Projet BIDASA ”Big DAta en SAnté”

Journées de restitution
MASTODONS’ 2017
Paris
Vendredi 14 juin 2019

Nicolas SAVY

Institut de Mathématiques de Toulouse

Content of the talk

Context of BiDaSa Project

- Health data, is an **unusal Big Data**

- Because 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
  - ...

  $\Longrightarrow$ Components of the "definition" of "Big Data" are partially met
  Volume ($\pm$) - **V**elocity (-) - **V**ariety (+) - **V**eracity ($\pm$) - **V**alue (++)

  $\Longrightarrow$ To Exploit of these multitudes of databases ask questions
  - **Data chaining** to construct care path
  - **Data merging** to enlarge database
  - **Data Reuse**
  - ...

Context of BiDaSa Project

- Health data, is an **unusal Big Data**

- Because 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
  - ...

$\Longrightarrow$ Components of the "definition" of "Big Data" are partially met
Volume ($\pm$) - **V**elocity (-) - **V**ariety (+) - **V**eracity ($\pm$) - **V**alue (++)

$\Longrightarrow$ To Exploit of these multitudes of databases ask questions
- **Data chaining** to construct care path
- **Data merging** to enlarge database
- **Data Reuse**
- ...

Context of BiDaSa Project

- Health data, is an **unusal Big Data**

- Because 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
  - ...

  $\Longrightarrow$ Components of the "definition" of "Big Data" are partially met
  Volume ($\pm$) - **V**elocity (-) - **V**ariety (+) - **V**eracity ($\pm$) - **V**alue (++)

  $\Longrightarrow$ To Exploit of these multitudes of databases ask questions
  - **Data chaining** to construct care path
  - **Data merging** to enlarge database
  - **Data Reuse**
  - ...

Context of BiDaSa Project

- Health data, is an **unusal Big Data**

- Because 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
  - ...

  $\Longrightarrow$ Components of the "definition" of "Big Data" are partially met
  Volume ($\pm$) - **V**elocity (-) - **V**ariety (+) - **V**eracity ($\pm$) - **V**alue (++)

  $\Longrightarrow$ To Exploit of these multitudes of databases ask questions
  - **Data chaining** to construct care path
  - **Data merging** to enlarge database
  - **Data Reuse**
  - ...

Context of BiDaSa Project

In this context, **BiDaSa project** pays attention to

- **Axe 1: Data ReUse**
  The calibration of models for In Silico Clinical Trials
- **Axe 2: Data Merging**
  Recoding of variables

To do so, **the consortium** is composed of

Context of BiDaSa Project

In this context, **BiDaSa project** pays attention to

- **Axe 1: Data ReUse**

  The calibration of models for In Silico Clinical Trials

- **Axe 2: Data Merging**

  Recoding of variables

To do so, **the consortium** is composed of

- **Institut de Mathmatiques de Toulouse**
  - Nicolas SAVY (MCF) - PI
  - Sébastien DEJEAN (IR)
  - Philippe SAINT-PIERRE (MCF)
- **Unit INSERM Epidmiologie Toulouse**
  - Thierry LANG (PU-PH)
  - Grégory GUERNEC (IR)
  - Chloé DIMEGLIO (IR)
  - Valérie GARES (IR)
  - Benoit LEPAGE (MCU-PH)

Context of BiDaSa Project

In this context, **BiDaSa project** pays attention to

- **Axe 1: Data ReUse**

  The calibration of models for In Silico Clinical Trials

- **Axe 2: Data Merging**

  Recoding of variables

To do so, **the consortium** is composed of

- **Institut de Mathématiques de Toulouse**
  - Nicolas SAVY (MCF) - PI
  - Sébastien DEJEAN (IR)
  - Philippe SAINT-PIERRE (MCF)
- **Unité INSERM Epidémiologie Toulouse**
  - Thierry LANG (PU-PH)
  - Grégory GUERNEC (IR)
  - Benoit LEPAGE (MCU-PH)
- **Institut de Recherche en Mathématiques de Rennes**
  - Valérie GARES (MCF)
  - Jérémy OMER (MCF)

Content of the talk

Axe 1: In Silico Clinical Trials

Make use of the **available knowledge** about

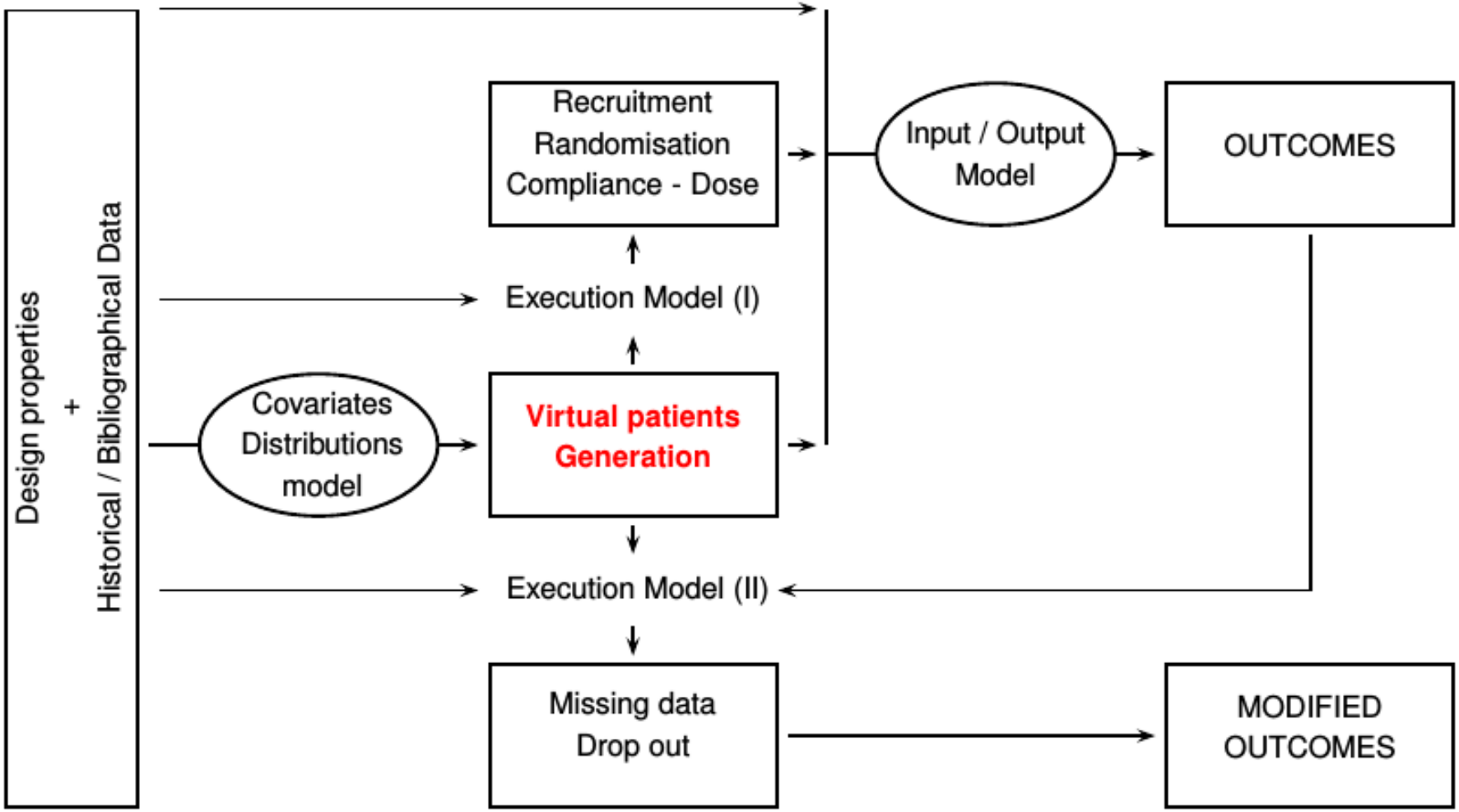
- drug
- patients
- disease progression
- clinical program

to investigate **in silico** aspects of the clinical study plan

- dose regimen
- study design
- patients population
- ...

