Simulation en Recherche Clinique
Généralités, exemple et problème connexe

Nicolas SAVY

Institut de Mathématiques de Toulouse

Séminaire MIAT - INRAE - Toulouse

Le 15 mai 2020

Content of the talk

1. Generalities on Simulated Clinical Trials

2. What an In Silico Clinical Trial is?

3. An example in Economics Health

4. Databases Merging issue

Contexte

Projet **”Big Data en Santé”** 2016-2018

- Financé par la Région Occitanie
- Porté par l’IMT (NS) et l’IFERISS (T. Lang)
- Projet interdisciplinaire
- Objectif : données massives en santé

IFERISS
Institut Fédératif d’Etudes et de Recherches Interdisciplinaires Santé Société

INSTITUT de MATHEMATIQUES de TOULOUSE

Contexte

Projet **"Big Data en Santé"** 2016-2018

- Financé par la Région Occitanie
- Porté par l'IMT (NS) et l'IFERISS (T. Lang)
- Projet interdisciplinaire
- Objectif : données massives en santé
- Axe : **Essai clinique simulé**

IFERISS

Institut Fédératif d'Etudes et de Recherches Interdisciplinaires Santé Société

INSTITUT de MATHEMATIQUES de TOULOUSE

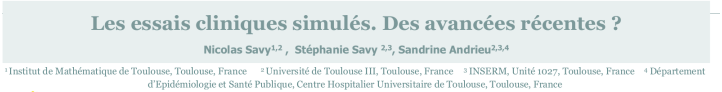
Contexte

Projet **"Big Data en Santé"** 2016-2018

- Financé par la Région Occitanie
- Porté par l'IMT (NS) et l'IFERISS (T. Lang)
- Projet interdisciplinaire
- Objectif : données massives en santé
- Axe : **Essai clinique simulé**

IFERISS
Institut Fédératif d'Etudes et de Recherches Interdisciplinaires Santé Société

INSTITUT de MATHEMATIQUES de TOULOUSE



Les essais cliniques simulés. Des avancées récentes ?

Nicolas Savy[1,2] , Stéphanie Savy [2,3], Sandrine Andrieu[2,3,4]

[1] Institut de Mathématique de Toulouse, Toulouse, France [2] Université de Toulouse III, Toulouse, France [3] INSERM, Unité 1027, Toulouse, France [4] Département d'Epidémiologie et Santé Publique, Centre Hospitalier Universitaire de Toulouse, Toulouse, France

Poster EpiClin 2015

Contexte

Projet **”Big Data en Santé”** 2016-2018

- Financé par la Région Occitanie
- Porté par l’IMT (NS) et l’IFERISS (T. Lang)
- Projet interdisciplinaire
- Objectif : données massives en santé
- Axe : **Essai clinique simulé**

IFERISS
Institut Fédératif d’Etudes et de Recherches Interdisciplinaires Santé Société

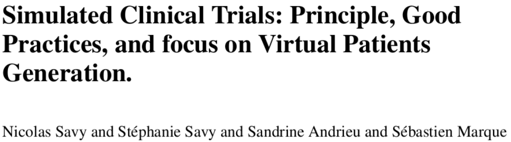
INSTITUT de MATHEMATIQUES de TOULOUSE

Les essais cliniques simulés. Des avancées récentes ?

Nicolas Savy[1,2] , Stéphanie Savy [2,3], Sandrine Andrieu[2,3,4]

[1] Institut de Mathématique de Toulouse, Toulouse, France [2] Université de Toulouse III, Toulouse, France [3] INSERM, Unité 1027, Toulouse, France [4] Département d’Epidémiologie et Santé Publique, Centre Hospitalier Universitaire de Toulouse, Toulouse, France

Poster EpiClin 2015



Simulated Clinical Trials: Principle, Good Practices, and focus on Virtual Patients Generation.

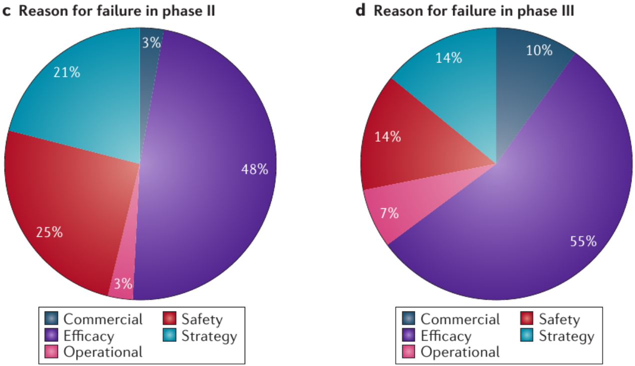
Nicolas Savy and Stéphanie Savy and Sandrine Andrieu and Sébastien Marque

Proceedings in Mathematics and Statistics, Springer Verlag Chapter 21, 2018

Content of the talk

1. Generalities on Simulated Clinical Trials
2. What an In Silico Clinical Trial is?
3. An example in Economics Health
4. Databases Merging issue

Why so many Clinical Trials fail?



- **∼70% of trial failures** due to Efficacy and Safety ((Harrison (2016), Fogel (2018)))
  - ⇒ **explanations** may be found in the Clinical Trial Design
  - ⇒ **solution** may come from optimization or at least challenging trials' designs

Why a Clinical Trial fails?

The FDA estimates that just a 10% improvement in the ability to predict drug failures before clinical trials begin could save $100 million in development costs per drug.

Dr. Amar Thyagarajan, 2015

It is appealing but how to proceed ?

A solution may come from performing In Silico Clinical Trials !

Why a Clinical Trial fails?

The FDA estimates that just a 10% improvement in the ability to predict drug failures before clinical trials begin could save $100 million in development costs per drug.

Dr. Amar Thyagarajan, 2015

It is appealing but how to proceed ?

A solution may come from performing In Silico Clinical Trials !

Content of the talk

1 Generalities on Simulated Clinical Trials

2 What an In Silico Clinical Trial is?

3 An example in Economics Health

4 Databases Merging issue

Our definition

An **In Silico Clinical Trial** is

an **agent-based model** which

uses **Virtual patients**

to **mimic** their behaviour in

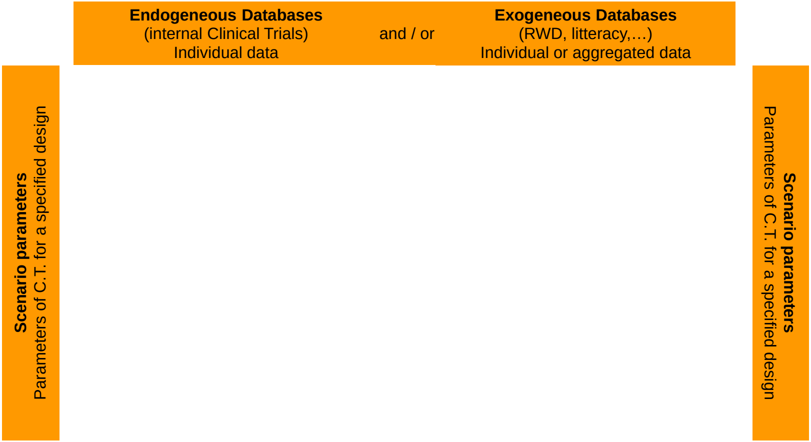
a **Virtual Clinical Trial** in order

to **challenge** trial's design in terms of

feasibility and **probability of success** of the trial.

Example of what an In Silico Clinical Trial can be?

General Schema



Endogeneous Databases
(internal Clinical Trials)
Individual data

and / or

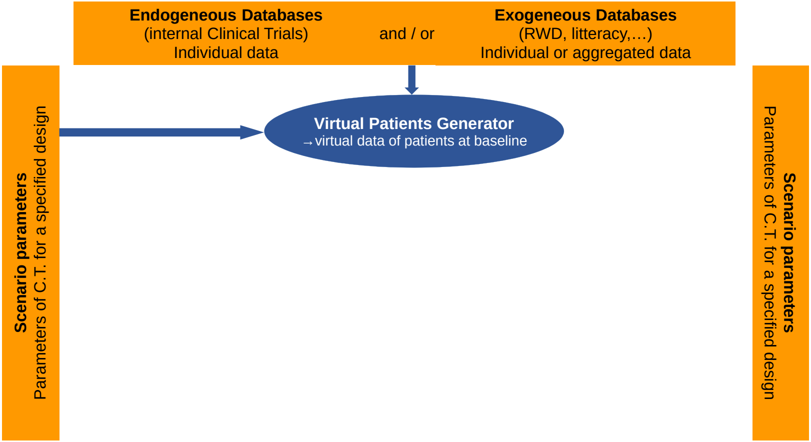
Exogeneous Databases
(RWD, litteracy,…)
Individual or aggregated data

Scenario parameters
Parameters of C.T. for a specified design

Scenario parameters
Parameters of C.T. for a specified design

Example of what an In Silico Clinical Trial can be?

General Schema



Endogeneous Databases
(internal Clinical Trials)
Individual data

and / or

Exogeneous Databases
(RWD, litteracy,…)
Individual or aggregated data

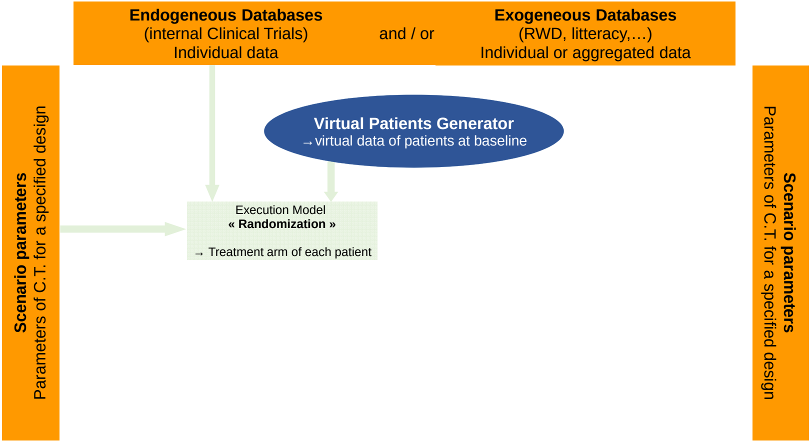
Virtual Patients Generator
→virtual data of patients at baseline

Scenario parameters
Parameters of C.T. for a specified design

Scenario parameters
Parameters of C.T. for a specified design

Example of what an In Silico Clinical Trial can be?

