programme de Recherche sur la Performance des Soins, obtained, a datathon has been organized with Clinicians and Premedical to test the machine learning methods that will be implemented in the decision support tool. A clinical trial will be launched in February 2024 to test the tool in sixteen dispatch centers in France.

– The partnership bringing together Traumabase, CNRS, EHESS, École Polytechnique, CNRS, INRIA, APHP, CHU Grenoble and Capgemini Invent is extended until June 2025.

7 New software, platforms, open data

7.1 New software

7.1.1 factominer

Keywords: Dimensionality reduction, PCA, Text mining, Clustering

Functional Description: The FactoMineR package is dedicated to performing principal components methods to explore, sum-up and visualize data. Dimensionality reduction methods include PCA, correspondence analysis (CA) for count data such as documents-words data, multiple correspondence analysis (MCA) for categorical data such as survey data, factorial analysis of mixed data (FAMD) for both types of variables as well as methods for groups of variables, of individuals (multiple factorial analysis, MFA), for hierarchy …

References: https://husson.github.io/MOOC_AnaDo/index.html https://husson.github.io/MOOC.html#PCAcourse

URL: http://factominer.free.fr/index_fr.html

Contact: Julie Josse

Partner: AGROCAMPUS

7.1.2 missMDA

Keyword: Missing data

Functional Description: The missMDA package is dedicated to missing values in and with Multivariate Data Analysis. It allows one to apply PCA, MCA, FAMD and MFA on incomplete data. It performs single and multiple imputation for continuous, categorical and mixed data based on principal components methods

URL: http://factominer.free.fr/missMDA/index.html

Contact: Julie Josse

Partner: AGROCAMPUS

7.1.3 metric-learn

Keywords: Machine learning, Python, Metric learning

Functional Description: Distance metrics are widely used in the machine learning literature. Traditionally, practicioners would choose a standard distance metric (Euclidean, City-Block, Cosine, etc.) using a priori knowledge of the domain. Distance metric learning (or simply, metric learning) is the sub-field of machine learning dedicated to automatically constructing optimal distance metrics.

This package contains efficient Python implementations of several popular metric learning algorithms.

URL: https://github.com/scikit-learn-contrib/metric-learn

Contact: Aurélien Bellet

Partner: Parietal

7.1.4 declearn

Keyword: Federated learning

Scientific Description: declearn is a python package providing with a framework to perform federated learning, i.e. to train machine learning models by distributing computations across a set of data owners that, consequently, only have to share aggregated information (rather than individual data samples) with an orchestrating server (and, by extension, with each other).

The aim of declearn is to provide both real-world end-users and algorithm researchers with a modular and extensible framework that:

(1) builds on abstractions general enough to write backbone algorithmic code agnostic to the actual computation framework, statistical model details or network communications setup

(2) designs modular and combinable objects, so that algorithmic features, and more generally any specific implementation of a component (the model, network protocol, client or server optimizer...) may easily be plugged into the main federated learning process - enabling users to experiment with configurations that intersect unitary features

(3) provides with functioning tools that may be used out-of-the-box to set up federated learning tasks using some popular computation frameworks (scikit- learn, tensorflow, pytorch...) and federated learning algorithms (FedAvg, Scaffold, FedYogi...)

(4) provides with tools that enable extending the support of existing tools and APIs to custom functions and classes without having to hack into the source code, merely adding new features (tensor libraries, model classes, optimization plug-ins, orchestration algorithms, communication protocols...) to the party.

Parts of the declearn code (Optimizers,...) are included in the FedBioMed software.

At the moment, declearn has been focused on so-called "centralized" federated learning that implies a central server orchestrating computations, but it might become more oriented towards decentralized processes in the future, that remove the use of a central agent.

Functional Description: This library provides the two main components to perform federated learning:

(1) the client, to be run by each participant, performs the learning on local data et releases only the result of the computation

(2) the server orchestrates the process and aggregates the local models in a global model

URL: https://gitlab.inria.fr/magnet/declearn/declearn2

Contact: Aurélien Bellet

Participants: Paul Andrey, Aurélien Bellet, Nathan Bigaud, Marc Tommasi, Nathalie Vauquier

Partner: CHRU Lille

7.2 New platforms

- Causal inference taskview: to list and organize all the R packages on causal inference

- R-miss-tastic **a platform for missing data to gather and create resources for users**, researchers and students who often don't have lecture on missing values: bibliography, courses, tutorials, implementations, pipelines of analysis in R and Python.

  Participants: Julie Josse, Pan Zhao.

8   New results

8.1   Treatment effect estimation

Results: Choice of the causal measure - Publications [2]

> **Participants:**   Benedicte Colnet, Julie Josse.

There are many measures to report so-called treatment or causal effect: absolute difference, ratio, odds ratio, number needed to treat, and so on. The choice of a measure, e.g. absolute versus relative, is often debated because it leads to different appreciations of the same phenomenon; but it also implies different heterogeneity of treatment effect. In addition some measures – but not all – have appealing properties such as collapsibility, matching the intuition of a population summary. We review common measures and their pros and cons typically brought forward. Doing so, we clarify notions of collapsibility and treatment effect heterogeneity, unifying different existing definitions. Our main contribution is to propose to reverse the thinking: rather than starting from the measure, we start from a non-parametric generative model of the outcome. Depending on the nature of the outcome, some causal measures disentangle treatment modulations from baseline risk. Therefore, our analysis outlines an understanding of what heterogeneity and homogeneity of treatment effect mean, not through the lens of the measure, but through the lens of the covariates. Our goal is the generalization of causal measures. We show that different sets of covariates are needed to generalize an effect to a different target population depending on (i) the causal measure of interest, (ii) the nature of the outcome, and (iii) the generalization's method itself (generalizing either conditional outcome or local effects).

Results: Variable importance for causal and distributional random forest - Publications [18] [19]

> **Participants:**   Julie Josse, Jeffrey Naf.