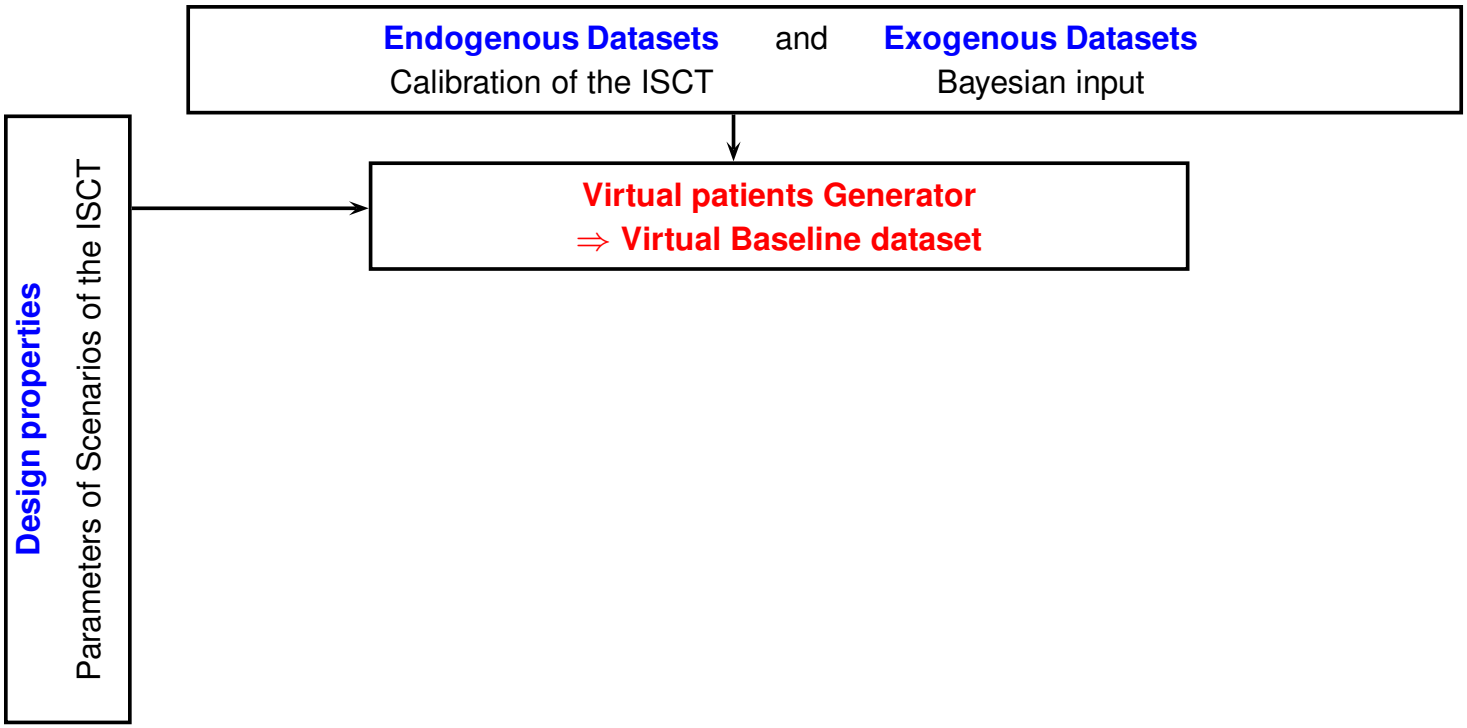
in order to make **rational, informed decision** with regards to **optimizing** the development plan of a new compound

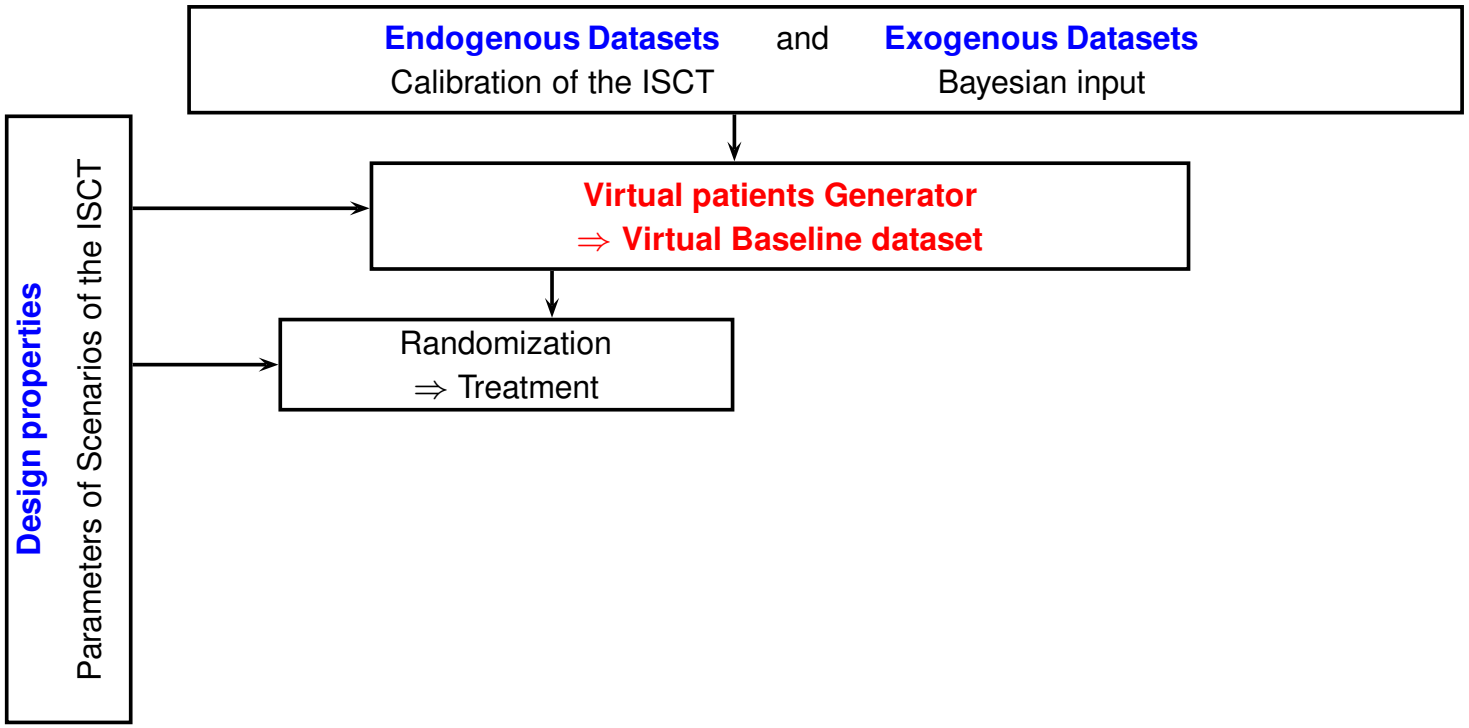
Axe 1: In Silico Clinical Trials



Endogenous Datasets and **Exogenous Datasets**
Calibration of the ISCT | Bayesian input

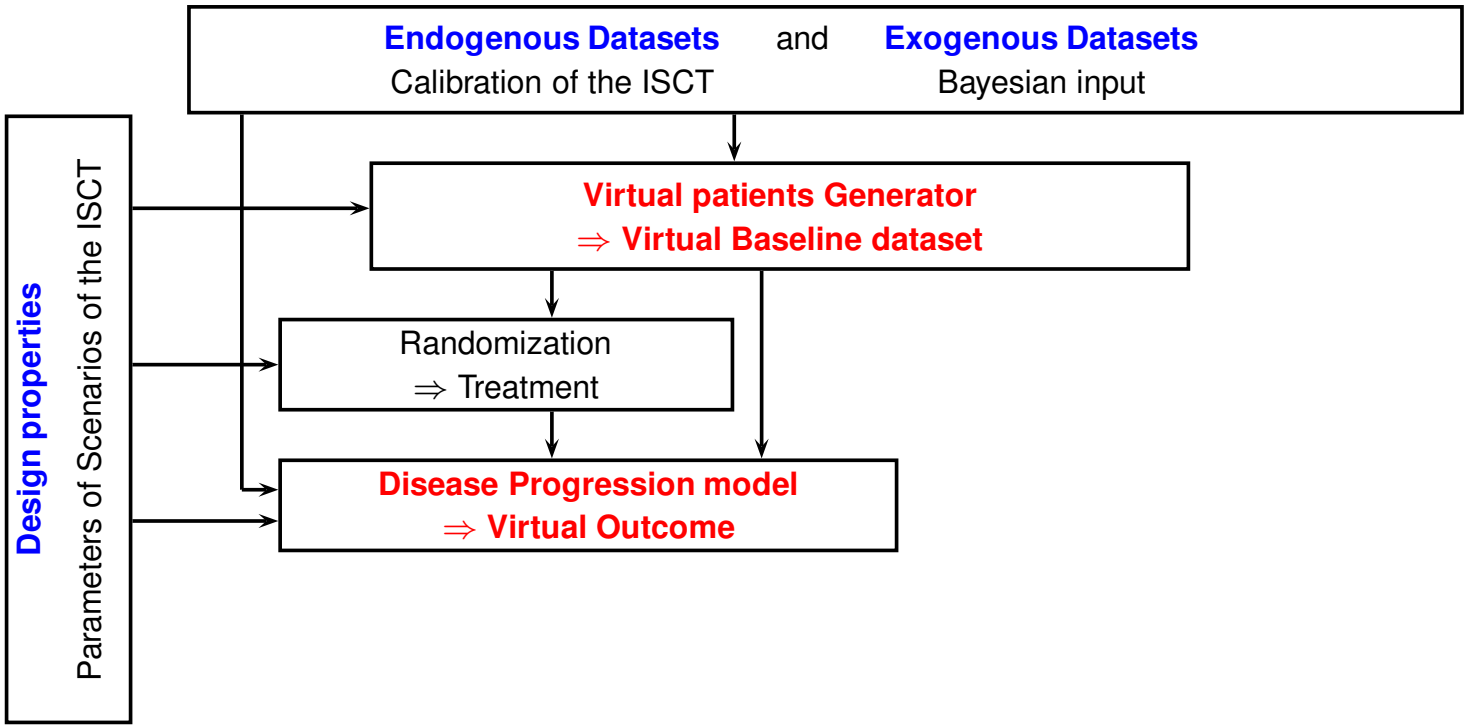
Design properties
Parameters of Scenarios of the ISCT