General Schema



Endogeneous Databases
(internal Clinical Trials)
Individual data

and / or

Exogeneous Databases
(RWD, litteracy,...)
Individual or aggregated data

Scenario parameters
Parameters of C.T. for a specified design

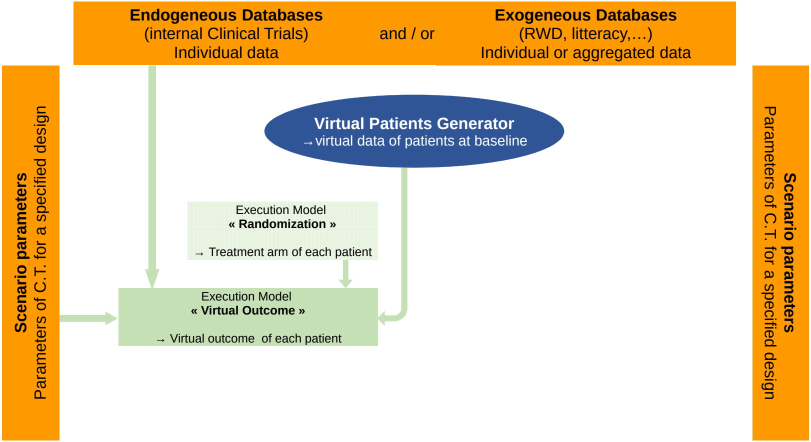
Virtual Patients Generator
→ virtual data of patients at baseline

Execution Model
« Randomization »
→ Treatment arm of each patient

Scenario parameters
Parameters of C.T. for a specified design

Example of what an In Silico Clinical Trial can be?

General Schema



Endogeneous Databases
(internal Clinical Trials)
Individual data

and / or

Exogeneous Databases
(RWD, litteracy,...)
Individual or aggregated data

Scenario parameters
Parameters of C.T. for a specified design

Virtual Patients Generator
→virtual data of patients at baseline

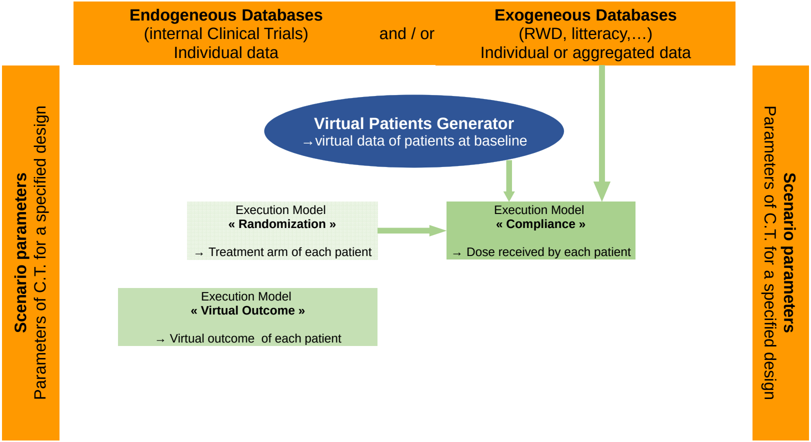
Execution Model
« Randomization »
→ Treatment arm of each patient

Execution Model
« Virtual Outcome »
→ Virtual outcome of each patient

Scenario parameters
Parameters of C.T. for a specified design

Example of what an In Silico Clinical Trial can be?

General Schema



Endogeneous Databases
(internal Clinical Trials)
Individual data

and / or

Exogeneous Databases
(RWD, litteracy,…)
Individual or aggregated data

Scenario parameters
Parameters of C.T. for a specified design

Virtual Patients Generator
→ virtual data of patients at baseline

Execution Model
« Randomization »
→ Treatment arm of each patient

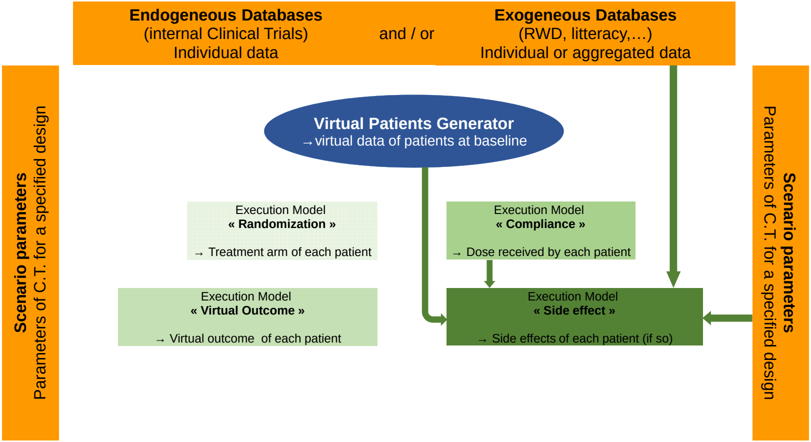
Execution Model
« Compliance »
→ Dose received by each patient

Execution Model
« Virtual Outcome »
→ Virtual outcome of each patient

Scenario parameters
Parameters of C.T. for a specified design

Example of what an In Silico Clinical Trial can be?

General Schema



Endogeneous Databases
(internal Clinical Trials)
Individual data

and / or

Exogeneous Databases
(RWD, litteracy,…)
Individual or aggregated data

Scenario parameters
Parameters of C.T. for a specified design

Virtual Patients Generator
→virtual data of patients at baseline

Execution Model
« Randomization »
→ Treatment arm of each patient

Execution Model
« Compliance »
→ Dose received by each patient

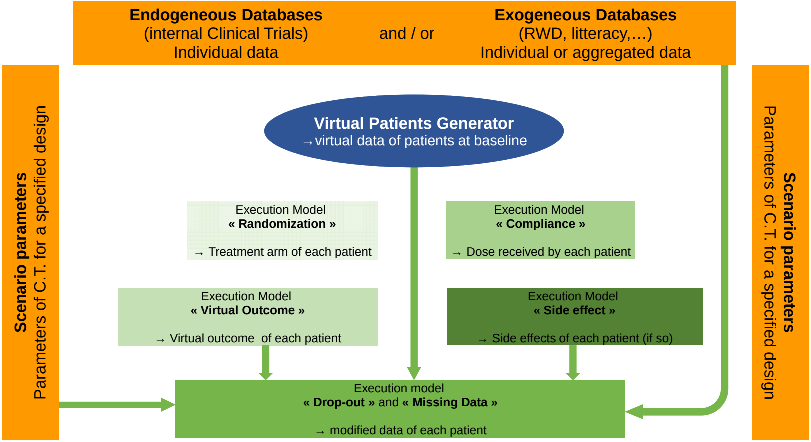
Execution Model
« Virtual Outcome »
→ Virtual outcome of each patient

Execution Model
« Side effect »
→ Side effects of each patient (if so)

Scenario parameters
Parameters of C.T. for a specified design

Example of what an In Silico Clinical Trial can be?

General Schema



Endogeneous Databases
(internal Clinical Trials)
Individual data

and / or

Exogeneous Databases
(RWD, litteracy,…)
Individual or aggregated data

Scenario parameters
Parameters of C.T. for a specified design

Virtual Patients Generator
→ virtual data of patients at baseline

Execution Model
« Randomization »
→ Treatment arm of each patient

Execution Model
« Compliance »
→ Dose received by each patient

Execution Model
« Virtual Outcome »
→ Virtual outcome of each patient

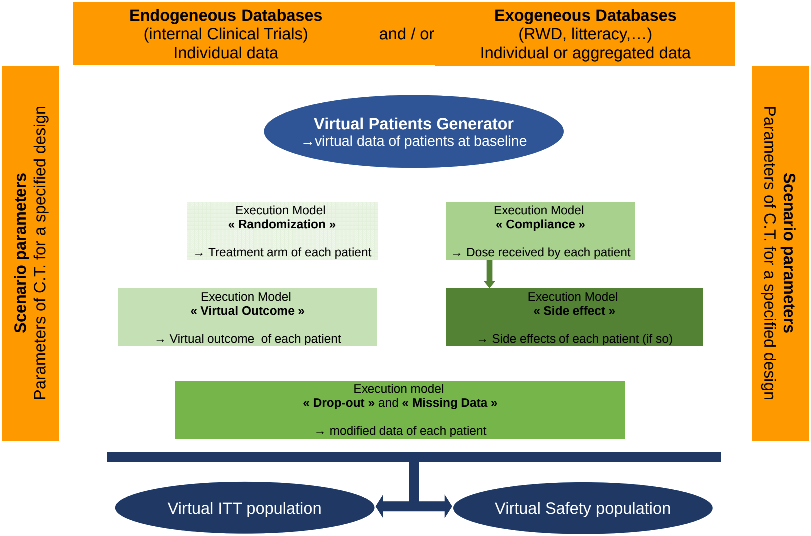
Execution Model
« Side effect »
→ Side effects of each patient (if so)

Execution model
« Drop-out » and **« Missing Data »**
→ modified data of each patient

Scenario parameters
Parameters of C.T. for a specified design

Example of what an In Silico Clinical Trial can be?

General Schema



Endogeneous Databases
(internal Clinical Trials)
Individual data

and / or

Exogeneous Databases
(RWD, litteracy,…)
Individual or aggregated data

Scenario parameters
Parameters of C.T. for a specified design

Virtual Patients Generator
→virtual data of patients at baseline

Execution Model
« Randomization »
→ Treatment arm of each patient

Execution Model
« Compliance »
→ Dose received by each patient

Execution Model
« Virtual Outcome »
→ Virtual outcome of each patient

Execution Model
« Side effect »
→ Side effects of each patient (if so)

Execution model
« Drop-out » and **« Missing Data »**
→ modified data of each patient

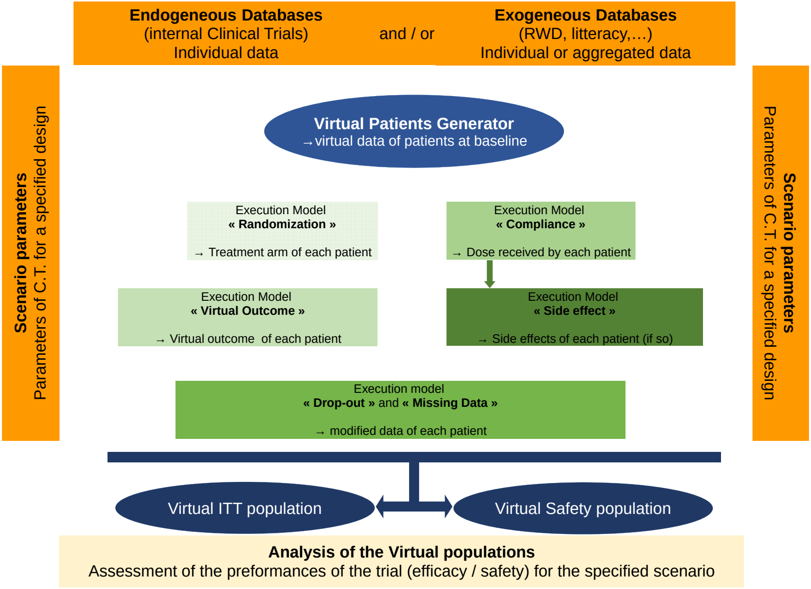
Virtual ITT population

Virtual Safety population

Scenario parameters
Parameters of C.T. for a specified design

Example of what an In Silico Clinical Trial can be?