Causal random forests provide efficient estimates of heterogeneous treatment effects. However, forest algorithms are also well-known for their black-box nature, and therefore, do not characterize how input variables are involved in treatment effect heterogeneity, which is a strong practical limitation. In [18], we develop a new importance variable algorithm for causal forests, to quantify the impact of each input on the heterogeneity of treatment effects. The proposed approach is inspired from the drop and relearn principle, widely used for regression problems. Importantly, we show how to handle the forest retrain without a confounding variable. If the confounder is not involved in the treatment effect heterogeneity, the local centering step enforces consistency of the importance measure. Otherwise, when a confounder also impacts heterogeneity, we introduce a corrective term in the retrained causal forest to recover consistency. Additionally, experiments on simulated, semi-synthetic, and real data show the good performance of our importance measure, which outperforms competitors on several test cases to recover important variables. Experiments also show that our approach can be efficiently extended to groups of variables, providing key insights in practice.

Distributional Random Forest (DRF) is a flexible forest-based method to estimate the full conditional distribution of a multivariate output of interest given input variables. In [19], we introduce a variable importance algorithm for DRFs, based on the same drop and relearn principle and MMD distance. While traditional importance measures only detect variables with an influence on the output mean, our algorithm detects variables impacting the output distribution more generally. We show that the introduced importance measure is consistent, exhibits high empirical performance on both real and simulated data, and outperforms competitors. In particular, our algorithm is highly efficient to select variables through recursive feature elimination, and can therefore provide small sets of variables to build accurate estimates of conditional output distributions.

Results: Distribution on Distribution Regression to model Treatment Response Assessment in Asthma Patients

Participants:	Marie Felicia Beclin, Nicolas Molinari, Pierre Lafaye De Micheaux.

Medical imaging plays a crucial role in evaluating treatment efficacy. While practitioners traditionally rely on specific biomarkers and clinical data, incorporating informative features derived from medical imaging can enhance treatment response prediction. This research focuses on thoracic scans taken in expiration and inspiration before and after one year of Benralizumab treatment for asthma patients.

Following image segmentation, histograms are calculated to represent the distribution of voxel intensities. The underlying hypothesis posits that patients with improved conditions will exhibit enhanced expiration scans after treatment, evident in the histograms through a rightward shift, indicating higher Hounsfield Unit (HU) values. To predict treatment's response, we develop an histogram on histogram regression. Unlike existing methods, our proposed model goes beyond point-wise estimation of coefficient, offering an inferential framework to obtain p-values and confidence intervals for assessing treatment effects.

Results: Assessing Safety of Mechanical Ventilation Weaning in Patients Receiving Continuous Vasopressors: An Emulated Target Trial

Participants:	Maxime Fosset, Nicolas Molinari, Julie Josse.

Purpose: Safety of weaning critically ill patients from mechanical ventilation while receiving vasopressor is uncertain. To identify the optimal strategy of vasopressor discontinuation at the time of weaning in adult critically ill medical patients, we conducted an emulated target trial to investigate the total causal effect of four different treatment strategies on the risk of weaning failure by day 7. We hypothesized that weaning while receiving a low-dose of vasopressor would be the strategy with the lowest risk of weaning failure.

Methods: We performed an emulated trial with consecutive observational data from patients admitted to the intensive care units at Beth Israel Deaconess Medical Center from January 2011 to June 2022. We compared the risk of death or reintubation within 7 days of weaning among four a priori defined weaning strategies: delayed weaning after vasopressor discontinuation, early weaning, low-dose, and high-dose vasopressor infusion.

Results: The results showed that weaning while receiving low-dose vasopressors reduced the duration of mechanical ventilation but increased the risk of weaning failure compared to a delayed strategy where vasopressors were discontinued for more than 24 hours. On the other hand, weaning while receiving a high-dose of vasopressors increased the risk of weaning failure. The study suggests that future randomized trials may evaluate weaning in critically ill patients while receiving low-dose vasopressor strategy.

8.2   Uncertainty quantification with conformal prediction

Participants:	Julie Josse.

Results: Conformal prediction for time series - publication [26]

Conformal Inference (CI) is a popular approach for generating finite sample prediction intervals based on the output of any point prediction method when data are exchangeable. Adaptive Conformal Inference (ACI) algorithms extend CI to the case of sequentially observed data, such as time series, and exhibit strong theoretical guarantees without having to assume exchangeability of the observed data. The common thread that unites algorithms in the ACI family is that they adaptively adjust the width of the

generated prediction intervals in response to the observed data. We provide a detailed description of five ACI algorithms and their theoretical guarantees, and test their performance in simulation studies. We then present a case study of producing prediction intervals for influenza incidence in the United States based on black-box point forecasts. Implementations of all the algorithms are released as an open-source R package, AdaptiveConformal, which also includes tools for visualizing and summarizing conformal prediction intervals.

8.3 Learning with privacy guarantees

Participants: Aurélien Bellet.

Results: Rényi Pufferfish Privacy [23]

Pufferfish privacy is a flexible generalization of differential privacy that allows to model arbitrary secrets and adversary's prior knowledge about the data (e.g., correlation across individuals). Unfortunately, designing general and tractable Pufferfish mechanisms that do not compromise utility is challenging. Furthermore, this framework does not provide the composition guarantees needed for a direct use in iterative machine learning algorithms. To mitigate these issues, we introduce a Rényi divergence-based variant of Pufferfish and show that it allows us to extend the applicability of the Pufferfish framework. We first generalize the Wasserstein mechanism to cover a wide range of noise distributions and introduce several ways to improve its utility. We also derive stronger guarantees against out-of-distribution adversaries. Finally, as an alternative to composition, we prove privacy amplification results for contractive noisy iterations and showcase the first use of Pufferfish in private convex optimization. A common ingredient underlying our results is the use and extension of shift reduction lemmas.