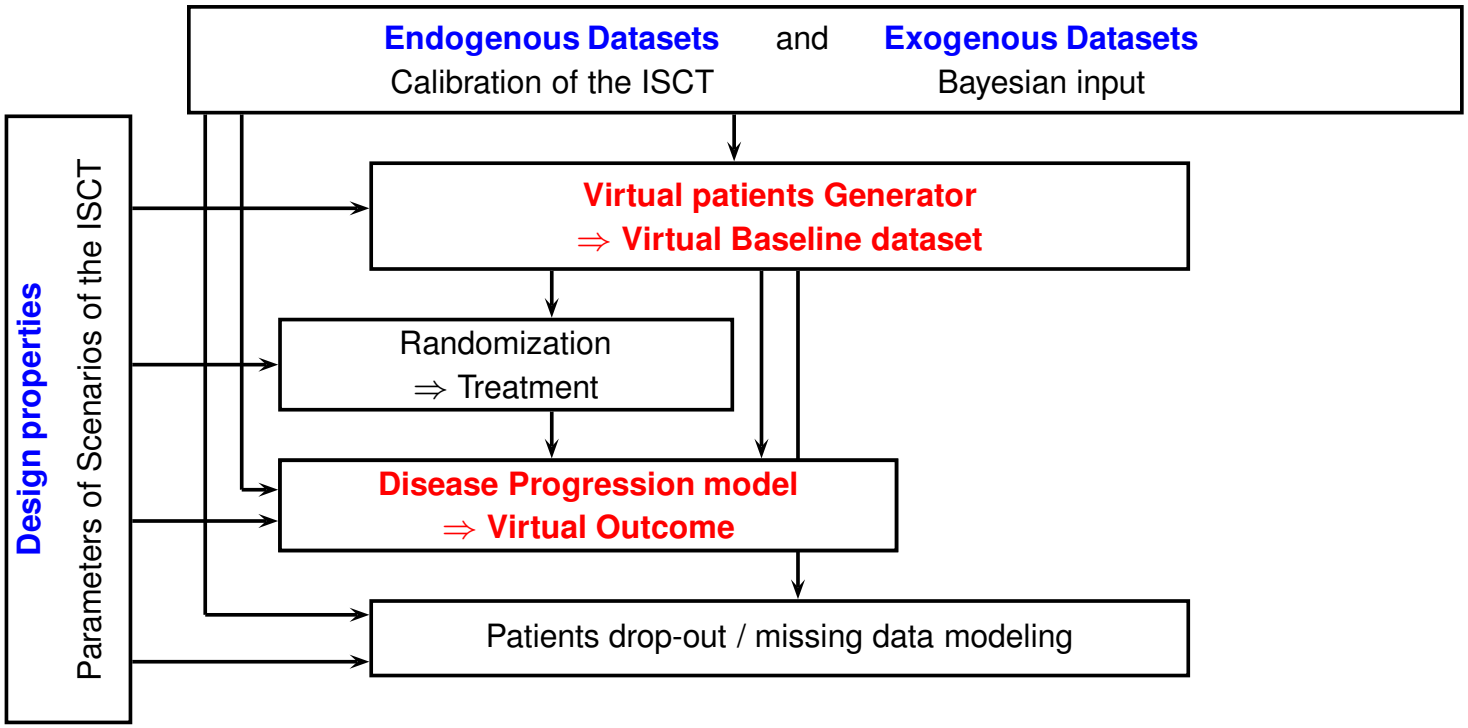

Endogenous Datasets
Calibration of the ISCT
and
Exogenous Datasets
Bayesian input
Virtual patients Generator
⇒ Virtual Baseline dataset
Design properties
Parameters of Scenarios of the ISCT

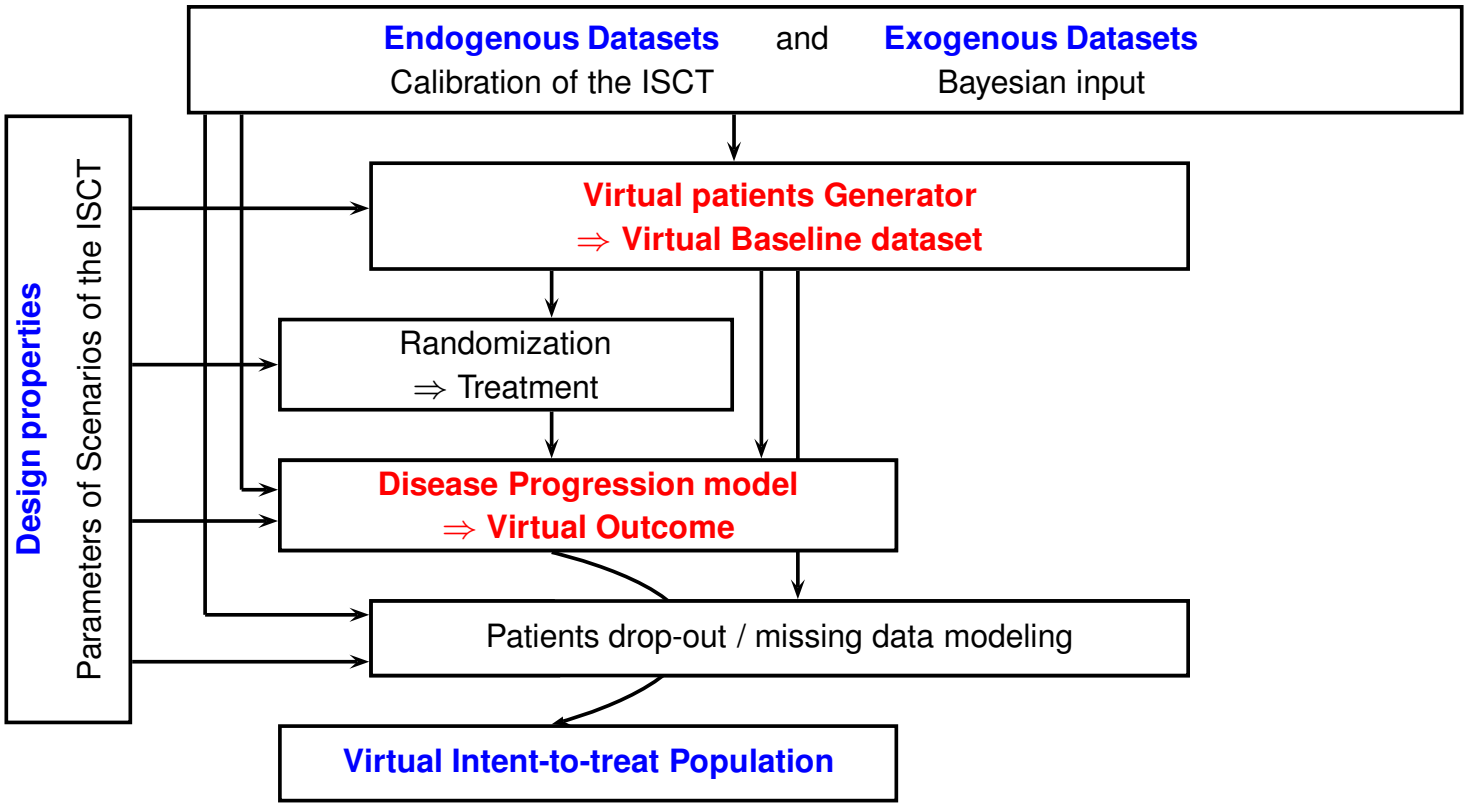

Endogenous Datasets and Exogenous Datasets
Calibration of the ISCT
Bayesian input
Virtual patients Generator
⇒ Virtual Baseline dataset
Randomization
⇒ Treatment
Design properties
Parameters of Scenarios of the ISCT