General Schema



Endogeneous Databases
(internal Clinical Trials)
Individual data

and / or

Exogeneous Databases
(RWD, litteracy,…)
Individual or aggregated data

Scenario parameters
Parameters of C.T. for a specified design

Virtual Patients Generator
→virtual data of patients at baseline

Execution Model
« Randomization »
→ Treatment arm of each patient

Execution Model
« Compliance »
→ Dose received by each patient

Execution Model
« Virtual Outcome »
→ Virtual outcome of each patient

Execution Model
« Side effect »
→ Side effects of each patient (if so)

Execution model
« Drop-out » and **« Missing Data »**
→ modified data of each patient

Scenario parameters
Parameters of C.T. for a specified design

Virtual ITT population

Virtual Safety population

Analysis of the Virtual populations
Assessment of the preformances of the trial (efficacy / safety) for the specified scenario

What an In Silico Clinical Trial is?

Virtual patients Generator

- Generation of **Baseline** data for each patient

- Data **Modeled** as
  - a vector of covariates
  - by Monte Carlo procedure

- **Pros:** a virtual patient is a **good boy**
  - No ethical problem
  - No problem with ”General Data Protection Regulation”
  - Can follow various treatment arms at the same time
  - Perfectly adherent to want we want him to do

- **Cons:** Much more simple as a real patient
  - Many covariates to include for being realistic
  - Correlation between covariates

What an In Silico Clinical Trial is?

Virtual patients Generator

- Generation of **Baseline** data for each patient
- Data **Modeled** as
  - a vector of covariates
  - by Monte Carlo procedure
- **Pros:** a virtual patient is a **good boy**
  - No ethical problem
  - No problem with ”General Data Protection Regulation”
  - Can follow various treatment arms at the same time
  - Perfectly adherent to want we want him to do
- **Cons:** Much more simple as a real patient
  - Many covariates to include for being realistic
  - Correlation between covariates

- Generation of **Longitudinal** data for each patient **mimicking** the course of the trial by means of **Execution models**

- **Examples of execution models**
  - Outcome model
  - Disease progression model
  - Drug action model
  - Recruitment model
  - Side effect model
  - Parameters evolving model
  - ...

- **Keypoint**
  - Wide variety of models available
  - Models must be **Clinically Realistic** rather than highly statistically accurate
  - Each model may be improved separately (**Modularity**)

What an In Silico Clinical Trial is?

Execution models

- Generation of **Longitudinal** data for each patient **mimicking** the course of the trial by means of **Execution models**

- **Examples of execution models**
  - Outcome model
  - Disease progression model
  - Drug action model
  - Recruitment model
  - Side effect model
  - Parameters evolving model
  - ...

- **Keypoint**
  - Wide variety of models available
  - Models must be **Clinically Realistic** rather than highly statistically accurate
  - Each model may be improved separately (**Modularity**)

What an In Silico Clinical Trial may bring to drug development?

Perform sensitivity analyses of Clinical Trial endpoints

- **Models** (Generator and Execution) depend on various **parameters**:
  - Parameters linked to patients
  - Parameters linked to models (tuning parameters)
  - Parameters linked to the design
- Those parameters can be considered as
  - **Deterministic** or **random**
  - **Fixed by IHM** (scenario) or **Estimated** from data (calibration)
- Finally results come from **Monte Carlo simulation** accounting for
  - The model chosen
  - The values of whole the parameters
- Allow to Perform **sensitivity analyses** of Clinical Trial endpoints: **Impact of varying parameter(s)** on the performances of a clinical trials
  - Based on a parametric modelling of Execution Models
  - Based on restriction of domains of baseline covariates
  - Bayesian approach using distribution of variables
  - Specifying scenarii by fixing parameters

What an In Silico Clinical Trial may bring to drug development?

Perform performances analyses of a predefined trial

- Assessment of the performance of a predefined trial (Is the difference observed between treated and untreated patients is due to intervention (drug) ?) can be formally stated in terms of **potential outcomes** setting:
  For any patient $i = 1, 2, \ldots, N$ a potential outcome is $Y_i(T_i)$ where $T_i = 1$ if patient is treated and $T_i = 0$ otherwise.
- The effect of treatment is assessed by means of **The Average Treatment Effect (ATE)**:

$$ATE = \mathbb{E}\left[Y(1) - Y(0)\right]$$

- In practice, **potential outcomes** $Y_i(1)$ **et** $Y_i(0)$ cannot be observed simultaneously and *ATE* cannot be estimated properly.

- Thanks to ISCT, **performances analyses** of a predefined trial can be assessed since *ATE* can be estimated by

$$\hat{ATE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$$

What are the technical locks in practice ?

Related to Virtual patients Generator

- Problems related to **Monte Carlo simulation of multidimensional** random variable

- **Trade-off** between **details** of Virtual Patient and **complexity** of the model
- Have to account for **Correlation structure** between covariates
- Have to account for different types of covariates (Categorical / Quantitatives)
  - Different techniques exist: Discrete, Continuous, Copula (Savy, 2017)

- Complexity depends on **number of covariates**
  - Issue for choosing variables
  - Issue for calibration
    - **need huge datasets** (curse of dimensionality)
    - and / or need to specify assumptions

What an In Silico Clinical Trial is?

The models

- Generation of (**longitudinal**) data for each patient **mimicking** the course of the trial

- Data are **Modeled**
  - from a wide variety of models
  - those models must be **Clinically Realistic** rather than highly statistically accurate
  - each execution model may be improved separately (**Modularity**)
  - attention must be paid to the difference between

model for Prediction and **model for Simulation**

What an In Silico Clinical Trial is?

The models

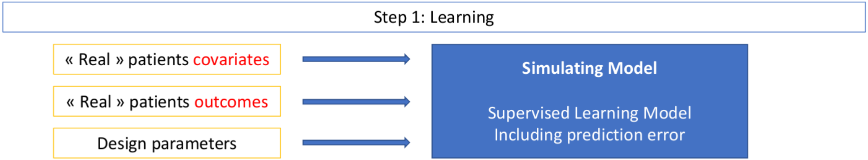
- Generation of (**longitudinal**) data for each patient **mimicking** the course of the trial

- Data are **Modeled**
  - from a wide variety of models
  - those models must be **Clinically Realistic** rather than highly statistically accurate
  - each execution model may be improved separately (**Modularity**)
  - attention must be paid to the difference between

model for Prediction and **model for Simulation**

Execution model: Simulation versus Prediction

Generalities



Step 1: Learning

« Real » patients covariates

« Real » patients outcomes

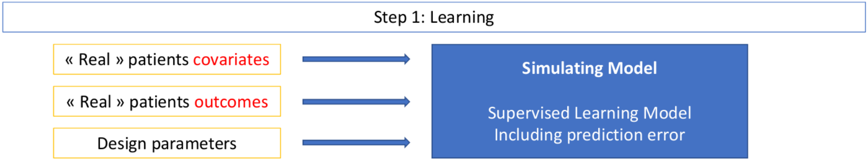
Design parameters

Simulating Model

Supervised Learning Model
Including prediction error

Execution model: Simulation versus Prediction

Generalities



Step 1: Learning

« Real » patients covariates

« Real » patients outcomes

Design parameters

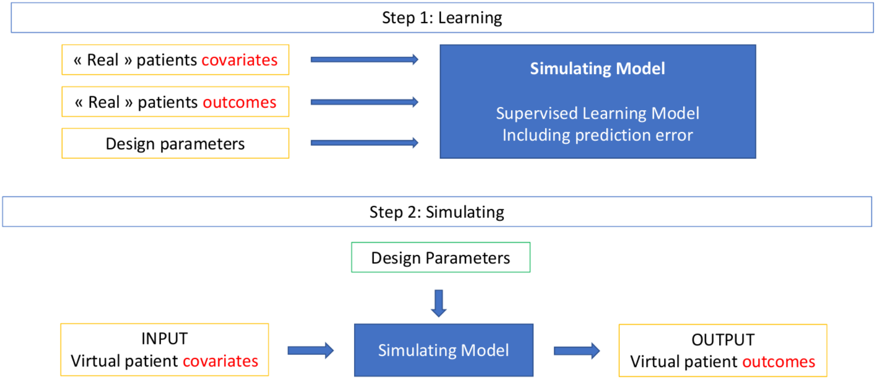
Simulating Model

Supervised Learning Model
Including prediction error

- Huge **diversity of models** may be considered
  - **Parametric models** (Markov, Cox, linear, logistics,...)
  - **Non-parametric models** (Machine learning)