Results: Relative Gaussian Mechanism [22]

The Gaussian Mechanism (GM), which consists in adding Gaussian noise to a vector-valued query before releasing it, is a standard privacy protection mechanism. In particular, given that the query respects some L2 sensitivity property (the L2 distance between outputs on any two neighboring inputs is bounded), GM guarantees Rényi Differential Privacy (RDP). Unfortunately, precisely bounding the L2 sensitivity can be hard, thus leading to loose privacy bounds. In this work, we consider a Relative L2 sensitivity assumption, in which the bound on the distance between two query outputs may also depend on their norm. Leveraging this assumption, we introduce the Relative Gaussian Mechanism (RGM), in which the variance of the noise depends on the norm of the output. We prove tight bounds on the RDP parameters under relative L2 sensitivity, and characterize the privacy loss incurred by using output-dependent noise. In particular, we show that RGM naturally adapts to a latent variable that would control the norm of the output. Finally, we instantiate our framework to show tight guarantees for Private Gradient Descent, a problem that naturally fits our relative L2 sensitivity assumption.

8.4 Application domain: allergies

Participants: Pascal Demoly, Nicolas Molinari.

Results: ResultSkin Test Reactivity Patterns in Patients Allergic to Iodinated Contrast Media: A Refined View [29]

Background: Two-dimensional (2D) classifications of iodinated contrast media (ICM) are insufficient to explain the observed skin test (ST) reactivity patterns in patients with drug hypersensitivity reactions (DHRs) to ICM.

Objective: To refine the current view on allergic DHRs to ICM by analyzing ST reactivity patterns in patients with previous reactions to ICM.

Methods: Patients with a history of DHR to ICM and positive STs, who presented at the University Hospital of Montpellier between 2004 and 2022, were included in the study. The relative difference between every two ICM products was measured by Manhattan distance and odds ratios were computed for all pairs of products in the immediate reaction (IR) and non-immediate reaction (NIR) ST groups.

Results: A total of 181 patients were included in the study. Odds ratio analysis identified significant associations between classical cross-reactive ICM, such as iohexol-ioversol, iohexol-iomeprol, iomeprol-ioversol, and iohexol-iodixanol in the IR ST group and iohexol-ioversol, iopromide-iohexol, and iomeprol-ioversol in the NIR ST group. We also identified uncommon associations, such as ioxitalamate-amidotrizoate in the IR ST group and amidotrizoate-iopamidol and amidotrizoate-ioxitalamate in the NIR ST group. The results were reflected by the Manhattan distance, which suggested the existence of clusters containing the same classically associated ICM as well as uncommon associations, which we hypothesize to be related to similarities in the 3D structure of the respective ICM.

Conclusions: Current chemical (2D) classifications cannot explain all observed ST reactivity patterns. Whether the 3D structure can be integrated into the current classifications to interpret the observed ST reactivity patterns and predict tolerance to alternative ICM requires further research.

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

Participants: Julie Josse, Remi Khellaf.

- Title: Finite sample behavior of instrumental variables methods in causal inference

  We provide a comprehensive theoretical and empirical exploration of the integration of instrumental variables (IV) in causal analysis. We focus on the estimation of the Average Treatment Effect (ATE) when confronted with the challenge of unmeasured confounding variables. In addition, we detail a more flexible nonparametric approach that facilitates the computation of the Local Average Treatment Effect (LATE). This method requires an additional assumption, monotonicity, ensuring a monotonous relationship between treatment and the instrumental variable, and integrates it within the framework of Principal Stratification.

- Company: Quinten Health

- Duration: Feb 2023 - Aug 2023

Participants: Julie Josse, Helene Bonneau–Chloup, Pauline Bian.

- Title: Policy learning for personalized medicine. Finding the optimal dose of hormone for ovarian stimulation

  Infertility affects 1 in 5 couples of childbearing age. The most common solution is to resort to In Vitro Fertilization. However, the first challenge is to determine the initial dose and duration of gonadotropin hormone administration to maximize the number of oocytes obtained at the end of stimulation, under the constraint that estradiol levels must not be too high to avoid hyperstimulation. The second challenge is to determine the ideal day for ovulation induction, to maximize the number of oocytes retrieved, and this is done by looking at the biological results of each monitoring. To tackle these two challenges, we will leverage rich observational multi-centric and longitudinal data as well as techniques of causal inference. More precisely, we will consider

methods for learning optimal treatment policies and in particular for establishing the appropriate dose and duration of treatment for each patient. One of the challenges will be to propose methods to manage missing data in this framework. We will also consider techniques of dynamic treatment regimes to enrich the analysis with monitoring data, especially regarding hormone levels.

- Company: Elixir

- Duration: Feb 2023 -

Participants: Pascal Demoly.

- Participation to the Fondation TEZOS (Vigicard digital health card project) with the startup CodInsight

- Co-creation of the startup AdviceMedica (collective intelligence for solving complex cases in medicine)

Participants: Nicolas Molinari, Aurélien Bellet.

- Title: Learning and statistical modeling in sleep medicine: from diagnosis to treatment

  This study focuses on obstructive sleep apnea syndrome (OSAS) within the context of sleep medicine. OSAS, characterized by frequent interruptions or reductions in ventilation during sleep, is associated with anatomical collapse of upper airways. Clinical manifestations include drowsiness, snoring, and various health issues such as cardiovascular and metabolic comorbidities. Continuous positive airway pressure (PPC)) is the standard treatment. The project aims to enhance patient care by utilizing diagnostic data from university hospitals and telemonitoring data from service providers. To address diverse data formats and property issues, an extension of the study proposes using federated learning models.

- Company: groupe Adène

- Duration: May 2020 - May 2023

9.2 Bilateral Grants with Industry

Participants: Julie Josse, Pascal Demoly.

- Title: Combining RCT and observational data. Educational Grant

- Company: Allergologisk Laboratorium Kobenhavn (ALK)

- Duration: Sept 2022 - Sept 2023

Participants: Nicolas Molinari.

- Title: Study IRIS - **Real-life effectiveness of the Chronic Care ConnectTM remote medical monitoring system in patients with chronic heart failure**

- Company: IQVIA

- Duration: Oct 2022 - Oct 2023

- Title: **real-life study analyzing mortality in severe COPD patients**

- Company: Sanofi

- Duration: Nov 2022 - Dec 2023

- Title: **study « Home-Care SIMEOX »**,

- Company: Agir à Dom

- Duration: Nov 2019 - Nov 2023

- Title: **RESALA**

- Company: GSK

- Duration: 2023

_____

10 Partnerships and cooperations

Participants: Julie Josse, Margaux Zaffran, Pan Zhao, Maxime Fosset.