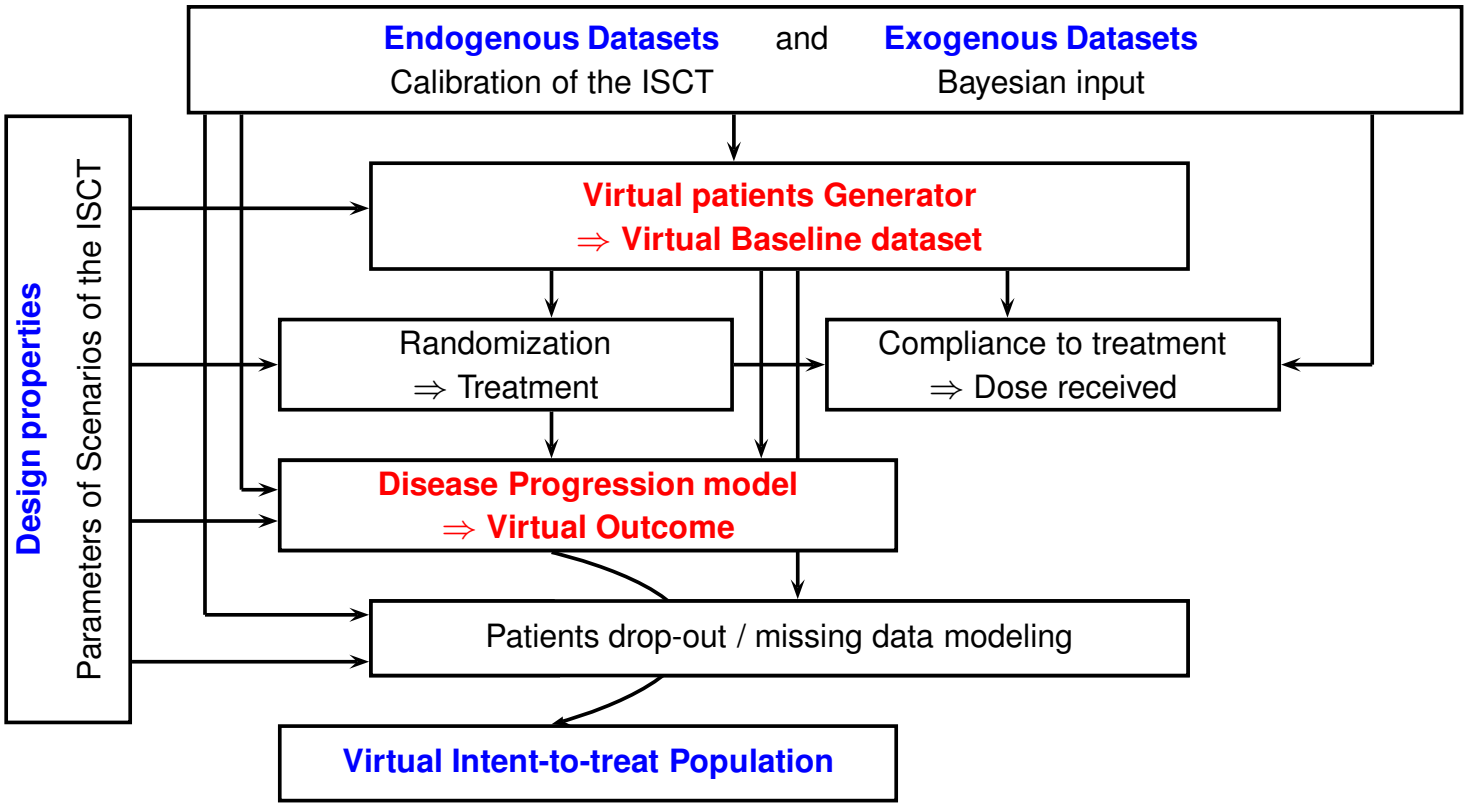
Axe 1: In Silico Clinical Trials

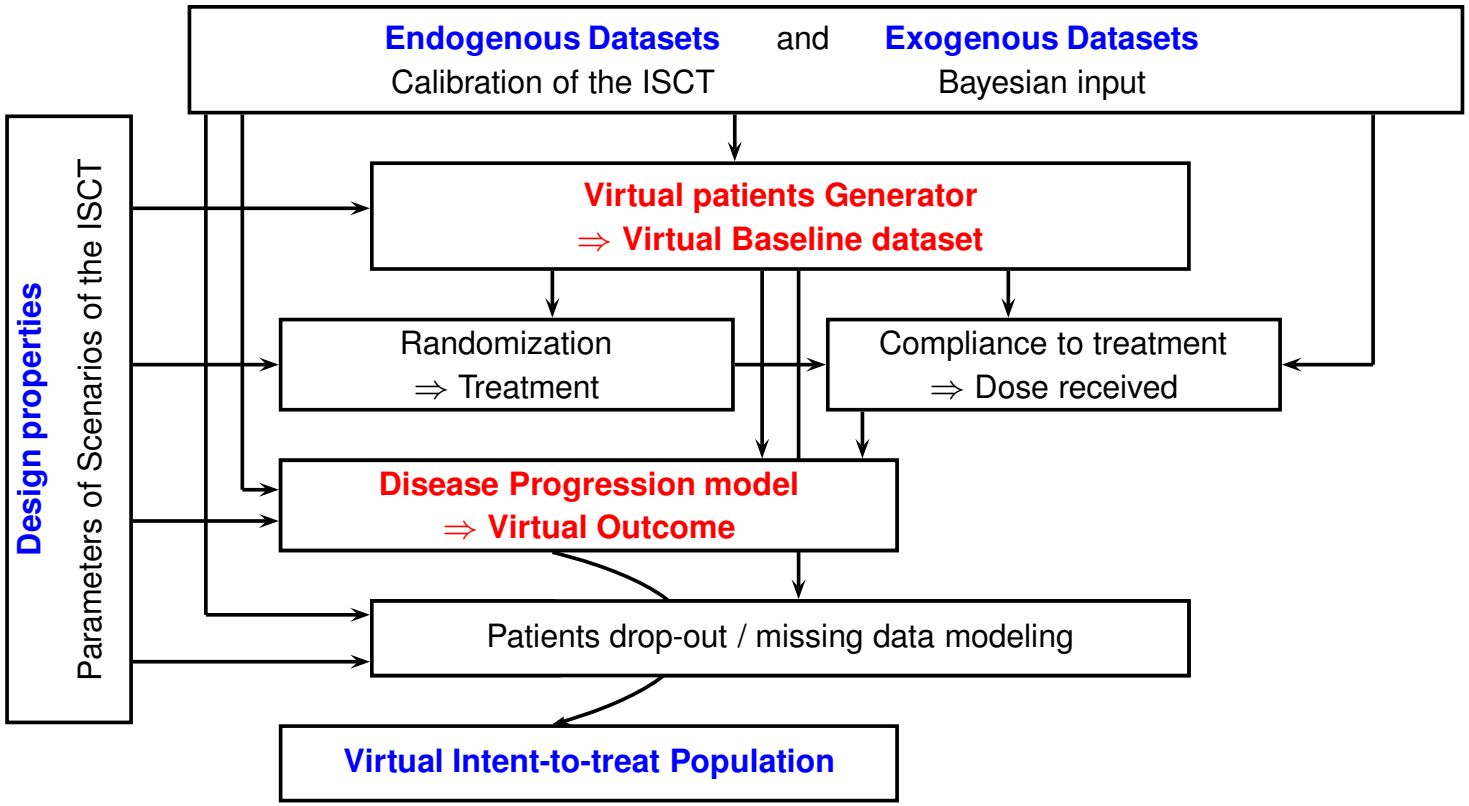
General Schema

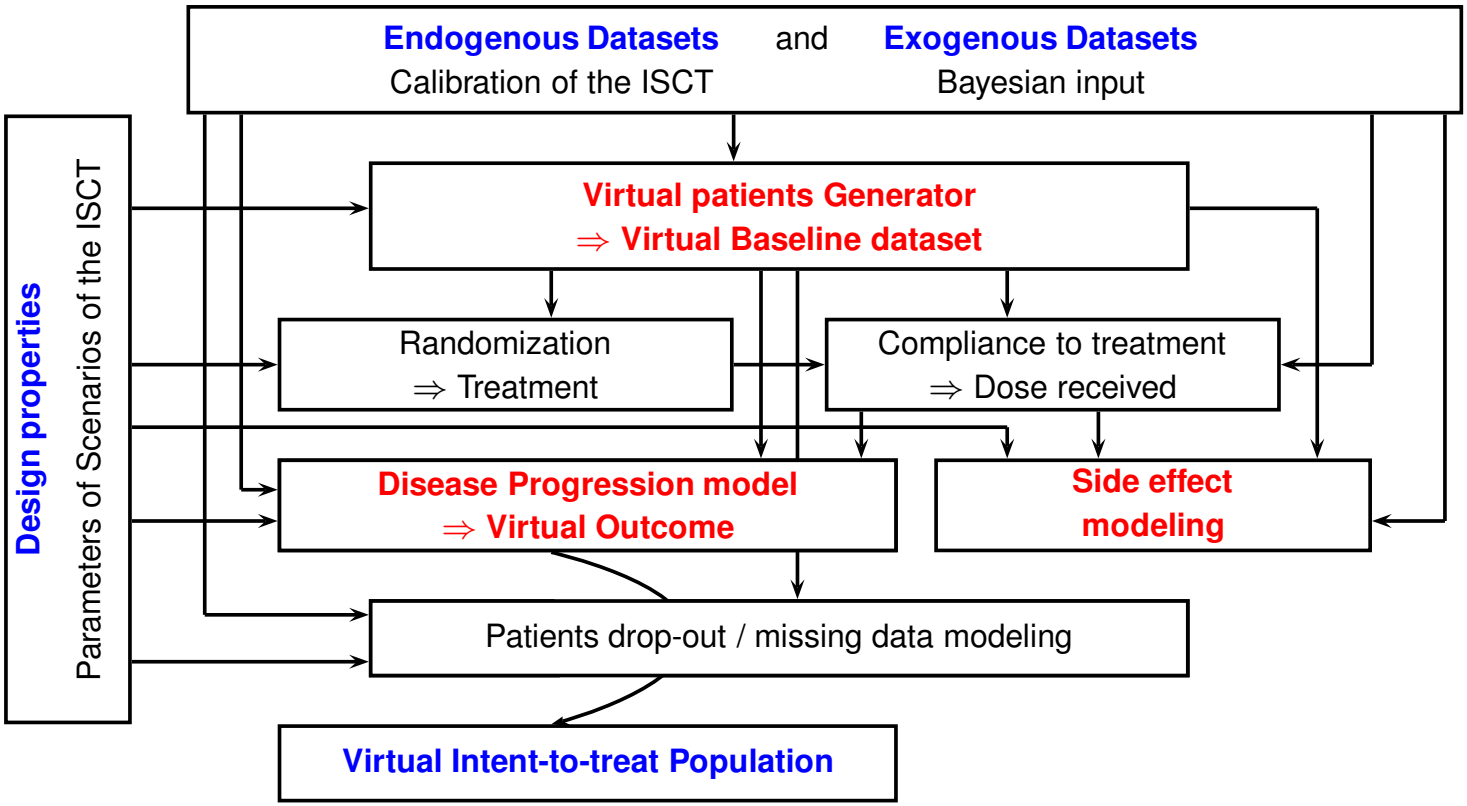


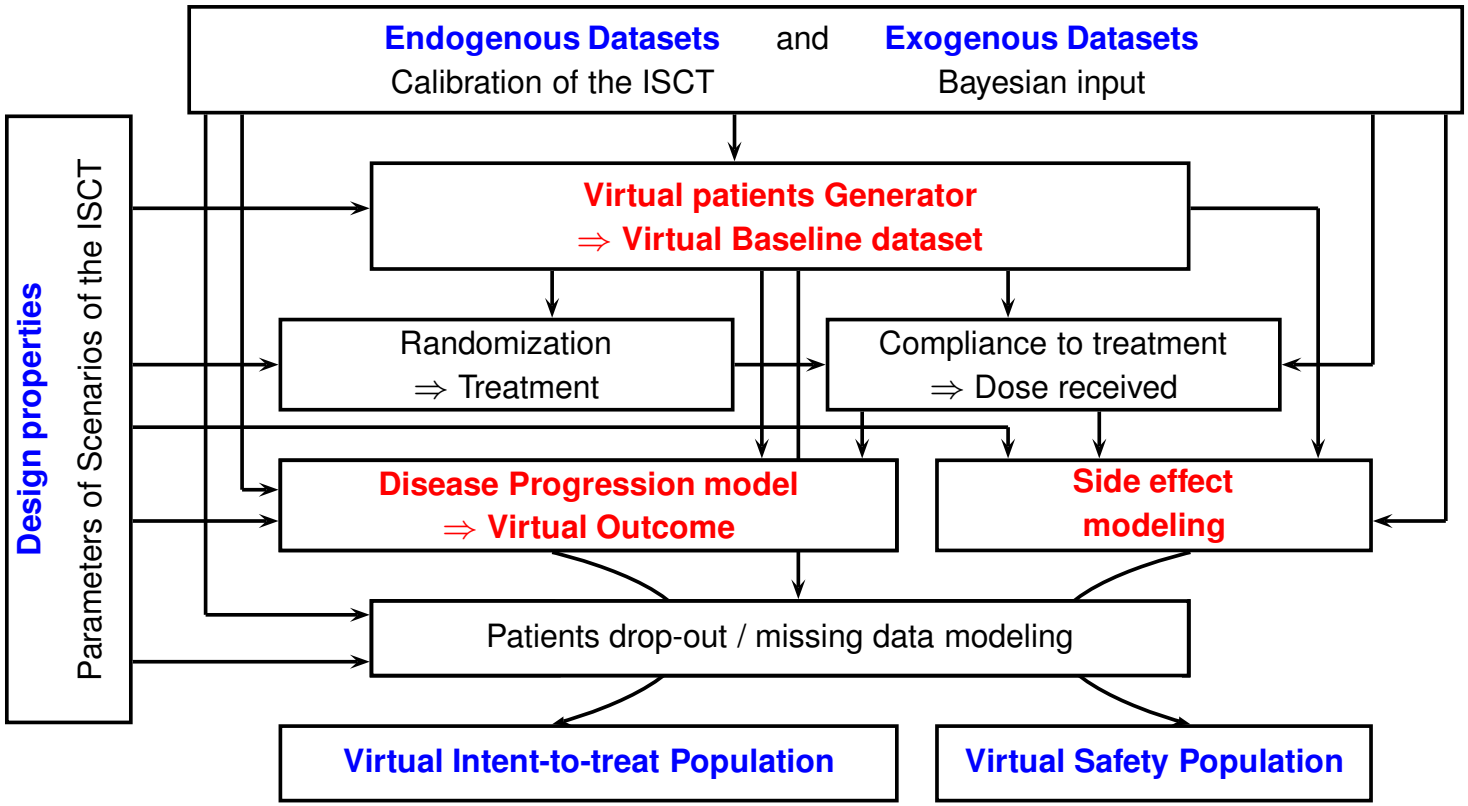

Endogenous Datasets and Exogenous Datasets
Calibration of the ISCT
Bayesian input
Design properties
Parameters of Scenarios of the ISCT
Virtual patients Generator
⇒ Virtual Baseline dataset
Randomization
⇒ Treatment
Disease Progression model