Execution model: Simulation versus Prediction

Generalities



Step 1: Learning

« Real » patients covariates

« Real » patients outcomes

Design parameters

Simulating Model

Supervised Learning Model
Including prediction error

Step 2: Simulating

Design Parameters

INPUT
Virtual patient covariates

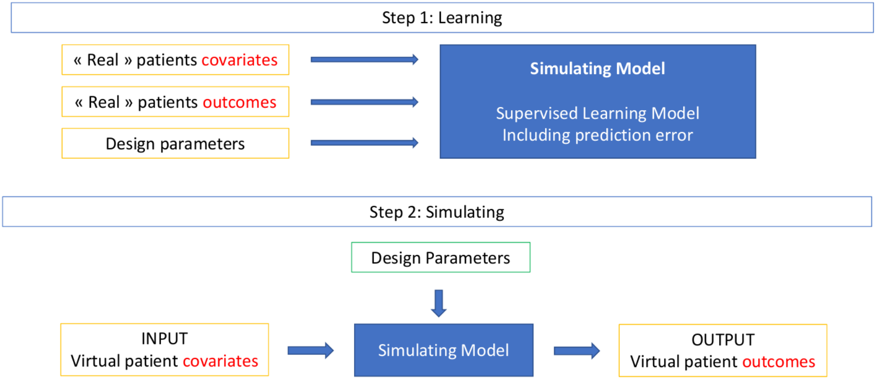
Simulating Model

OUTPUT
Virtual patient outcomes

- Huge **diversity of models** may be considered
  - **Parametric models** (Markov, Cox, linear, logistics,...)
  - **Non-parametric models** (Machine learning)

Execution model: Simulation versus Prediction

Generalities



Step 1: Learning

« Real » patients covariates

« Real » patients outcomes

Design parameters

Simulating Model

Supervised Learning Model
Including prediction error

Step 2: Simulating

Design Parameters

INPUT
Virtual patient covariates

Simulating Model

OUTPUT
Virtual patient outcomes

- Huge **diversity of models** may be considered
  - **Parametric models** (Markov, Cox, linear, logistics,...)
  - **Non-parametric models** (Machine learning)
- The aim of an **execution model** is to simulate not only to predict outcomes
  - Necessitate model with **good predictive performances**
  - + **Modeling of the error** of prediction

- It is not enough to use a **predictive model**
- Patients with the covariates implies patient with the same outcome
    - not realistic
    - Biological variability
    - Need to model the error of prediction
- **Continuous outcome**
    - model for prediction error distribution
    - Monte Carlo simulation according to this distribution
- **Categorical outcome**
    - confusion matrix for prediction error
    - Monte Carlo simulation according to Mutlinomial distribution

Execution model: Simulation versus Prediction

Continuous setting : Linear Regression

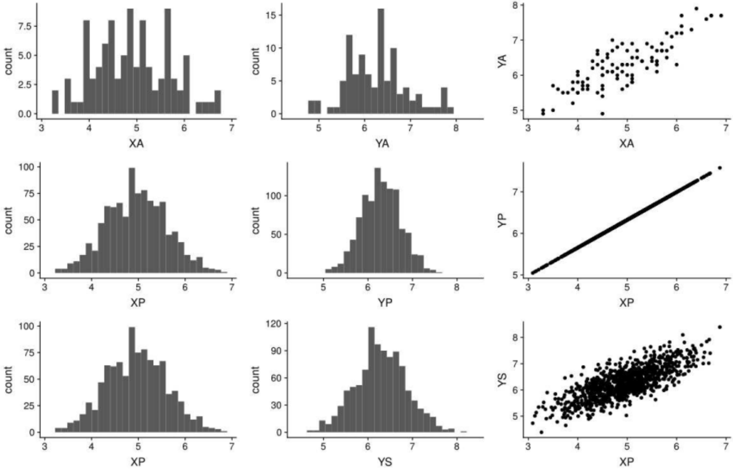


Figure 4: Illustration of the error made by considering only predictive performances. On the top the learning data, in the middle simulated abscissa and predicted ordinates, on the bottom simulated abscissa and simulated ordinates.

Execution model: Simulation versus Prediction

Categorical setting : Random Forest

- **Database:** Pima Indians Diabetes (267 patients)
- **Covariates:**
  - Number of times pregnant
  - Plasma glucose concentration a 2 hours in an oral glucose tolerance test
  - Diastolic blood pressure (mm Hg)
  - Triceps skinfold thickness (mm)
  - 2-Hour serum insulin (mu U/ml)
  - Body mass index
  - Age (year)
- **Outcome:** Diabete Yes or No
- **Key point of interest**
  - Group D+: Group of patients Diabete = Yes
  - Group D-: Group of patients Diabete = No
  - t.test to compare DBP for group D+ and D-: **P-value = 0.020**

Execution model: Simulation versus Prediction

Categorical setting : Random Forest

- **Predictive model:** Random Forest learned from a training set of 267 patients
- **Virtual patients:** random generation of 267 Virtual patients

  $\Longrightarrow$ continuous method

- **Virtual outcome:** using the random forest

  $\Longrightarrow$ **without** taking into account prediction error

  - Group VPD+: Group of virtual patients Diabete = Yes
  - Group VPD-: Group of virtual patients Diabete = No
  - t.test to compare DBP for group VPD+ and VPD-: **P-value = 0.000141**

- **Virtual outcome:** using the random forest

  $\Longrightarrow$ taking into account prediction error

  - Group VPED+: Group of virtual patients Diabete = Yes
  - Group VPED-: Group of virtual patients Diabete = No
  - t.test to compare DBP for group VPED+ and VPED-: **P-value = 0.041**

Take Home Message

- **ISCT** is a fantastic opportunity for drug development especially to **challenge trials' designs** **but**

- many **methodological challenges** remains especially
  - How to generate relevant virtual patients ?
  - How to build relevant execution models ?
  - How to calibrate or train those models ?
  - How to properly generate virtual outcomes ?

- many **technical challenges** remains especially
  - How to identify the right data ?
  - How to access the right databases ?
  - How to exploit properly those databases ?

Take Home Message

- **ISCT** is a fantastic opportunity for drug development especially to **challenge trials' designs** **but**

- many **methodological challenges** remains especially
  - How to generate relevant virtual patients ?
  - How to build relevant execution models ?
  - How to calibrate or train those models ?
  - How to properly generate virtual outcomes ?

- many **technical challenges** remains especially
  - How to identify the right data ?
  - How to access the right databases ?
  - How to exploit properly those databases ?

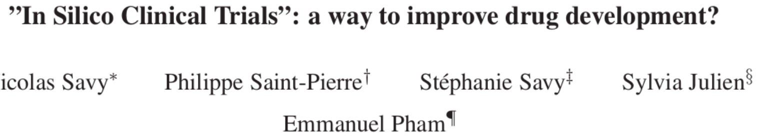
Take Home Message

- **ISCT** is a fantastic opportunity for drug development especially to **challenge trials' designs but**

- many **methodological challenges** remains especially
  - How to generate relevant virtual patients ?
  - How to build relevant execution models ?
  - How to calibrate or train those models ?
  - How to properly generate virtual outcomes ?

- many **technical challenges** remains especially
  - How to identify the right data ?
  - How to access the right databases ?
  - How to exploit properly those databases ?

References



”In Silico Clinical Trials”: a way to improve drug development?

Nicolas Savy[*] Philippe Saint-Pierre[†] Stéphanie Savy[‡] Sylvia Julien[§]
Emmanuel Pham[¶]

Proceedings of JSM conference, Denver, 2019

Content of the talk

1. Generalities on Simulated Clinical Trials

2. What an In Silico Clinical Trial is?

3. An example in Economics Health

4. Databases Merging issue

Context

- From 2012 to 2016 only **four generics** of ARV were available
    - These generics are **less and less** used
- During the year 2017, **three new generics** of ARVs were launched on the market
    - corresponding to **combo generics**
    - more used than the first wave of generics
- In the coming years, new generics of ARVs as well as patent expiry for other molecules are expected
- **What are the expected money savings generated by this switch to generics?**

Usual strategy - Results

- Very **few studies** have assessed the economical impact of generic arrivals

- An Italian study of 2015

  187.4 million euros of savings between 2015 and 2019
- A british study of 2014
  - 1.46 billion euros of savings between 2015 and 2019

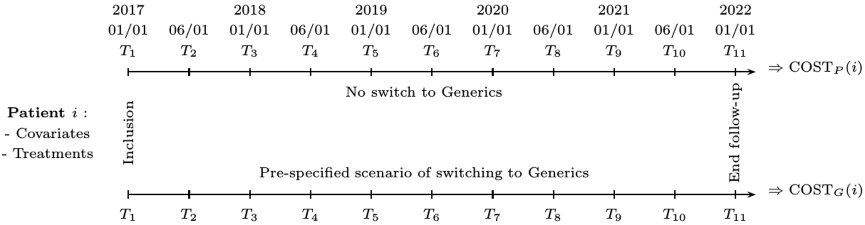
- These studies use
  - **simplistic approaches** based on **strong assumptions** and **poorly documented** data sources
  - **population approach** considering an average behavior of patient
  - **population approach** does not take into account **inter-patients variability** of various parameters influencing the model

- **Results** express in terms of **punctual estimation** of the variations of costs embedded by the switch to the generics

An agent-based strategy

Here is a different approch. Consider

- a cohort of patients
- followed during 5 years
- by steps of six months
  - coincides with a standard medical monitoring of such patients
- Each patient follows ”virtually” two trajectories:
  - without switching to generic
  - with a switch to generic with specific scenario



The material: Dat'Aids

The **Dat'Aids database** has been used

- **to construct**
  - The baseline characteristics of patients (time T_1) `PATIENT`
  - A database of medications, denoted as `MEDICATION`
  - A database of treatments, denoted as `TREATMENT`
- **to calibrate** the execution models considered
  - for updating, at each T_u, the patients characteristics,
  - for updating, at each T_u, the treatment
  - for updating, at each T_u, the cohort (deaths and incident cases).

- Dat'Aids contains the following information for 27341 patients:
  - Civil status of the patient,
  - Social record,
  - HIV / Hepatitis clinical record,
  - Pathological and therapeutic history,
  - Clinical examinations,
  - Biological results,
  - Antiretroviral genotypes and dosings,
  - Medicinal prescriptions,
  - Examinations and checkup prescriptions,
  - Consultation and diagnosis motivations,
  - PMSI (Programme de Médicalisation des Systèmes dInformation).

The cohort of patients at baseline

- **MEDICATION database** is constructed from the DatAids database
  - involves 31 of the main medications used for HIV management
  - For each medication is collected:
    - NAMEM: The name of the medication,
    - REFO: The amount refounded by ”Assurance Maladie”,
    - AMMT: The marketing authorization date.
  - enriched by **parameters of the simulation scenarios**:
    - AMMGM: The marketing authorization date of the generic version of the medication,
    - PENRATE: The maximal penetration rate of the generic
    - PENTIME: The penetration time of the generic
    - PROBCONV: The probability of conversion to generic assumed to increase linearly between AMMGM and PENTIME.