10.1 International research visitors

10.1.1 Visits of international scientists

Nicolas W Hengartner

Status Senior Researcher

Institution of origin: Los Alamos National Laboratory

Country: USA

Dates: December 2023

Context of the visit: Work with Nicolas Molinari and discussion of Meta-Analysis and Federated learning with Julie Josse, Aurélien Bellet and Nicolas Molinari

Mobility program/type of mobility: research stay

10.1.2 Visits to international teams

Research stays abroad

Margaux Zaffran

Visited institution: University of Stanford

Country: United States

Dates: August 9, 2023 - August 18, 2023

Context of the visit: hosted by Madeleine Udell

Mobility program/type of mobility: research stay

Visited institution: University of Berkeley

Country: United States

Dates: August 21, 2023 - August 25, 2023

Context of the visit: visiting the Statistics department

Mobility program/type of mobility: research stay

Maxime Fosset

Visited institution: Beth Israel Deaconess Medical Center, Harvard Medical School

Country: United States

Dates: January 12, 2023 - April 30, 2023

Context of the visit: Conducting a project using causal inference tools on the local ICU patients database: the VASOWEAN study. The aim of this project is to conduct an emulated target trial on an observational database to estimate the effect of vasopressors, a class of drug, on the success rate of extubation in the intensive care unit. To estimate the average treatment effect, several techniques are used such as IPTW, G-formula and causal forest models.

Mobility program/type of mobility: Research stay

Visited institution: Beth Israel Deaconess Medical Center, Harvard Medical School, Harvard T.H Chan School of Public Health

Country: United States

Dates: June 20, 2023 - August 1, 2023

Context of the visit: CAUSALAB Summer Course on Advanced Confounding Adjustment and Target Trial Emulation

Mobility program/type of mobility: Research stay and Course

Pan Zhao

Visited institution: Ghent University

Country: Belgium

Dates: October, 2023 - November, 2023

Context of the visit: Research collaboration with Stijn Vansteelandt

Mobility program/type of mobility: Research stay to work on nonparametric instrumental variables in causal inference with continuous treatment.

10.2   European initiatives

10.2.1   Other european programs/initiatives

- Julie Josse Advisory Board of HORIZON EUROPE(HORIZON-HLTH-2022-TOOL-11-02), more-europa The aim of the project is to develop, implement and establish evidentiary standards and methods to address the data and evidentiary needs of regulatory authorities and health technology assessment (HTA) bodies towards a more efficient use of Real Word Data for the development, registration and assessment of medicinal products in Europe.

10.3   National initiatives

10.3.1   PEPR Digital Health

The "PEPR Santé Numérique", launched in June 2023 as part of the Plan Innovation Santé 2030, is a major initiative in the "Digital Health" acceleration strategy with a program dedicated to stimulating scientific research in this field.

PreMeDICaL is involved in three projects that have been lauched:

- SMATCH "Statistical and AI Methods for the Challenges of Modern Clinical Trials in Digital Health" - Julie Josse, Pascal Demoly

  - New clinical trial methods and designs based on animal-to-human, research-based disease models,
  - Enriching clinical trials with multi-source, multi-dimensional ancillary data,
  - Next-generation designs for clinical evaluation of digital medical devices based on AI algorithms,
  - Regulation, feasibility and dissemination of clinical trials

- Digital Pharmacological Twins "Multi-scale and longitudinal data modelling in pharmacology: toward digital pharmacological twins" - Julie Josse

- Secure, safe and fair machine learning for healthcare - Aurélien Bellet

10.3.2   PEPR Cybersecurity

PreMeDICaL is involved in project IPoP (Interdisciplinary Project on Privacy) - Aurélien Bellet. The objectives of this project are to study the threats on privacy that have been introduced by these new services, and to conceive theoretical and technical privacy-preserving solutions that are compatible with French and European regulations, that preserve the quality of experience of the users. These solutions will be deployed and assessed, both on the technological and legal sides, and on their societal acceptability. In order to achieve these objectives, we adopt an interdisciplinary approach, bringing together many diverse fields: computer science, technology, engineering, social sciences, economy and law.

The project's scientific program focuses on new forms of personal information collection, on the learning of Artificial Intelligence (AI) models that preserve the confidentiality of personal information used, on data anonymization techniques, on securing personal data management systems, on differential privacy, on personal data legal protection and compliance, and all the associated societal and ethical considerations. This unifying interdisciplinary research program brings together internationally recognized research teams (from universities, engineering schools and institutions) working on privacy, and the French Data Protection Authority (CNIL).

This holistic vision of the issues linked to personal data protection will on one hand let us propose solutions to the scientific and technological challenges and on the other help, us confront these solutions in many different ways, in the context of interdisciplinary collaborations, thus leading to recommendations and proposals in the field of regulations or legal frameworks. This comprehensive consideration of all the issues aims at encouraging the adoption and acceptability of the solutions proposed by all stakeholders, legislators, data controllers, data processors, solution designers, developers all the way to end-users.

10.3.3 Inria Challenge FedMalin

Aurélien Bellet leads FedMalin. FedMalin is a research project that spans 11 Inria research teams and aims to push FL research and concrete use-cases through a multidisciplinary consortium involving expertise in ML, distributed systems, privacy and security, networks, and medicine. We propose to address a number of challenges that arise when FL is deployed over the Internet, including privacy & fairness, energy consumption, personalization, and location/time dependencies. FedMalin will also contribute to the development of open-source tools for FL experimentation and real-world deployments, and use them for concrete applications in medicine and crowdsensing.

The FedMalin Inria Challenge is supported by Groupe La Poste, sponsor of the Inria Foundation.