⇒ Virtual Outcome
Patients drop-out / missing data modeling

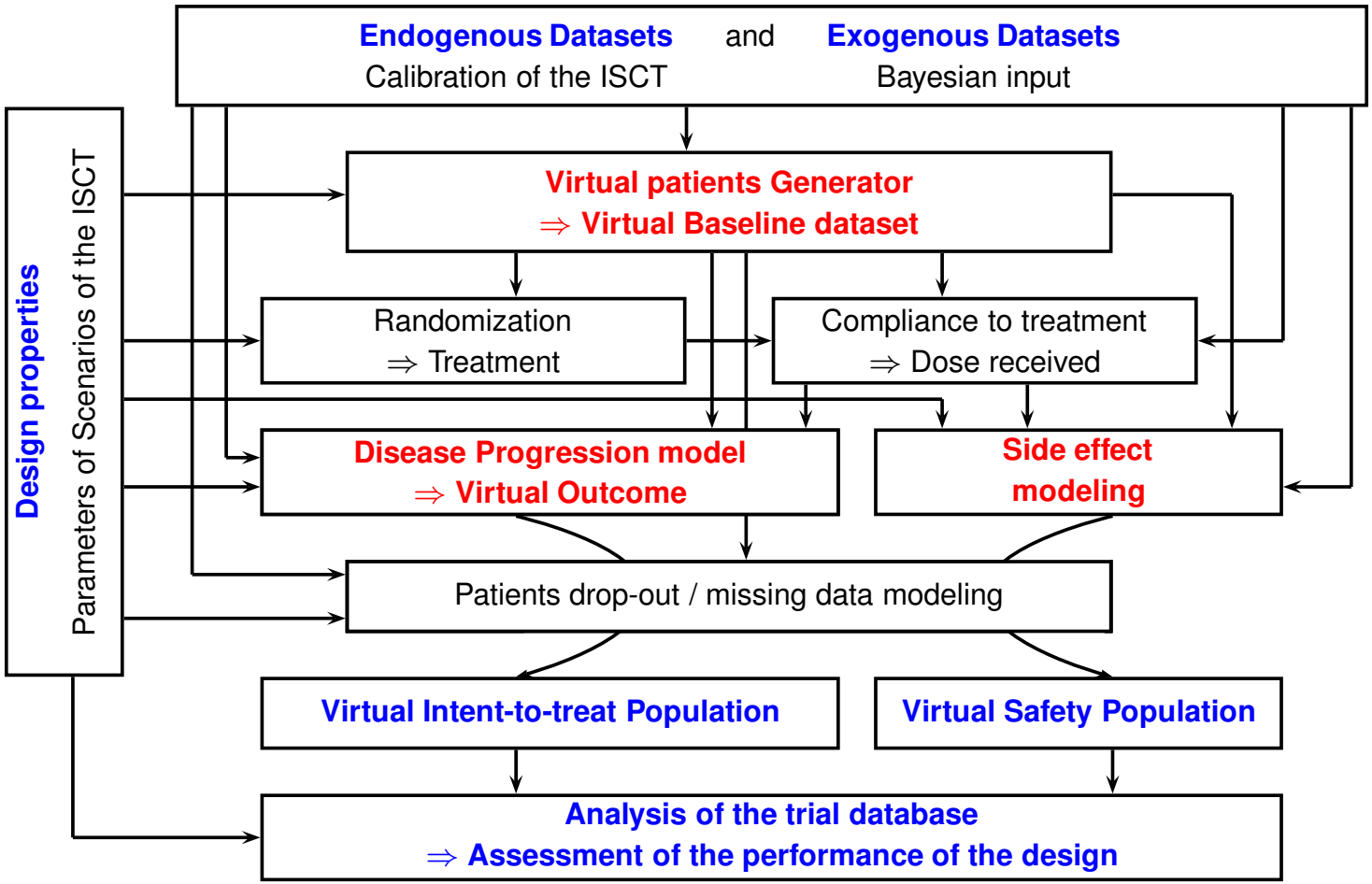

Endogenous Datasets and Exogenous Datasets
Calibration of the ISCT
Bayesian input
Design properties
Parameters of Scenarios of the ISCT
Virtual patients Generator
⇒ Virtual Baseline dataset
Randomization
⇒ Treatment
Disease Progression model
⇒ Virtual Outcome
Patients drop-out / missing data modeling
Virtual Intent-to-treat Population