- **TREATMENT database** is constructed from the MEDICATION database
  - a treatment is a combination of medications due to multi-therapy
  - a total of more than 800 different treatments in the database
  - for each treatment, is collected:
    - NAMET: The name of the treatment which is a combination of medications,
    - MEDCOSTP: The cost for the princeps version of the treatment,
    - MEDCOSTG: The cost for the generic version of the treatment.

The cohort of patients

- $(\texttt{PATIENT}(u,i), \texttt{TREATMENT}(u,i))$ defines covariates of patient i at times T_u
- Covariates $\texttt{TREATMENT}(u,i)$ are the treatments of patient i at time T_u
- Covariates $\texttt{PATIENT}(u,i)$ are characteristics of patient i at time T_u and are
  - the ones available in **Dat'Aids database**
  - the ones that may have an impact on the choice of the medication
  - the ones that may have an impact on the evolution of other covariates
- Baseline values $(u = 1)$ are directly picked in **Dat'Aids database**.

The cohort of patients

- Covariates PATIENT$(u, .)$ can be classified in three categories:
  - Demographic covariates,
    - SEX, the sex of the patient (0 for male and 1 for female)
    - AGE, the age of the patient (in months)
    - BC, the country of birth, (1 for France and 0 for elsewhere)
  - Covariates linked to the pathology and its history,
    - CONTA, the way of contamination of the patient (1 for homosexual and 0 for not)
    - VIHS, the status of the infection (1 for SIDA and 0 for not)
    - VIHD, the duration of the HIV infection(in months)
    - TREATD, the duration of the last treatment (in months)
    - ARN, the viral load.
  - Covariates linked to the comorbidities,
    - HEART, cardiovascular illnesses (1 for Yes and 0 for No),
    - DIAB diabetes (1 for Yes and 0 for No)
    - IR, Renal failure (1 for Yes and 0 for No)
    - DEATH indicates whether a patient is alive (1 for alive and 0 for dead)

The algorithm

The algorithm is split into four steps:

- **Step 1:** Updating of the patient covariates: transition from $\texttt{PATIENT}(u-1,i)$ to $\texttt{PATIENT}(u,i)$
- **Step 2:** Updating of the patient treatment: transition from $\texttt{TREATMENT}(u-1,i)$ to $\texttt{TREATMENT}(u,i)$ according to $\texttt{PATIENT}(u,i)$
- **Step 3:** Updating of the cohort of patients by considering possibility of death and by including incident cases
- **Step 4:** Assessment of the costs for each scenario during the period $[T_u, T_{u+1}[$ denoted $\mathrm{COST}_P(u,i)$ and $\mathrm{COST}_G(u,i)$.

The execution models 1/3: Updating of the patient covariates

- **Covariates fixed in time.**

  For $u = 2, \dots, 10$, we have:

$$\begin{aligned} \texttt{SEX}(T_u) &= \texttt{SEX}(T_1), \\ \texttt{CONTA}(T_u) &= \texttt{CONTA}(T_1), \\ \texttt{BC}(T_u) &= \texttt{BC}(T_1). \end{aligned}$$

- **Covariates with deterministic dependence on time.**

  For $u = 2, \dots, 10$, we have:

$$\begin{aligned} \texttt{AGE}(T_u) &= \texttt{AGE}(T_{u-1}) + 6, \\ \texttt{VIHD}(T_u) &= \texttt{VIHD}(T_{u-1}) + 6. \end{aligned}$$

$$\texttt{TREATD}(T_u) = \begin{cases} \texttt{TREATD}(T_{u-1}) + 6, & \text{if there is no switch of treatment,} \\ 0, & \text{if there is a change of treatment at time } T_u. \end{cases}$$

The execution models 1/3: Updating of the patient covariates

- **Covariates with random dependence in time.**
  HEART, DIAB, VIHS and DEATH may change during the patient follow-up
  - Changes modeled by **Markov chains**
  - The matrices of transition are chosen to be **constant**
  - **Coefficients estimated** from Dat'Aids database

- **Covariates with random dependence in time and randomness depending of covariates.**
  ARN and CREA may change during the patient follow-up
  - Changes modeled by **Markov chains with transitions depending of covariates**
  - The transitions are **modeled by logistic or polytomic regressions**
  - **Coefficients of the regressions are estimated** from Dat'Aids database
  - the covariates involved in the model are selected by a backward stepwise strategy

  - for ARN(T_u), the covariates involved are ARN(T_{u-1}), IR, CONTA, HEART, VIHS, AGE, SEX, VIHD and TREATD.
  - for CREA(T_u), the covariates involved are CREA(T_{u-1}), SEX, ARN(T_u), AGE, HEART, TREATD, VIHS and VIHD.

- These changes can lead to a **modification in patient treatment**.

The execution models 2/3: updating of the treatment

- **Four rules:**
  - a patient can keep his treatment
  - a patient can switch for a treatment to another
  - a patient can switch to the generic version of his treatment
  - a patient cannot convert back to the princeps if he converts its medication to generic

- Patients can **change his treatment** according to a Markov chain
  - the transition matrix is **estimated** from the `Dat'Aids` database
    - if a transition is rare ($<$ 100 observations), the probability is considered as constant
    - else a logistic regression model is adjusted, the choice of covariates involved is done case-by-case following a backward step-by-step strategy

- Patients can **switch to generic** with probability depending of time *t* defined by:

$$\mathrm{PROBCONV}(t) = \begin{cases} 0 & \text{if } t < \mathrm{AMMGM}, \\ \mathrm{PENRATE}\,\frac{t - \mathrm{AMMGM}}{\mathrm{PENTIME} - \mathrm{AMMGM}} & \text{if } \mathrm{AMMGM} \leq t < \mathrm{PENTIME}, \\ \mathrm{PENRATE} & \text{if } t > \mathrm{PENTIME}. \end{cases}$$

The execution models 3/3: updating of the cohort

- **Including** 455 incident cases.
  - in adequacy with literacy results
  - for each incident case,
    - baseline values are randomly chosen from the `PATIENT`(1, .) vector
    - the treatment is randomly chosen in the updated `TREATMENT` database

- **Patients who died** during the period of interest
  - are not removed of the cohort
  - their future costs are fixed to 0.

Differential costs calculation

For each patient *i* can be computed:

- The **differential cost** DC:

$$\mathrm{DC}(i) = \mathrm{COST}_P(i) - \mathrm{COST}_G(i) = \sum_{u=1}^{10} \left(\mathrm{COST}_P(u,i) - \mathrm{COST}_G(u,i)\right),$$

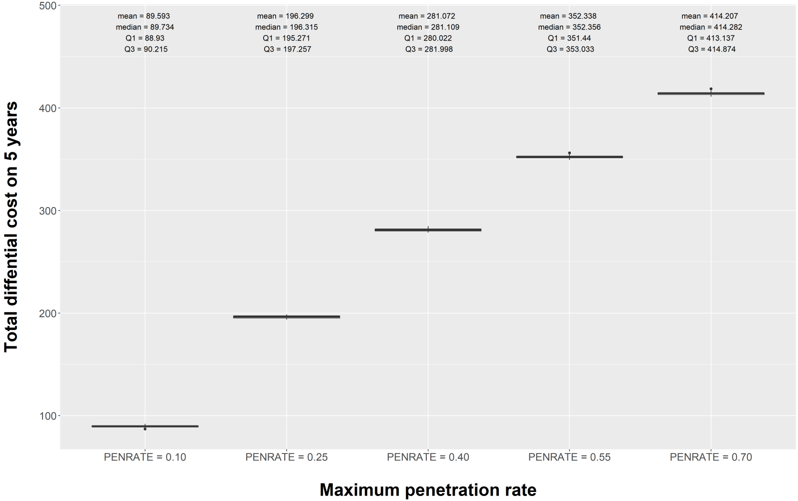
- The **normalized differential cost** NDC:

$$\mathrm{NDC}(i) = \frac{\mathrm{DC}(i)}{\mathrm{FD}(i)}.$$

  - $\mathrm{COST}_P(i)$: the cost considering no switch to generics
  - $\mathrm{COST}_G(i)$: the cost assuming a pre-specified scenario of switching to generics
  - $\mathrm{FD}(i)$: the duration of the follow-up

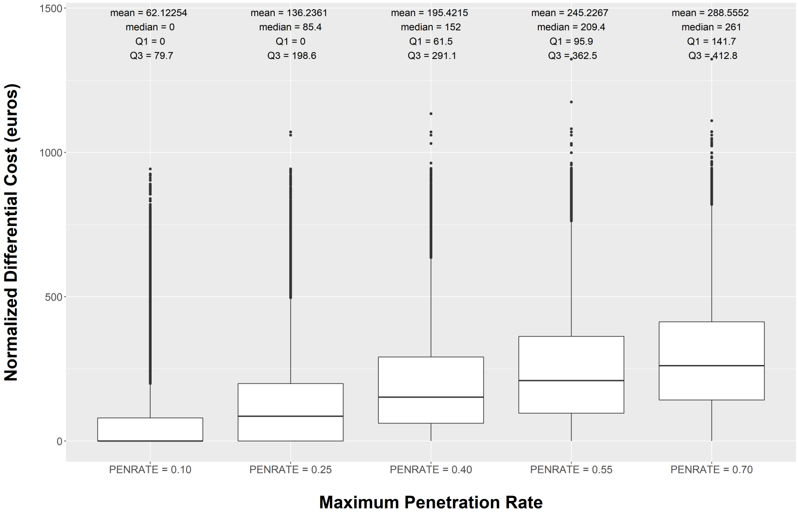
- **100 simulation runs** are performed
  - yield to the **empirical distribution of each parameter**
  - derive 90% **prediction intervals** considering the 5th and 95th values of the sorted distribution.