10.3.4 ANR JCJC PRIDE

Aurélien Bellet leads PRIDE, a JCJC ANR project on privacy-preserving decentralized machine learning. The goal of PRIDE is to develop theoretical and algorithmic tools that enable differentially-private ML methods operating on decentralized datasets, through three complementary objectives:

- Prove that decentralized learning protocols naturally amplify DP guarantees;

- Propose algorithms at the intersection of decentralized ML and secure multi-party computation;

- Design data-adaptive communication schemes to speed up the convergence on heterogeneous datasets.

10.3.5 Allergen-Chip-Challenge

The challenge L'allergen-chip-challenge aimed at creating a national dataset for artificial intelligence-assisted allergy diagnosis using semantic attributes and allergen multiplex technology. The challenge was supported by the Health Data Hub in collaboration with the company Trustee - Pascal Demoly

Two follow-up projects:

- grant PNRIA 2023 with Olivier Saut

- AAP MESSIDORE 2023 submitted, Pascal Demoly and Julie Josse lead one research axis

10.3.6 Grant from the National Interministerial Road Safety Observatory

Julie Josse - In collaboration with Traumabase. Grant for the SPOTE project (Specificities of Populations and Impact of Territories) aimed at studying the intra-hospital outcome of victims of road accidents treated, in critical care, in France, between 2013 and 2027.

10.3.7 Grant from PHRC

Nicolas Molinari leads 3 work packages

- Evaluation of early venous stenting treatment of patients with newly diagnosed idiopathic intracranial hypertension

- Evaluation of venous stenting treatment of patients with idiopathic intracranial hypertension to pursue acetazolamide withdrawal

- REVERT - Reversing airway remodeling with Tezepelumab

10.3.8 Grant from Institut Exposum Doctoral Nexus

Nicolas Molinari obtained a grant from Doctoral Nexus for Phd students on modeling the effect of plastic nanoparticles on respiratory health and epidemiological effects of low-emission zones on human health

10.4   Regional initiatives

Julie Josse, Nicolas Molinari, Pascal Demoly are in the scientific committee of "IA for health", supported by Occitanie Region and Aniti (Toulouse)

11   Dissemination

11.1   Promoting scientific activities

11.1.1   Scientific events: organisation

- Margaux Zaffran: Organization of the events of the Young Statisticians group during the Journées de la Société Française de Statistique, June 2023, Brussels, Belgium

- Margaux Zaffran: 11th Young Statisticians and Probabilists day, January 2023, Institut Henri Poincaré, Paris, France

- Aurélien Bellet: co-organization of the Federated Learning One World webinar (1100+ registered attendees) since May 2020.

- Julie Josse, Jeffrey Naf, Pan Zhao: organization of the missing data and causality group

11.1.2   Scientific events: selection

Member of the conference program committees

- Julie Josse: IMS International Conference on Data Science, Lisbonne, Portugal, December 2023.

- Julie Josse: Journées de la Société Française de Statistique, Brussels, Belgium 2023.

- Julie Josse: Statlearn, Montpellier, France, April 2023.

- Aurélien Bellet: Area Chair for Neural Information Processing Systems, NeurIPS 2023

- Aurélien Bellet: Area Chair for Artificial Intelligence and Statistics, AISTATS 2024

Reviewer

- Aurélien Bellet: ecure and Trustworthy Machine Learning, SaTML 2024

- Aurélien Bellet: Federated learning for training and tuning foundation models, FL@FM-NeurIPS'23

- Aurélien Bellet: Workshop on Privacy-Preserving Artificial Intelligence, PPAI@AAAI 2024

- Margaux Zaffran: AISTATS, Valencia, Spain, April 2023 (top 10% reviewer)

- Margaux Zaffran: NeurIPS, New Orleans, United States, December 2023

11.1.3   Journal

Member of the editorial boards

- Aurélien Bellet is Action Editor for Transactions of Machine Learning Research (TMLR).

11.1.4 Invited talks

- Julie Josse: Journées de biostatistique November 2023, Toulouse.

- Julie Josse: Labex NUMEV, Montpellier, Feb. 2023.

- Aurélien Bellet: Distributed ML workshop, Paris, December 2023.

- Margaux Zaffran: Institut Mathématiques de Toulouse – Statistics and Optimization Seminar, Toulouse, France, October 2023.

- Margaux Zaffran: FAST-BIG – Statistics Workshop, Paris, France, October 2023.

- Margaux Zaffran: ENBIS Annual European Conference, Valencia, Spain, September 2023.

- Margaux Zaffran: Journées de la société Française de Statistique, Brussels, Belgium, July 2023.

- Margaux Zaffran: INRIA – MIND Seminar, Saclay, France, June 2023.

- Margaux Zaffran: Agro ParisTech and INRAE – MIA Seminar, Toulouse, France, February 2023.

- Marie Felicia Beclin: International Conference on Statistics and Data Science, ICSDS 2023 Lisbonne, Dec. 2023.

- Jeffrey Naf: Statistic Seminar, Lund, November 2023.

- Jeffrey Naf: Statistic Seminar, Sorbonne University, Paris, November 2023.

- Pan Zhao: Statistic Seminar, Ghent, December 2023.

- Pan Zhao: Joint Statistical Meetings, Ontario (Online), August 2023.

- Pan Zhao: When Causal Inference meets Statistical Analysis, Conservatory of Arts and Crafts, Paris, April 2023. Contributed talk.

11.1.5 Leadership within the scientific community

- Margaux Zaffran: president of the Young Statisticians group of the French Statistical Society

- Pascal Demoly: president of the "Société Française d'Allergologie"

- Pascal Demoly: Animation of the network e-allergies

- Julie Josse elected as a member of the R foundation and of the R Foundation Conference Committee. She is in the board of the French R committee (organization for coordinating R conferences "Les rencontres R") and involved in a task Forwards force on behalf of the R Foundation with the aim of increasing the participation of women and under-represented groups in the STEM community (founding member in 2015).