Endogenous Datasets and Exogenous Datasets
Calibration of the ISCT
Bayesian input
Design properties
Parameters of Scenarios of the ISCT
Virtual patients Generator
⇒ Virtual Baseline dataset
Randomization
⇒ Treatment
Compliance to treatment
⇒ Dose received
Disease Progression model
⇒ Virtual Outcome
Patients drop-out / missing data modeling
Virtual Intent-to-treat Population


Endogenous Datasets and Exogenous Datasets
Calibration of the ISCT
Bayesian input
Design properties
Parameters of Scenarios of the ISCT
Virtual patients Generator
⇒ Virtual Baseline dataset
Randomization
⇒ Treatment
Compliance to treatment
⇒ Dose received
Disease Progression model
⇒ Virtual Outcome
Patients drop-out / missing data modeling
Virtual Intent-to-treat Population


Endogenous Datasets and Exogenous Datasets
Calibration of the ISCT
Bayesian input
Design properties
Parameters of Scenarios of the ISCT
Virtual patients Generator
⇒ Virtual Baseline dataset
Randomization
⇒ Treatment
Compliance to treatment
⇒ Dose received
Disease Progression model
⇒ Virtual Outcome
Side effect
modeling
Patients drop-out / missing data modeling
Virtual Intent-to-treat Population


Endogenous Datasets and Exogenous Datasets
Calibration of the ISCT
Bayesian input
Design properties
Parameters of Scenarios of the ISCT
Virtual patients Generator
⇒ Virtual Baseline dataset
Randomization
⇒ Treatment
Compliance to treatment
⇒ Dose received
Disease Progression model
⇒ Virtual Outcome
Side effect
modeling
Patients drop-out / missing data modeling
Virtual Intent-to-treat Population
Virtual Safety Population


Endogenous Datasets and Exogenous Datasets
Calibration of the ISCT
Bayesian input
Design properties
Parameters of Scenarios of the ISCT
Virtual patients Generator
⇒ Virtual Baseline dataset
Randomization
⇒ Treatment
Compliance to treatment
⇒ Dose received
Disease Progression model
⇒ Virtual Outcome
Side effect
modeling
Patients drop-out / missing data modeling
Virtual Intent-to-treat Population
Virtual Safety Population
Analysis of the trial database
⇒ Assessment of the performance of the design

- 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

- 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

- 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

- Huge **diversity of models** may be considered essentially based on
  - **Parametric models** (Markov, Cox, linear, logistics,...)
    - Problem of database for parameters estimation
  - **Non-parametric models** (Machine learning)
    - Problem of database for learning model

- Main difficulty comes from **Calibration / Learning** database
  - The size of learning database (may be irrelevent with the objective)
  - Origin of the data (completed trials data, real word databases)
  - Problem of portability of those data
    - For example: How to use US data to simulate an European trial

- 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

- Huge **diversity of models** may be considered essentially based on
  - **Parametric models** (Markov, Cox, linear, logistics,...)
    - Problem of database for parameters estimation
  - **Non-parametric models** (Machine learning)
    - Problem of database for learning model

- Main difficulty comes from **Calibration / Learning** database
  - The size of learning database (may be irrelevent with the objective)
  - Origin of the data (completed trials data, real word databases)
  - Problem of portability of those data
    - For example: How to use US data to simulate an European trial

- 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

- Huge **diversity of models** may be considered essentially based on
  - **Parametric models** (Markov, Cox, linear, logistics,...)
    - Problem of database for parameters estimation
  - **Non-parametric models** (Machine learning)
    - Problem of database for learning model

- Main difficulty comes from **Calibration / Learning** database
  - The size of learning database (may be irrelevent with the objective)
  - Origin of the data (completed trials data, real word databases)
  - Problem of portability of those data
    - For example: How to use US data to simulate an European trial

- 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

Content of the talk

Variables recoding issue

Database A

	C_1	C_2	...	C_p	Y^A	Y^B
1					Observed	Unobserved
...						
...						
n_A						

Database B

	C_1	C_2	...	C_p	Y^A	Y^B
1					Unobserved	Observed
...						
...						
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	...	C_p	Y^A	Y^B
1					Observed	Unobserved
...						
...						
n_A						

Database B

	C_1	C_2	...	C_p	Y^A	Y^B
1					Unobserved	Observed
...						
...						
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	...	C_p	Y^A	Y^B
1					Observed	Unobserved
...						
...						
n_A						

Database B

	C_1	C_2	...	C_p	Y^A	Y^B
1					Unobserved	Observed
...						
...						
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	...	C_p	Y^A	Y^B
1					Observed	Unobserved
...						
...						
n_A						

Database B

	C_1	C_2	...	C_p	Y^A	Y^B
1					Unobserved	Observed
...						
...						
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

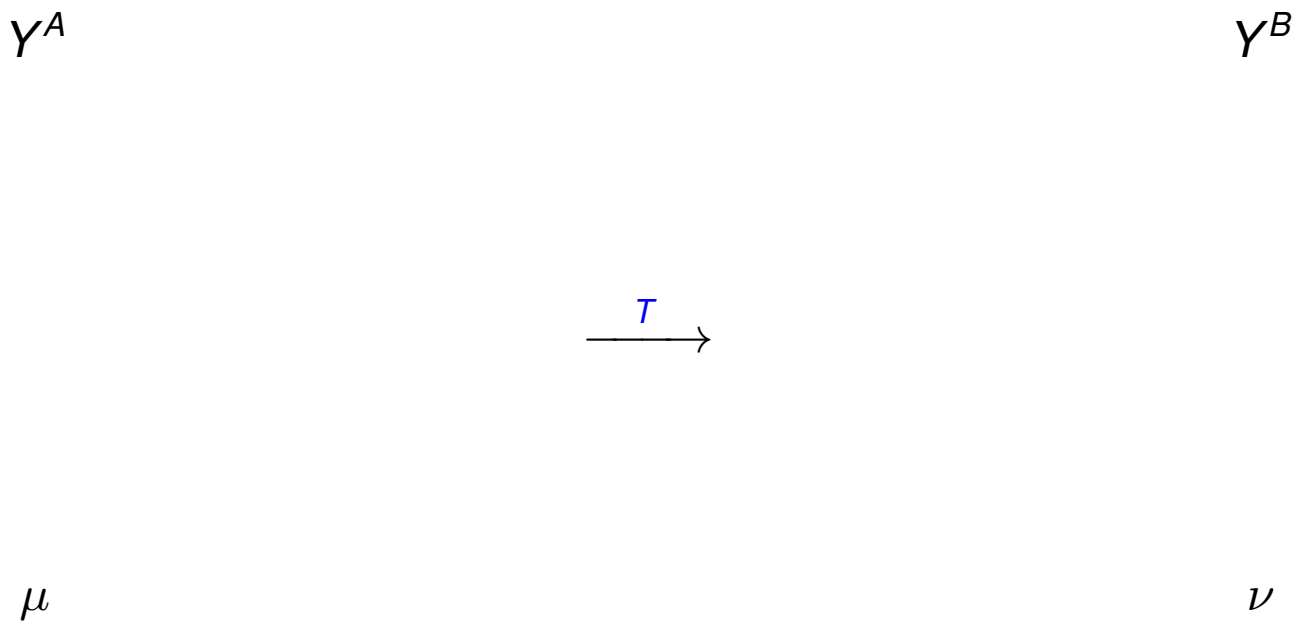
- Ideas
  - Missing data problem (MAR)
  - Latent variables models (class latent analysis, trait latent analysis)
  - Estimation (polytomous regression) / Prediction
  - **Optimal transportation**

Optimal Transportation

Y^A Y^B

μ ν

Optimal Transportation



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

Optimal Transportation

Y^A Y^B

T

μ ν

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

Y^A Y^B

T

μ 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

- $\mathbb{Y}^A$ 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*

- $\mathbb{Y}^A$ 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:

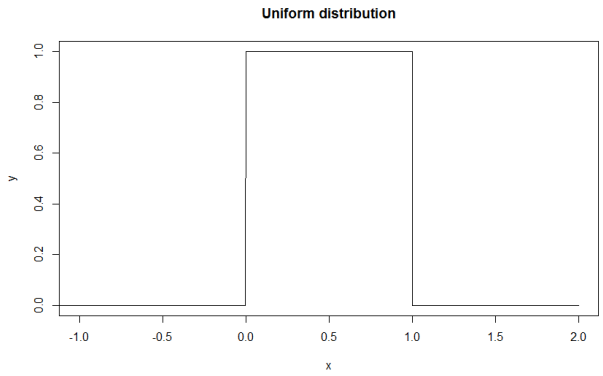
$$\left\{ \int_{\mathbb{Y}^A \times \mathbb{Y}^B} c\left(y^A, y^B\right) \mathrm{d}\gamma\left(y^A, y^B\right) \;\middle|\; \gamma \in \gamma(\mu, \nu) \right\},$$