The results 1/3



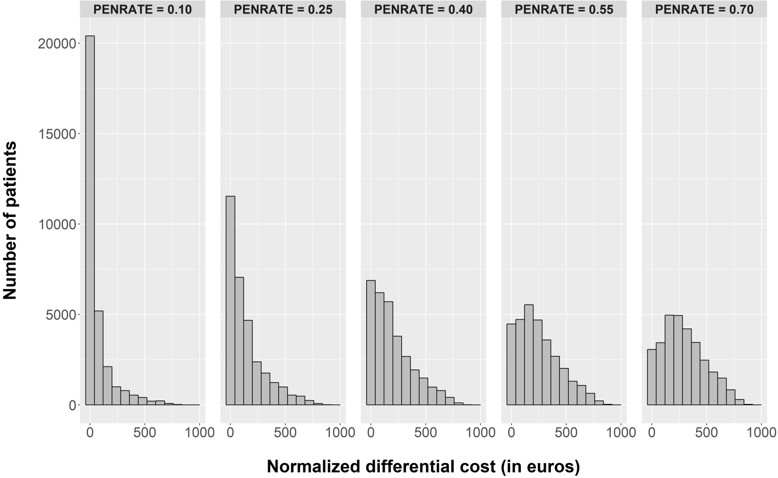
Boxplot of the total differential cost on five year for French population as a function of the penetration rate (from 100 simulation runs, in millions on euros)

The results 2/3



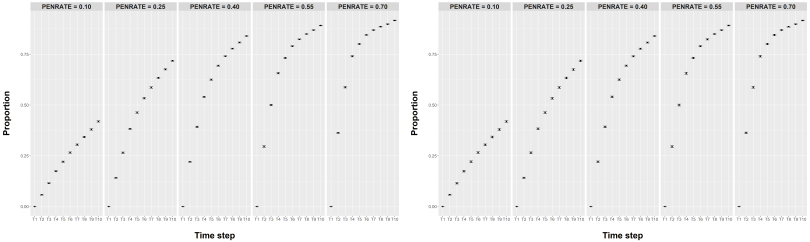
Boxplot of the normalized differential costs per patient per year as a function of scenarios defined by penetration rate (from 100 simulation runs, in euros)

The results 3/3



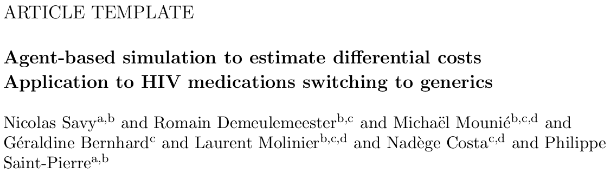
Distribution of the normalized differential costs per patient as a function of scenarios defined by penetration rates (from 100 simulation runs, in euros)

The results 3/3



Boxplots of the proportion of patients who were prescribed a generic at least once during the follow up together with their prediction intervals (from 100 simulation runs). 90% prediction intervals - 80% prediction intervals

References



ARTICLE TEMPLATE

Agent-based simulation to estimate differential costs
Application to HIV medications switching to generics

Nicolas Savy[a,b] and Romain Demeulemeester[b,c] and Michaël Mounié[b,c,d] and Géraldine Bernhard[c] and Laurent Molinier[b,c,d] and Nadège Costa[c,d] and Philippe Saint-Pierre[a,b]

In revision for Journal of Statistical Computation and Simulation, 2019

Content of the talk

1. Generalities on Simulated Clinical Trials

2. What an In Silico Clinical Trial is?

3. An example in Economics Health

4. Databases Merging issue

Variables recoding issue

- Health components are diverse

$\Longrightarrow$ ”Big Data” is built from **various sources of information**:

  - clinical trials
  - medico-administrative databases
  - patients cohort
  - medical file
  - patients data (from connected devices for example)
  - social networks
  - ...

$\Longrightarrow$ To exploit of these multitudes of databases ask questions

  - **Data chaining** to construct care path
  - **Data merging** to enlarge database
  - **Variable recoding** to unify the information contained in different databases
  - ...

Variables recoding issue

- Health components are diverse

⟹ ”Big Data” is built from **various sources of information**:

  - clinical trials
  - medico-administrative databases
  - patients cohort
  - medical file
  - patients data (from connected devices for example)
  - social networks
  - ...

⟹ To exploit of these multitudes of databases ask questions

  - **Data chaining** to construct care path
  - **Data merging** to enlarge database
  - **Variable recoding** to unify the information contained in different databases
  - ...

Variables recoding issue

- Health components are diverse

$\Longrightarrow$ ”Big Data” is built from **various sources of information**:

  - clinical trials
  - medico-administrative databases
  - patients cohort
  - medical file
  - patients data (from connected devices for example)
  - social networks
  - ...

$\Longrightarrow$ To exploit of these multitudes of databases ask questions

  - **Data chaining** to construct care path
  - **Data merging** to enlarge database
  - **Variable recoding** to unify the information contained in different databases
  - ...

Introducing example

- **NCDS (The National Child Development Study)**
  - a continuing survey which follows the lives of over 17,000 people born in England, Scotland and Wales in a same week of the year 1958
  - collects specific information on many distinct fields
    - *physical and educational development, economic circumstances, employment, family life, health behaviour, well-being, social participation and attitudes*
  - 9 waves (0, 7, 11, 16, 22, 33, 42, 50 and 55 years old)

- **Outcome**: social status of the participants :
  - Two scales built from profession :
  - *Goldthorp social class'90 scale* (`GSS90`) : a scale in 11 categories
  - *RGs social Class'91 scale* (`RGS91`) : a scale in 6 categories.

- Social status assessed by these scales at some waves
- Necessitate **to recode the variable Social Status** to perform analysis

Variables recoding issue

Database A

	C_1	C_2	$\dots$	C_p	Y^A	Y^B
1					Observed	Unobserved
$\dots$						
$\dots$						
n_A						

Database B

	C_1	C_2	$\dots$	C_p	Y^A	Y^B
1					Unobserved	Observed
$\dots$						
$\dots$						
n_B						

- Y evaluated in both databases but not assessed on the same variable

Aim : Complete Y^A on database B and/or complete Y^B on database A

- Ideas
  - Missing data problem (MAR)
  - Latent variables models (class latent analysis, trait latent analysis)
  - Estimation (polytomous regression) / Prediction

Variables recoding issue

Database A

	C_1	C_2	$\dots$	C_p	Y^A	Y^B
1					Observed	Unobserved
$\dots$						
$\dots$						
n_A						

Database B

	C_1	C_2	$\dots$	C_p	Y^A	Y^B
1					Unobserved	Observed
$\dots$						
$\dots$						
n_B						

- Y evaluated in both databases but not assessed on the same variable

Aim : Complete Y^A on database B and/or complete Y^B on database A

- Ideas
  - Missing data problem (MAR)
  - Latent variables models (class latent analysis, trait latent analysis)
  - Estimation (polytomous regression) / Prediction

Variables recoding issue

Database A

	C_1	C_2	$\dots$	C_p	Y^A	Y^B
1					Observed	Unobserved
$\dots$						
$\dots$						
n_A						

Database B

	C_1	C_2	$\dots$	C_p	Y^A	Y^B
1					Unobserved	Observed
$\dots$						
$\dots$						
n_B						

- Y evaluated in both databases but not assessed on the same variable

Aim : Complete Y^A on database B and/or complete Y^B on database A

- Ideas
  - Missing data problem (MAR)
  - Latent variables models (class latent analysis, trait latent analysis)
  - Estimation (polytomous regression) / Prediction

Variables recoding issue

Database A

	C_1	C_2	$\dots$	C_p	Y^A	Y^B
1					Observed	Unobserved
$\dots$						
$\dots$						
n_A						

Database B

	C_1	C_2	$\dots$	C_p	Y^A	Y^B
1					Unobserved	Observed
$\dots$						
$\dots$						
n_B						

- Y evaluated in both databases but not assessed on the same variable

Aim : Complete Y^A on database B and/or complete Y^B on database A

- Ideas
  - Missing data problem (MAR)
  - Latent variables models (class latent analysis, trait latent analysis)
  - Estimation (polytomous regression) / Prediction
  - **Optimal transportation**

Optimal Transportation

Y^A Y^B

μ ν

Optimal Transportation

Y^A Y^B

$\xrightarrow{T}$

μ ν

- T such that $\nu = T\mu$ is a **transportation map** from μ to ν

Optimal Transportation

Y^A $\qquad$ Y^B

$$\xrightarrow{T}$$

μ $\qquad$ ν

- *T* such that $\nu = T\mu$ is a **transportation map** from μ to ν

Optimal Transportation

Y^A $\qquad\qquad\qquad$ Y^B

$\xrightarrow{T}$

μ $\qquad\qquad\qquad$ ν

- T such that $\nu = T\mu$ is a **transportation map** from μ to ν
- **Optimal transportation**
  - Let c a cost function measuring the displacement from y^A to y^B
  - Find a map T such that the average displacement is minimal

Optimal Transportation

- $\mathbb{Y}^A$ and $\mathbb{Y}^B$: two Radon spaces
- $c : \mathbb{Y}^A \times \mathbb{Y}^B \longrightarrow [0, \infty]$ a Borel-measurable function given probability measures μ on $\mathbb{Y}^A$ and ν on $\mathbb{Y}^B$ (**cost function**)

- **Monge's formulation** (1781): Find a transport map $T : \mathbb{Y}^A \to \mathbb{Y}^B$ that realizes the infimum:

$$\left\{ \int_{\mathbb{Y}^A} c\left(y^A, T\left(y^A\right)\right) \mathrm{d}\mu\left(y^A\right) \;\middle|\; T(\mu) = \nu \right\},$$

  - **Optimal transportation map** : map T realizing this infimum
  - *Non-linear optimization problem, rigid assumptions on the regularity of T*

- **Kantovorich's formulation** (1942): Find a measure $\gamma \in \gamma(\mu, \nu)$ that realizes the infimum:

$$\left\{ \int_{\mathbb{Y}^A \times \mathbb{Y}^B} c\left(y^A, y^B\right) \mathrm{d}\gamma\left(y^A, y^B\right) \;\middle|\; \gamma \in \gamma(\mu, \nu) \right\},$$

where $\gamma(\mu, \nu)$ denote the set of measures on $\mathbb{Y}^A \times \mathbb{Y}^B$ with marginals μ on $\mathbb{Y}^A$ and ν on $\mathbb{Y}^B$

  - *Linear problem, solution achievable with compacity (volume fitting) argument*

Optimal Transportation

- $\mathbb{Y}^A$ and $\mathbb{Y}^B$: two Radon spaces
- $c : \mathbb{Y}^A \times \mathbb{Y}^B \longrightarrow [0, \infty]$ a Borel-measurable function given probability measures μ on $\mathbb{Y}^A$ and ν on $\mathbb{Y}^B$ (**cost function**)