11.1.6 Scientific expertise

- Julie Josse, Nicolas Molinari: member of the CSE of the Montpellier University Hospital (Comité scientifique et éthique du CHU de Montpellier). December 2023

- Aurélien Bellet: review of G7 case study on synthetic data, October-November 2023

- Aurélien Bellet: ethics advisor for the European Strategy Forum on Research Infrastructures (ESFRI) project SLICES-PP

- Nicolas Molinari: jury member for HCERES

- Nicolas Molinari: president of the Institutional Review Board (IRB) of the Adène group

11.1.7   Research administration

- Aurélien Bellet: member of the Operational Committee for the assessment of Legal and Ethical risks (COERLE).

- Julie Josse: member of CSD ("Comité Suivi Doctoral")

- Nicolas Molinari: elected member of "Commissions scientifiques spécialisées" (CSS) 6 of INSERM

11.2   Teaching - Supervision - Juries

11.2.1   Teaching

- Master: ESEEC, 19.5eqTD, Les outils statistiques du diagnostic (modélisation et statistiques économiques et sociales), Université Paul Valery, Montpellier - Pierre Lafaye De Micheaux

- Master: ESEEC, 17.5eqTD, Analyse de Données Multidimensionnelles SHS, Université Paul Valery, Montpellier - Pierre Lafaye De Micheaux

- Master: MIASHS, 22.5eqTD, Statistique et probabilités bivariées, Université Paul Valery, Montpellier - Pierre Lafaye De Micheaux

- Bachelor: MIASHS, 26eqTD, Science des données 2, Université Paul Valery, Montpellier - Pierre Lafaye De Micheaux

- Master: MIASHS, head of projects of " Travaux d'Études et de Recherche (TER)", Université Paul Valery, Montpellier - Pierre Lafaye De Micheaux

- Master: MIASHS, head of "Marathon du Web", Université Paul Valery, Montpellier - Pierre Lafaye De Micheaux

- Master: Institut de formation en masso-kinésithérapie, 9eqTD, statistics, Montpellier - Nicolas Molinari

- Master: Institut de formation en masso-kinésithérapie, head of the program, statistics, Montpellier - Nicolas Molinari

- Ecoles d'étiopathie, head of the program, statistics, Montpellier - Nicolas Molinari

- Master: EDSB « Epidémiologie, Données de Santé, Biostatistique », head of « Grands enjeux en santé » , Université de Montpellier - Pascal Demoly

- Engineering School: First year (equivalent to L3), 13eqTD, Introduction to probability, ENSTA Paris – Institut Polytechnique de Paris - Margaux Zaffran

- Engineering School: First year (equivalent to L3), 14eqTD, Introduction to statistics, ENSTA Paris – Institut Polytechnique de Paris - Margaux Zaffran

11.2.2   Supervision

- Postdoc: Batiste Le Bars, since Jun 2023. Supervised by Aurélien Bellet, Kevin Scaman, Giovanni Neglia and Marc Tommasi

- PhD defended in Oct 2022: Paul Mangold. Supervised by Aurélien Bellet

- PhD in progress: Gaurav Maheshwari, since Nov 2020. Supervised by Aurélien Bellet, Pascal Denis and Mikaela Keller

- CIFRE PhD in progress: Jean-Remy Conti, since Sep 2021. Supervised by Aurélien Bellet, Stéphan Clémençon and Vincent Despiegel.

- PhD in progress: Edwige Cyffers, since Oct 2021. Supervised by Aurélien Bellet

- PhD in progress: Tudor Cebere, since Nov 2022. Supervised by Aurélien Bellet

- CIFRE PhD in progress: Clément Pierquin, since Jun 2023. Supervised by Aurélien Bellet, Marc Tommasi and Matthieu Boussard

- PhD in progress: Brahim Erraji, since Sep 2023. Supervised by Aurélien Bellet, Catuscia Palamidessi and Mickael Perrot.

- PhD in progress: Remi Khellaf, since Oct 2023. Supervised by Aurélien Bellet and Julie Josse

- Engineer: Paul Andrey. Supervised by Aurélien Bellet

- PhD in progress: Ahmed Boughdiri, since Sep 2023. Supervised by Julie Josse and Erwan Scornet

- PhD in progress: Pan Zhao, since Sep 2021. Supervised by Julie Josse and Antoine Chambaz

- PhD in progress: Maxime Fosset, since Sep 2022. Supervised by Julie Josse and Nicolas Molinari

- CIFRE EDF PhD in progress: Margaux Zaffran, since Sep 2021. Supervised by Julie Josse, Yannig Goude, Olivier Ferron and Aymeric Dieuleveut

- CIFRE Sanofi PhD in progress: Charlotte Voinot, since Sep 2023. Supervised by Julie Josse and Bernard Sebastien

- Postdoc: Herbert Susmann, since Jan 2023. Supervised by Antoine Chambaz and Julie Josse

- Postdoc: Jeffrey Naf, since Jan 2023. Supervised by Julie Josse

- Postdoc: Houssam Zenati, since Dec 2023. Supervised by Bertrand Thirion, Judith Abecassis and Julie Josse

- CIFRE Adene group PhD Celia Vidal. Supervised by Nicolas Molinari.

- PhD in Progress: Martin Puig, since Sep 2023. Supervised by Nicolas Molinari.

- PhD defended in Dec 2023: F Bertelli. Supervised by Nicolas Molinari.

- PhD in Progress: Marie Felicia Beclin, since Sep 2022. Supervised by Pierre Lafaye De Micheaux and Nicolas Molinari

11.2.3   Juries

- Aurélien Bellet: reviewer for the PhD thesis of Christian Lebeda (Univ Copenhagen), November 2023

- Aurélien Bellet: reviewer for the PhD thesis of Vincent Plassier (IP Paris), October 2023

- Aurélien Bellet: reviewer for the PhD thesis of Sayan Biswas (IP Paris), October 2023

- Aurélien Bellet: reviewer for the PhD thesis of Clément Lalanne (ENS Lyon), October 2023

- Aurélien Bellet: examiner for the PhD thesis of Fatima El Hattab (Univ Lyon), November 2023

- Julie Josse: examiner for the PhD thesis of Patrick Saux (Lille), December 2023

- Julie Josse: examiner for the PhD thesis of Giulia Marchello (Nice)

- Julie Josse: reviewer for the PhD thesis of François Grolleau (CRESS Paris), November 2023

- Julie Josse: reviewer for the PhD thesis Lucas Etourneau (Grenoble)

11.3 Popularization

11.3.1 Articles and contents

- Aurélien Bellet was interviewed for an article "La mainmise des géants de la tech sur la recherche en intelligence artificielle" in French newspaper La Croix (13/09/2023).