  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 $\mathbb{Y}^B$

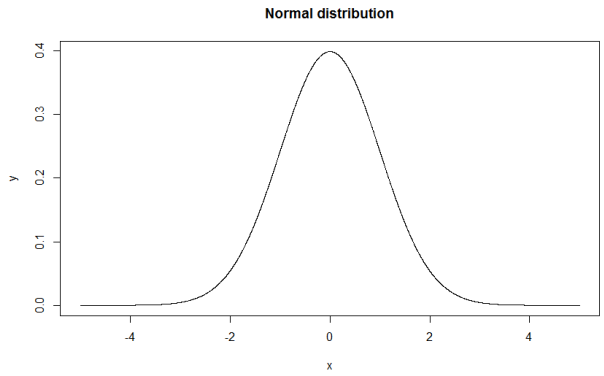
  - *Linear problem, solution achievable with compacity (volume fitting) argument*

- **Continuous case**

$\mathcal{U}[0,1]$

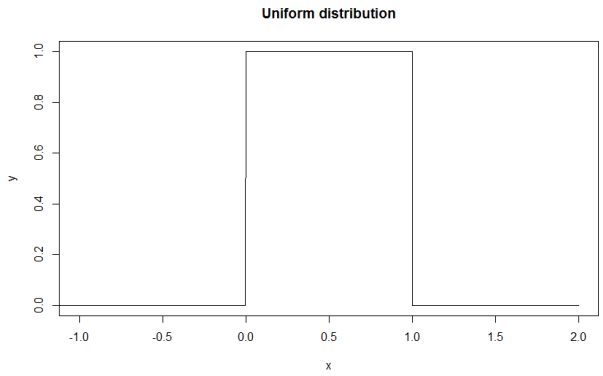


$\mathcal{N}(0,1)$

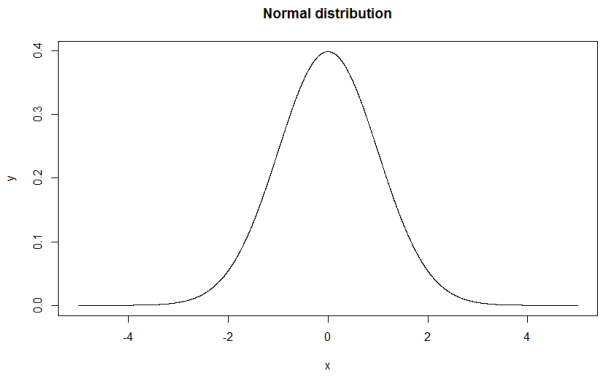


- **Continuous case**

$\mathcal{U}[0,1]$ $\qquad\qquad$ $\mathcal{N}(0,1)$

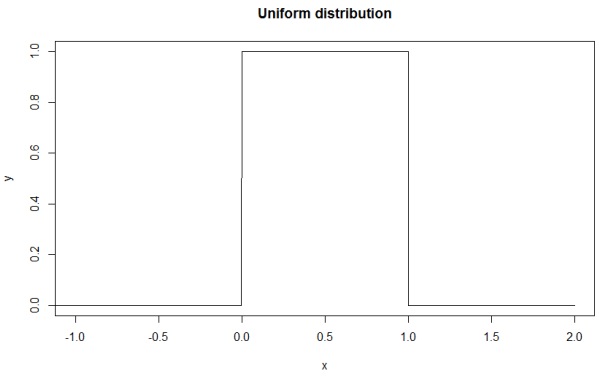


$$\xrightarrow[c(x,y)=(x-y)^2]{T^*(x)=\frac{1}{2\pi}\exp\left(-\frac{x^2}{2}\right)}$$

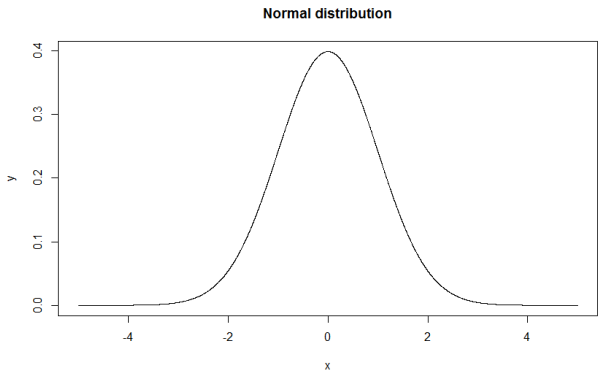


- **Continuous case**

$\mathcal{U}[0,1]$ $\qquad\qquad$ $\mathcal{N}(0,1)$



$$\xrightarrow[c(x,y)=(x-y)^2]{T^*(x)=\frac{1}{2\pi}\exp\left(-\frac{x^2}{2}\right)}$$



- 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

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

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

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

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

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

On the use of optimal transportation theory to recode variables and application to database merging.

Valérie Garès, Chloé Dimeglio, Grégory Guernec, Romain Fantin, Benoit Lepage, Michael R. Kosorok and Nicolas Savy

Forthcoming in International Journal of Biostatistics and Available on HAL

This paper presents 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

On the use of optimal transportation theory to recode variables and application to database merging.

Valérie Garès, Chloé Dimeglio, Grégory Guernec, Romain Fantin, Benoit Lepage, Michael R. Kosorok and Nicolas Savy

Forthcoming in International Journal of Biostatistics and Available on HAL

This paper presents 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

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

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

under the following constraints

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

- 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

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

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

- 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

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

$$= 0 \text{ otherwise}$$

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

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

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

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

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

- The cost function is estimated by

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

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

- 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

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

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

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				

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

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 ?

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.

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.

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.

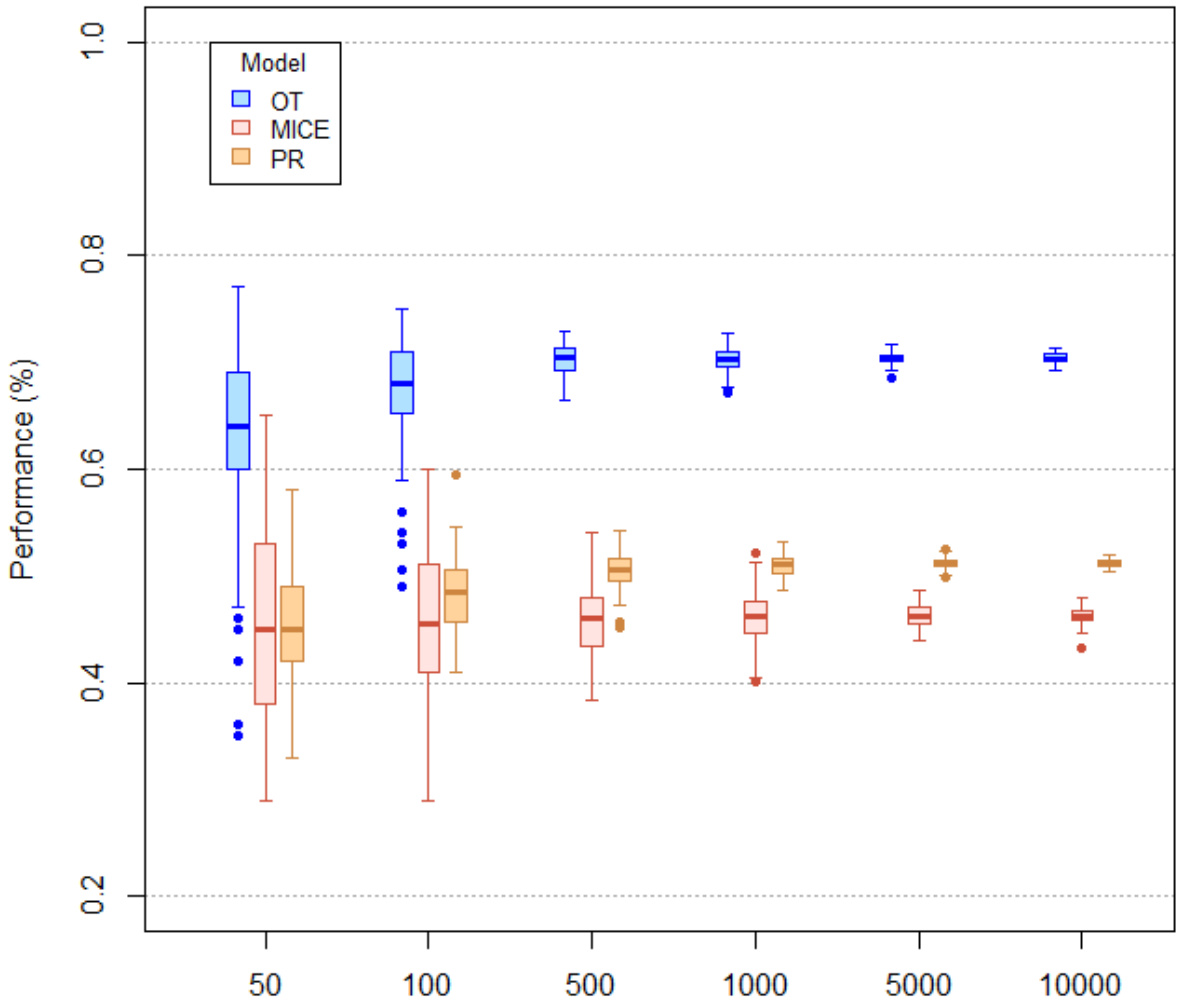
- Y^A is the discretization of Y by quartiles
- Y^B is the discretization of Y by tertiles

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.