- **Monge's formulation** (1781): Find a transport map $T : \mathbb{Y}^A \to \mathbb{Y}^B$ that realizes the infimum:

$$\left\{ \int_{\mathbb{Y}^A} c\left(y^A, T\left(y^A\right)\right) \mathrm{d}\mu\left(y^A\right) \,\middle|\, T(\mu) = \nu \right\},$$

  - **Optimal transportation map** : map T realizing this infimum
  - *Non-linear optimization problem, rigid assumptions on the regularity of T*

- **Kantovorich's formulation** (1942): Find a measure $\gamma \in \gamma(\mu, \nu)$ that realizes the infimum:

$$\left\{ \int_{\mathbb{Y}^A \times \mathbb{Y}^B} c\left(y^A, y^B\right) \mathrm{d}\gamma\left(y^A, y^B\right) \middle| \gamma \in \gamma(\mu, \nu) \right\},$$

where $\gamma(\mu, \nu)$ denote the set of measures on $\mathbb{Y}^A \times \mathbb{Y}^B$ with marginals μ on $\mathbb{Y}^A$ and ν on $\mathbb{Y}^B$

  - *Linear problem, solution achievable with compacity (volume fitting) argument*

- **Continuous case**
  - The optimal transportation map **exists** and is **unique** if h is strictly convex with $c(x,y) = h(x-y)$
- **Discrete case: Hitchcock's problem (1941)**
  - Y^A the assessment of Y on database $D = A$
    - with distribution μ discrete with modalities $\{m_1^A, \ldots, m_R^A\}$
    - $a_r = \mathbb{P}(Y^A = m_r^A)$, $r = 1, \ldots, R$

$$\mu = \sum_{r=1}^{R} a_r \delta_{m_r^A}$$

  - Y^B the assessment of Y on database $D = B$
    - with distribution ν discrete with modalities $\{m_1^B, \ldots, m_S^B\}$
    - $b_s = \mathbb{P}(Y^B = m_s^B)$, $s = 1, \ldots, S$

$$\nu = \sum_{s=1}^{S} b_s \delta_{m_s^B}$$

  - $\mathbf{X} = (C_1, C_2, \ldots, C_p)$, covariates

OT Algorithm: Optimal Transportation for Variable recoding

The optimal joint distribution γ^{opt} of (Y^A, Y^B) is solution to the linear programming:

$$\gamma^{opt} \text{ minimizes } \gamma = \{\gamma_{r,s}, r = 1, \ldots, R, s = 1, \ldots, S\} \to \sum_{r=1}^{R} \sum_{s=1}^{S} \gamma_{r,s}\, c\,(p_r, q_s)\,,$$

under the following constraints

$$\begin{cases} \sum_{r=1}^{R} \gamma_{r,s} = \mu_s, \quad \forall s = 1, \ldots S \\ \sum_{s=1}^{S} \gamma_{r,s} = \nu_r, \quad \forall r = 1, \ldots R \\ \gamma_{r,s} \geq 0, \quad \forall r = 1, \ldots R, \forall s = 1, \ldots S. \end{cases}$$

with

$$\begin{aligned} c(p_r, q_s) &= \mathbb{E}\left[d(\bar{\mathbf{X}}, \bar{\bar{\mathbf{X}}}) | Y^A = p_r, Y^B = q_s\right] \quad \text{if } \mathbb{P}[Y^A = p_r, Y^B = q_s] \neq 0 \\ &= 0 \text{ otherwise} \end{aligned}$$

where $\bar{\mathbf{X}}$ and $\bar{\bar{\mathbf{X}}}$ are two independent copies of $\mathbf{X}$.

Variable recoding - Algorithm OT

It is possible to derive an **algorithm for variable recoding** in two steps:

1. Estimation of γ^{opt} the joint distribution of (Y^A, Y^B)

2. Allocation of a new code to each patient

OT Algorithm: Optimal Transportation for Variable recoding

γ^{opt} **is estimated by** $\hat{\gamma}^{opt}$ solution to the linear programming:

$$\hat{\gamma}^{opt} \text{ minimizes } \{\gamma_{r,s}, r = 1, \dots, R, s = 1, \dots, S\} \to \sum_{r=1}^{R} \sum_{s=1}^{S} \gamma_{r,s}\, \hat{c}_{n_A,n_B}(p_r, q_s)$$

under the following constraints

$$\begin{cases} \displaystyle\sum_{r=1}^{R} \gamma_{r,s} = (\hat{b}_{n_B})_s, \quad \forall s = 1, \dots S \\ \displaystyle\sum_{s=1}^{S} \gamma_{r,s} = (\hat{a}_{n_A})_r, \quad \forall r = 1, \dots R \\ \gamma_{r,s} \geq 0, \quad \forall r = 1, \dots R, \forall s = 1, \dots S \end{cases}$$

OT Algorithm: Optimal Transportation for Variable recoding

- The marginal distributions of Y^A and Y^B are estimated by

$$\begin{aligned}(\hat{a}_{n_A})_r &= \frac{\text{card}\left\{\{i|\,Y_i^A = m_r^A\}\right\}}{n_A}, \qquad r = 1,\dots R,\\ (\hat{b}_{n_B})_s &= \frac{\text{card}\left\{\{j|\,Y_j^B = m_s^B\}\right\}}{n_B}, \qquad s = 1,\dots S.\end{aligned}$$

⇒ **Assumption 1 :** $\mathcal{L}(Y^A|D=A) = \mathcal{L}(Y^A|D=B)$

- The cost function is estimated by

$$\begin{aligned}\hat{c}_{n_A,n_B}(p_r,q_s) &= \frac{1}{\kappa_{r,s}} \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d(\mathbf{X}_i^A,\mathbf{X}_j^B)\,\mathbb{I}_{\left\{Y_i^A=p_r\ ,\ Y_j^B=q_s\right\}} \qquad \text{if } \kappa_{r,s} \neq 0\\ &= 0 \text{ otherwise}\end{aligned}$$

with $\kappa_{r,s} =$ card $\left\{\{(i,j)|y_i^A = m_r^A\ ,\ y_j^B = m_s^B\}\right\}$.

⇒ **Assumption 2** : $\mathcal{L}(Y^A|C,D=A) = \mathcal{L}(Y^A|C,D=B)$.

OT Algorithm: Optimal Transportation for Variable recoding

- The marginal distributions of Y^A and Y^B are estimated by

$$\begin{aligned}(\hat{a}_{n_A})_r &= \frac{\text{card}\left\{\{i|\,Y_i^A = m_r^A\}\right\}}{n_A}, \qquad r = 1,\dots R,\\ (\hat{b}_{n_B})_s &= \frac{\text{card}\left\{\{j|\,Y_j^B = m_s^B\}\right\}}{n_B}, \qquad s = 1,\dots S.\end{aligned}$$

$\Rightarrow$ **Assumption 1 :** $\mathcal{L}(Y^A|D=A) = \mathcal{L}(Y^A|D=B)$

- The cost function is estimated by

$$\begin{aligned}\hat{c}_{n_A,n_B}(p_r,q_s) &= \frac{1}{\kappa_{r,s}} \sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d(\mathbf{X}_i^A,\mathbf{X}_j^B)\,\mathbb{I}_{\left\{Y_i^A=p_r\ ,\ Y_j^B=q_s\right\}} \qquad \text{if } \kappa_{r,s}\neq 0\\ &= 0 \text{ otherwise}\end{aligned}$$

with $\kappa_{r,s} = \text{card}\left\{\{(i,j)|y_i^A = m_r^A\ ,\ y_j^B = m_s^B\}\right\}$.

$\Rightarrow$ **Assumption 2** : $\mathcal{L}(Y^A|C,D=A) = \mathcal{L}(Y^A|C,D=B)$.

OT Algorithm: Optimal Transportation for Variable recoding

- The marginal distributions of Y^A and Y^B are estimated by

$$(\hat{a}_{n_A})_r = \frac{\text{card}\left\{\{i | Y_i^A = m_r^A\}\right\}}{n_A}, \qquad r = 1, \dots R,$$

$$(\hat{b}_{n_B})_s = \frac{\text{card}\left\{\{j | Y_j^B = m_s^B\}\right\}}{n_B}, \qquad s = 1, \dots S.$$

$\Rightarrow$ **Assumption 1 :** $\mathcal{L}(Y^A | D = A) = \mathcal{L}(Y^A | D = B)$

- The cost function is estimated by

$$\begin{aligned}\hat{c}_{n_A, n_B}(p_r, q_s) &= \frac{1}{\kappa_{r,s}} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(\mathbf{X}_i^A, \mathbf{X}_j^B)\, \mathbb{I}_{\left\{Y_i^A = p_r \ , \ Y_j^B = q_s\right\}} \quad \text{if } \kappa_{r,s} \neq 0 \\ &= 0 \text{ otherwise}\end{aligned}$$

with $\kappa_{r,s} = \text{ card}\left\{\{(i,j) | y_i^A = m_r^A \ , \ y_j^B = m_s^B\}\right\}$.

$\Rightarrow$ **Assumption 2** : $\mathcal{L}(Y^A | C, D = A) = \mathcal{L}(Y^A | C, D = B)$.

OT Algorithm: Optimal Transportation for Variable recoding

- The marginal distributions of Y^A and Y^B are estimated by

$$(\hat{a}_{n_A})_r = \frac{\text{card}\left\{\{i| Y_i^A = m_r^A\}\right\}}{n_A}, \qquad r = 1, \dots R,$$
$$(\hat{b}_{n_B})_s = \frac{\text{card}\left\{\{j| Y_j^B = m_s^B\}\right\}}{n_B}, \qquad s = 1, \dots S.$$

$\Rightarrow$ **Assumption 1 :** $\mathcal{L}(Y^A|D = A) = \mathcal{L}(Y^A|D = B)$

- The cost function is estimated by

$$\begin{aligned} \hat{c}_{n_A,n_B}(p_r, q_s) &= \frac{1}{\varkappa_{r,s}} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(\mathbf{X}_i^A, \mathbf{X}_j^B) \, \mathbb{I}_{\left\{Y_i^A = p_r \ , \ Y_j^B = q_s\right\}} \quad \text{if } \varkappa_{r,s} \neq 0 \\ &= 0 \text{ otherwise} \end{aligned}$$

with $\varkappa_{r,s} = \text{ card}\left\{\{(i,j)|y_i^A = m_r^A \ , \ y_j^B = m_s^B\}\right\}$.

$\Rightarrow$ **Assumption 2** : $\mathcal{L}(Y^A|C, D = A) = \mathcal{L}(Y^A|C, D = B)$.