- Pascal Demoly Interview in "la tribune Occitanie": Le CHU de Montpellier et Codinsight créent une application pour tracer les allergies médicamenteuses

11.3.2 Interventions

- Lycée Marseilleveyre, November 2023 - Margaux Zaffran

- For Girls in Science, Cité des Sciences et de l'Industrie, October 2023 - Margaux Zaffran

- Séphora Berrebi Association, MasterClass for high school girls, April 2023 - Margaux Zaffran

12 Scientific production

12.1 Major publications

[1] A. Bourdin, S. Bommart, G. Marin, I. Vachier, A. S. Gamez, E. Ahmed, C. Suehs and N. Molinari. 'Obesity in women with asthma: baseline disadvantage plus greater small-airway responsiveness'. In: *Allergy* (2022). DOI: 10.1111/all.15509. URL: https://hal.science/hal-03772214.

[2] B. Colnet, J. Josse, G. Varoquaux and E. Scornet. *Risk ratio, odds ratio, risk difference... Which causal measure is easier to generalize?* 2023. URL: https://hal.science/hal-04369607.

[3] B. Colnet, I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse and S. Yang. 'Causal inference methods for combining randomized trials and observational studies: a review'. In: *Statistical Science* (2024). URL: https://hal.science/hal-03008276.

[4] J. Josse, N. Prost, E. Scornet and G. Varoquaux. *On the consistency of supervised learning with missing values.* 29th June 2020. URL: https://hal.science/hal-02024202.

[5] M. Le Morvan, J. Josse, E. Scornet and G. Varoquaux. 'What's a good imputation to predict with missing values?' In: NeurIPS 2021 - 35th Conference on Neural Information Processing Systems. Virtual, France, 6th Dec. 2021. URL: https://hal.science/hal-03243931.

[6] I. Mayer, A. Sportisse, N. Tierney, N. Vialaneix and J. Josse. 'R-miss-tastic: a unified platform for missing values methods and workflows'. In: *The R Journal* (30th July 2022). URL: https://hal.archives-ouvertes.fr/hal-02879337.

[7] I. Mayer, E. Sverdrup, T. Gauss, J.-D. Moyer, S. Wager and J. Josse. 'Doubly robust treatment effect estimation with missing attributes'. In: *Annals of Applied Statistics* 14.3 (Sept. 2020), pp. 1409–1431. DOI: 10.1214/20-AOAS1356. URL: https://hal.science/hal-02879332.

[8] F. Roubille, E. Matzner-Lober, S. Aguilhon, M. Rene, L. Lecourt, M. Galinier, J.-e. Ricci and N. Molinari. 'Impact of global warming on weight in patients with heart failure during the 2019 heatwave in France'. In: *ESC Heart Failure* (2022). DOI: 10.1002/ehf2.14129. URL: https://hal.science/hal-03783874.

[9] M. Zaffran, A. Dieuleveut, J. Josse and Y. Romano. 'Conformal Prediction with Missing Values'. In: *Proceedings of Machine Learning Research*. ICML 2023 - 40 th International Conference on Machine Learning. Vol. PMLR202. Honolulu (Hawai), United States, 23rd July 2023, p. 40578. URL: https://hal.science/hal-03896384.

12.2   Publications of the year

International journals

[10]   S. Achard, J.-F. Coeurjolly, P. L. de Micheaux, H. Lbath and J. Richiardi. 'Inter-regional correlation estimators for functional magnetic resonance imaging'. In: *NeuroImage* 282 (Nov. 2023), p. 120388. DOI: 10.1016/j.neuroimage.2023.120388. URL: https://hal.science/hal-04242995.

[11]   A. Fillard, A. Licari, N. Molinari, G. Marseglia, P. Demoly and D. Caimmi. 'Sensitivity of FEV1 and Clinical Parameters in Children With a Suspected Asthma Diagnosis'. In: *Journal of Allergy and Clinical Immunology: In Practice* 11.1 (Jan. 2023), pp. 238–247. DOI: 10.1016/j.jaip.2022.10.0 11. URL: https://hal.umontpellier.fr/hal-03930657.

[12]   D. Jaffuel, Y. Bouchaut, J.-P. Mallet, C. Vidal, N. Molinari, A. Bourdin and F. Roubille. 'Dapagliflozin initiation in chronic heart failure patients improves central sleep apnoea'. In: *ERJ Open Research* 9.3 (2023), pp. 00123–2023. DOI: 10.1183/23120541.00123-2023. URL: https://hal.science /hal-04145196.

[13]   E. Macy, A. Trautmann, A. M. Chiriac, P. Demoly and E. Phillips. 'Advances in the Understanding of Drug Hypersensitivity: 2012 Through 2022'. In: *Journal of Allergy and Clinical Immunology: In Practice* 11.1 (Jan. 2023), pp. 80–91. DOI: 10.1016/j.jaip.2022.10.025. URL: https://hal.um ontpellier.fr/hal-03930648.

[14]   L. Martrille, S. Papadodima, C. Venegoni, N. Molinari, D. Gibelli, E. Baccino and C. Cattaneo. 'Age Estimation in 0–8-Year-Old Children in France: Comparison of One Skeletal and Five Dental Methods'. In: *Diagnostics* 13.6 (Mar. 2023), p. 1042. DOI: 10.3390/diagnostics13061042. URL: https://hal.science/hal-04128147.

[15]   A. Salsano, A. Nenna, N. Molinari, S. S. Avtaar Singh, C. Spadaccio, F. Santini, M. Chello, A. Fiore and F. Nappi. 'Impact of Mitral Regurgitation Recurrence on Mitral Valve Repair for Secondary Ischemic Mitral Regurgitation'. In: *Journal of Cardiovascular Development and Disease* 10.3 (Mar. 2023), p. 124. DOI: 10.3390/jcdd10030124. URL: https://hal.science/hal-04105466.