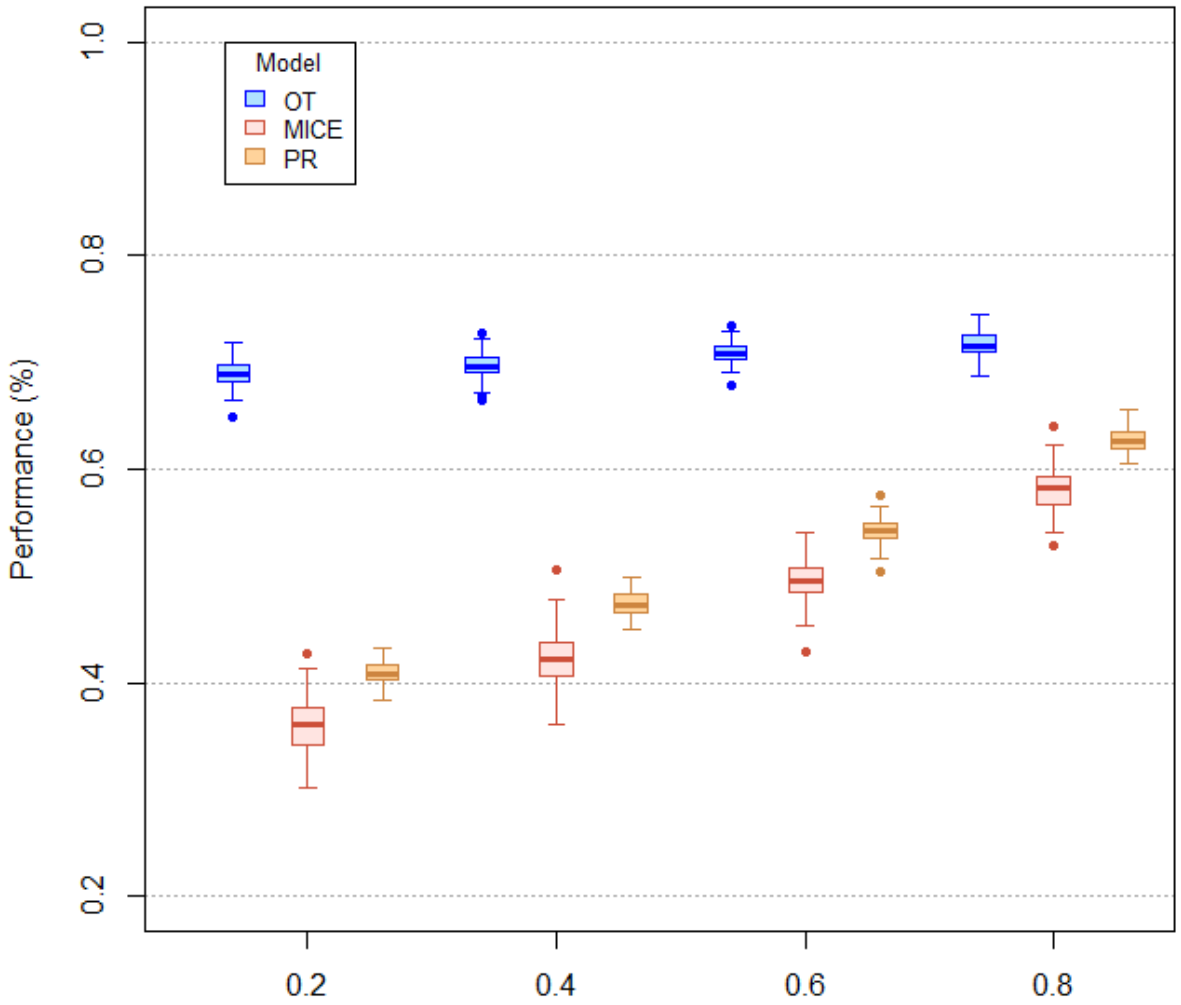
Variable recoding - simulation studies

$R^2 = 0.5$, n varies



Variable recoding - simulation studies

$n = 1000$, R^2 varies



Variable recoding - real life 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**: two measurements scales of the **social class** of the participants built from profession and collected at wave 5:
  - *Goldthorp social class'90 scale* (`GSS90`) : a scale in 11 categories
  - *RGs social Class'91 scale* (`RGS91`) : a scale in 6 categories.

- The initial database was randomly divided in two databases of the same size and we kept
  - the `GSS90` scale in the first database
  - the `RGS91` scale in the second database

Variable recoding - real life 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**: two measurements scales of the **social class** of the participants built from profession and collected at wave 5:
  - *Goldthorp social class'90 scale* (`GSS90`) : a scale in 11 categories
  - *RGs social Class'91 scale* (`RGS91`) : a scale in 6 categories.

- The initial database was randomly divided in two databases of the same size and we kept
  - the `GSS90` scale in the first database
  - the `RGS91` scale in the second database

Variable recoding - real life 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**: two measurements scales of the **social class** of the participants built from profession and collected at wave 5:
  - *Goldthorp social class'90 scale* (`GSS90`) : a scale in 11 categories
  - *RGs social Class'91 scale* (`RGS91`) : a scale in 6 categories.

- The initial database was randomly divided in two databases of the same size and we kept
  - the `GSS90` scale in the first database
  - the `RGS91` scale in the second database

Variable recoding - real life 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.

Variable recoding - real life 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.

Works in progress

- **Improvment** of OT-Algorithm by relaxation of hypotheses
  - ↬ V. Garès, J. Omer
    Recoding Variable with domain adaptation using optimal transport
    Submitted to Journal of American Statistician Association

- **Missing data** in covariates issue
  - ↬ G. Guernec, V. Garès, P. Saint-Pierre, B. Lepage, M. Kosorok, C. Diméglio and N. Savy
    - On the use of OT-algorithm when there are missing data in covariates. Application to clinical research
    - In preparation for Statistical Methods for Medical Research

- **Challenging** of OT algorithm by comparison with Machine Learning strategies
  - ↬ D. Vuillemenot, C. Diméglio, V. Garès, G. Guernec, B. Lepage, P. Muller, M. Kosorok, N. Savy and P. Saint-Pierre
    - Comparison of OT-algorithm and machine learning approach to merge databases. Application to clinical research.
    - In preparation for Computational Statistics & Data Analysis

- **Preparation** of a R packaqe

  - ↬ G. Guernec, V. Garès, P. Navaro, J. Omer, P. Saint-Pierre, N. Savy
    - OTRECOD: Using Optimal Transport theory in database fusion for recoding heterogeneous variables
    - Poster for UseR Conference, Toulouse, july 2019

Content of the talk

Project OT4DI

- Project **O**ptimal **T**ransportation for **D**ata **I**ntegration

- Appel Flash Science Ouverte

  ”Pratiques de recherche et données ouvertes” de l’ANR

- Aims to **investigate**
  - Feasibility of Extension to multidimensional framework
  - Various questions on Data Integration by means of Optimal Transportation
  - Various theoretical aspects on Statistics and Optimal Transportation theory
  - Feasibility to perform in high dimension setting

Project OT4DI

- Project **O**ptimal **T**ransportation for **D**ata **I**ntegration

- Appel Flash Science Ouverte

  ”Pratiques de recherche et données ouvertes” de l’ANR

- Aims to **investigate**
  - Feasibility of Extension to multidimensional framework
  - Various questions on Data Integration by means of Optimal Transportation
  - Various theoretical aspects on Statistics and Optimal Transportation theory
  - Feasibility to perform in high dimension setting

The consortium is:

- **Institut de Mathématiques de Toulouse**
  - Nicolas SAVY (MCF) - PI
  - Sébastien DEJEAN (IR)
  - Philippe SAINT-PIERRE (MCF)
  - **Jean-Michel Loubès (PR)**
- **Unité INSERM Epidémiologie Toulouse**
  - Grégory GUERNEC (IR)
- **Institut de Recherche en Mathématiques de Rennes**
  - Valérie GARES (MCF)
  - **Mounir Haddou (PR)**
- **University of North Carolina at Chapel Hill - USA**
  - **Michael Kosorok (PR)**
- **Mc Gill University - CAN**
  - **Erica Moodie (MCF)**

A big thank you to the CNRS and Mastodons for the support of the BiDaSa project

Thank you for your attention...

A big thank you to the CNRS and Mastodons for the support of the BiDaSa project

Thank you for your attention...