Variable recoding - Algorithm OT

Example :

Consider Y^A is observed and Y^B unobserved.

		Variable Y^A			
		m_1^A	m_2^A	m_3^A	m_4^A
Variable Y^B	m_1^B				
	m_2^B				
	m_3^B				

Variable recoding - Algorithm OT

Example :

Consider Y^A is observed and Y^B unobserved.

		Variable Y^A			
		m_1^A	m_2^A	m_3^A	m_4^A
Variable Y^B	m_1^B	84	23	2	5
	m_2^B	23	76	14	4
	m_3^B	3	2	55	13

Variable recoding - Algorithm OT

Example :

Consider Y^A is observed and Y^B unobserved.

		Variable Y^A			
		m_1^A	m_2^A	m_3^A	m_4^A
Variable Y^B	m_1^B	84	23	2	5
	m_2^B	23	76	14	4
	m_3^B	3	2	55	13

Which are the 84 individuals encoded m_1^A that will be recoded m_1^B ?

Variable recoding - Algorithm OT

Step 2 of OT algorithm: affectation

For each subject *i* of database A, a predicted value for $\hat{y}_i^B$ can be constructed by means of an **adapted nearest neighbor algorithm** accounting for the covariates with distance *d*.

Example :

		Variable Y^A			
		m_1^A	m_2^A	m_3^A	m_4^A
Variable Y^B	m_1^B	84	23	2	5
	m_2^B	23	76	14	4
	m_3^B	3	2	55	13

Among the 114 individuals encoded m_1^B will choose the 84 closest to the mean of individuals in modality m_1^A.

Variable recoding - Algorithm OT

Step 2 of OT algorithm: affectation

For each subject *i* of database A, a predicted value for $\hat{y}_i^B$ can be constructed by means of an **adapted nearest neighbor algorithm** accounting for the covariates with distance *d*.

Example :

		Variable Y^A			
		m_1^A	m_2^A	m_3^A	m_4^A
Variable Y^B	m_1^B	84	23	2	5
	m_2^B	23	76	14	4
	m_3^B	3	2	55	13

Among the 114 individuals encoded m_1^B will choose the 84 closest to the mean of individuals in modality m_1^A.

Variable recoding - Algorithm OT

Step 2 of OT algorithm: affectation

For each subject i of database A, a predicted value for $\hat{y}_i^B$ can be constructed by means of an **adapted nearest neighbor algorithm** accounting for the covariates with distance d.

Example :

		Variable Y^A			
		m_1^A	m_2^A	m_3^A	m_4^A
Variable Y^B	m_1^B	84	23	2	5
	m_2^B	23	76	14	4
	m_3^B	3	2	55	13

Among the 114 individuals encoded m_1^B will choose the 84 closest to the mean of individuals in modality m_1^A.

Variable recoding - simulation studies

We consider 3 dependent covariates:

- C_1 categorical with 2 modalities
- C_2 categorical with 3 modalities
- C_3 quantitative normally distributed

Construct Y from these covariates and a normally distributed error term.

- Y^A is the discretization of Y by quartiles
- Y^B is the discretization of Y by tertiles

Remark

By construction, the coefficient R^2 **measuring the association between the covariates and the outcome** depends on the simulation parameters and is easy to control.

Variable recoding - simulation studies

We consider 3 dependent covariates:

- C_1 categorical with 2 modalities
- C_2 categorical with 3 modalities
- C_3 quantitative normally distributed

Construct Y from these covariates and a normally distributed error term.

- Y^A is the discretization of Y by quartiles
- Y^B is the discretization of Y by tertiles

Remark

By construction, the coefficient R^2 **measuring the association between the covariates and the outcome** depends on the simulation parameters and is easy to control.

Variable recoding - simulation studies

We consider 3 dependent covariates:

- C_1 categorical with 2 modalities
- C_2 categorical with 3 modalities
- C_3 quantitative normally distributed

Construct Y from these covariates and a normally distributed error term.

- Y^A is the discretization of Y by quartiles
- Y^B is the discretization of Y by tertiles

Remark

By construction, the coefficient R^2 **measuring the association between the covariates and the outcome** depends on the simulation parameters and is easy to control.

Variable recoding - simulation studies

We consider 3 dependent covariates:

- C_1 categorical with 2 modalities
- C_2 categorical with 3 modalities
- C_3 quantitative normally distributed

Construct Y from these covariates and a normally distributed error term.

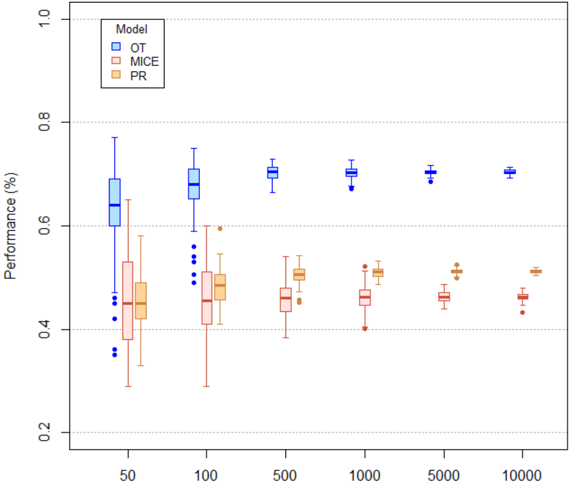
- Y^A is the discretization of Y by quartiles
- Y^B is the discretization of Y by tertiles

Remark

By construction, the coefficient R^2 **measuring the association between the covariates and the outcome** depends on the simulation parameters and is easy to control.

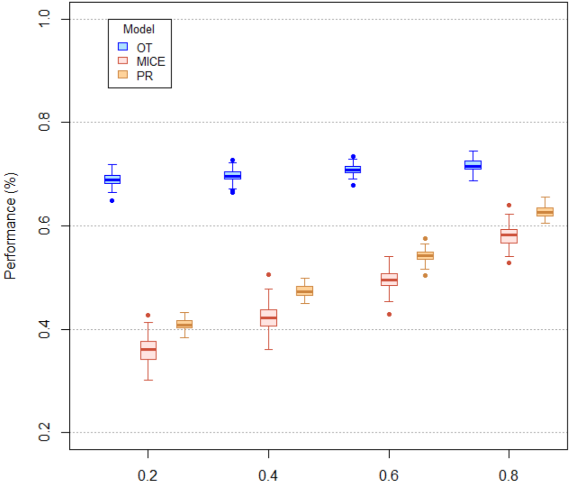
Variable recoding - simulation studies

$R^2 = 0.5$, n varies



Variable recoding - simulation studies

$n = 1000$, R^2 varies



Back to introducing example

- **NCDS (The National Child Development Study)**
  - a continuing survey which follows the lives of over 17,000 people born in England, Scotland and Wales in a same week of the year 1958
  - collects specific information on many distinct fields
  - 9 waves (0, 7, 11, 16, 22, 33, 42, 50 and 55 years old)

- **Outcome**: social status of the participants :
  - Two scales built from profession :
  - *Goldthorp social class'90 scale* (`GSS90`) : a scale in 11 categories
  - *RGs social Class'91 scale* (`RGS91`) : a scale in 6 categories.

- At wave 5, Social status assessed by **both scales**

- The database is **randomly divided in two** sub-databases of the same size
  - the `RGS91` scale is forgotten in the first sub-database
  - the `GSS90` scale is forgotten in the second sub-database
  - OT-algorithm is used to recode the variable
  - the true value of the recoded values is known : **evaluate the performances**

Back to introducing example

Social class GSS90	Database *A*	Database *B*	Social class RGS91	Database *A*	Database *B*
Modalities	n (%)	n (%)	Modalities	n (%)	n (%)
Not applicable	116 (2.89)	85 (2.12)	Not applicable	129 (3.21)	102 (2.54)
I	646 (16.09)	697 (17.36)	I	201 (5.01)	207 (5.16)
II	761 (18.95)	702 (17.48)	II	1241 (30.91)	1214 (30.24)
IIIa	650 (16.19)	683 (17.01)	IIIN	930 (23.16)	982 (24.46)
IIIb	349 (8.69)	311 (7.75)	IIIM	736 (18.33)	765 (19.05)
IVa	13 (0.32)	12 (0.30)	IV	617 (15.37)	580 (14.45)
IVb	146 (3.64)	146 (3.64)	V	161 (4.01)	165 (4.11)
IVc	27 (0.67)	31 (0.77)			
V	161 (4.01)	182 (4.53)			
VI	426 (10.61)	435 (10.83)			
VIIa	699 (17.41)	705 (17.56)			
VIIb	21 (0.52)	26 (0.65)			

OT	MICE
63.5%	29.3%

Table: NCDS study. % of well classified subjects.

Back to introducing example

Social class GSS90	Database *A*	Database *B*	Social class RGS91	Database *A*	Database *B*
Modalities	n (%)	n (%)	Modalities	n (%)	n (%)
Not applicable	116 (2.89)	85 (2.12)	Not applicable	129 (3.21)	102 (2.54)
I	646 (16.09)	697 (17.36)	I	201 (5.01)	207 (5.16)
II	761 (18.95)	702 (17.48)	II	1241 (30.91)	1214 (30.24)
IIIa	650 (16.19)	683 (17.01)	IIIN	930 (23.16)	982 (24.46)
IIIb	349 (8.69)	311 (7.75)	IIIM	736 (18.33)	765 (19.05)
IVa	13 (0.32)	12 (0.30)	IV	617 (15.37)	580 (14.45)
IVb	146 (3.64)	146 (3.64)	V	161 (4.01)	165 (4.11)
IVc	27 (0.67)	31 (0.77)			
V	161 (4.01)	182 (4.53)			
VI	426 (10.61)	435 (10.83)			
VIIa	699 (17.41)	705 (17.56)			
VIIb	21 (0.52)	26 (0.65)			

OT	MICE
63.5%	29.3%

Table: NCDS study. % of well classified subjects.

References

DE GRUYTER

The International Journal of Biostatistics. 2019; 20180106

Valérie Gares[1] / Chloé Dimeglio[2] / Grégory Guernec[3] / Romain Fantin[4] / Benoit Lepage[5] / Michael R. Kosorok[6] / Nicolas Savy[7]

On the Use of Optimal Transportation Theory to Recode Variables and Application to Database Merging

Thank you for your attention...