International peer-reviewed conferences

[16]   C. Biernacki, G. Celeux, J. Josse, F. Laporte, M. Marbac, A. Sportisse, V. Vandewalle and C. Boyer. 'Impact of missing data on mixtures and clustering with illustrations in Biology and Medicine'. In: SPSR 2023 - The 24th annual Conference of the Romanian Society of Probability and Statistics. Bucarest, Romania, 21st Apr. 2023. URL: https://inria.hal.science/hal-04370870.

[17]   M. Zaffran, A. Dieuleveut, J. Josse and Y. Romano. 'Conformal Prediction with Missing Values'. In: *Proceedings of Machine Learning Research*. ICML 2023 - 40 th International Conference on Machine Learning. Vol. PMLR202. Honolulu (Hawai), United States, 23rd July 2023, p. 40578. URL: https://hal.science/hal-03896384.

Reports & preprints

[18]   C. Bénard and J. Josse. *Variable importance for causal forests: breaking down the heterogeneity of treatment effects.* 4th Aug. 2023. URL: https://hal.science/hal-04177493.

[19]   C. Bénard, J. Naf and J. Josse. *MMD-based Variable Importance for Distributional Random Forest.* 2023. URL: https://hal.science/hal-04251588.

[20]   B. Colnet, J. Josse, G. Varoquaux and E. Scornet. *Risk ratio, odds ratio, risk difference... Which causal measure is easier to generalize?* 2023. URL: https://hal.science/hal-04369607.

[21]   B. Colnet, I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse and S. Yang. *Causal inference methods for combining randomized trials and observational studies: a review.* 10th Jan. 2023. URL: https://hal.science/hal-03008276.

[22]   H. Hendrikx, P. Mangold and A. Bellet. *The Relative Gaussian Mechanism and its Application to Private Gradient Descent.* 29th Aug. 2023. URL: https://hal.science/hal-04370596.

[23] C. Pierquin, A. Bellet, M. Tommasi and M. Boussard. *Rényi Pufferfish Privacy: General Additive Noise Mechanisms and Privacy Amplification by Iteration via Shift Reduction Lemmas.* 21st Dec. 2023. URL: https://inria.hal.science/hal-04363020.

[24] A. Sportisse, M. Marbac, F. Laporte, G. Celeux, C. Boyer, C. Biernacki and J. Josse. *Accompanying note : Model-based Clustering with Missing Not At Random Data.* 21st Dec. 2023. URL: https://hal.science/hal-04358192.

[25] A. Sportisse, M. Marbac, F. Laporte, G. Celeux, C. Boyer, J. Josse and C. Biernacki. *Model-based Clustering with Missing Not At Random Data.* 21st Dec. 2023. URL: https://hal.science/hal-03494674.

[26] H. Susmann, A. Chambaz and J. Josse. *AdaptiveConformal: An R Package for Adaptive Conformal Inference.* 30th Nov. 2023. URL: https://hal.science/hal-04316544.

[27] P. Zhao, A. Chambaz, J. Josse and S. Yang. *Positivity-free Policy Learning with Observational Data.* 2023. URL: https://hal.science/hal-04236430.

[28] P. Zhao and Y. Cui. *A Semiparametric Instrumented Difference-in-Differences Approach to Policy Learning.* 14th Oct. 2023. URL: https://hal.science/hal-04242007.

12.3 Cited publications

[29] I.-M. Ghiordanescu, N. Molinari, I. C. nea-Teodorescu, R. Schrijvers, C. Motei, A.-M. Forsea, P. Demoly and A. M. Chiriac. 'Skin Test Reactivity Patterns in Patients Allergic to Iodinated Contrast Media: A Refined View'. In: *The Journal of Allergy and Clinical Immunology: In Practice* (2023). DOI: \url{https://doi.org/10.1016/j.jaip.2023.10.040}. URL: https://www.sciencedirect.com/science/article/pii/S2213219823011959.

[30] P. Kairouz, H. B. McMahan, B. Avent, **Bellet, A.**, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings et al. 'Advances and Open Problems in Federated Learning'. In: *Foundations and Trends® in Machine Learning* 14.1–2 (2021), pp. 1–210.

[31] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani and L. Wasserman. 'Distribution-Free Predictive Inference for Regression'. In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1094–1111.

[32] L. Pahus, D. Jaffuel, I. Vachier, A. Bourdin, C. M. Suehs, N. Molinari and P. Chanez. 'Randomised controlled trials in severe asthma: selection by phenotype or stereotype'. In: *European Respiratory Journal* 53.2 (2019).

[33] B. Paige, J. Bell, **Bellet, A.**, A. Gascón and D. Ezer. 'Reconstructing Genotypes in Private Genomic Databases from Genetic Risk Scores'. In: *Journal of Computational Biology* 28.5 (2021), pp. 435–451.

[34] H. Papadopoulos, K. Proedrou, V. Vovk and A. Gammerman. 'Inductive Confidence Machines for Regression'. In: *Machine Learning: ECML 2002*. Springer, 2002, pp. 345–356.

[35] Y. Romano, E. Patterson and E. Candès. 'Conformalized Quantile Regression'. en. In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019. URL: https://papers.nips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html.

[36] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian and J. Vertesi. 'Fairness and Abstraction in Sociotechnical Systems'. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019, pp. 59–68.

[37] R. Shokri, M. Stronati, C. Song and V. Shmatikov. 'Membership Inference Attacks Against Machine Learning Models'. In: *IEEE Symposium on Security and Privacy*. 2017.

[38] V. Vovk, A. Gammerman and G. Shafer. *Algorithmic Learning in a Random World.* en. Springer US, 2005.

[39] M. B. Zafar, I. Valera, M. Gomez-Rodriguez and K. P. Gummadi. 'Fairness Constraints: A Flexible Approach for Fair Classification'. In: *Journal of Machine Learning Research* 20.75 (2019), pp. 1–42.