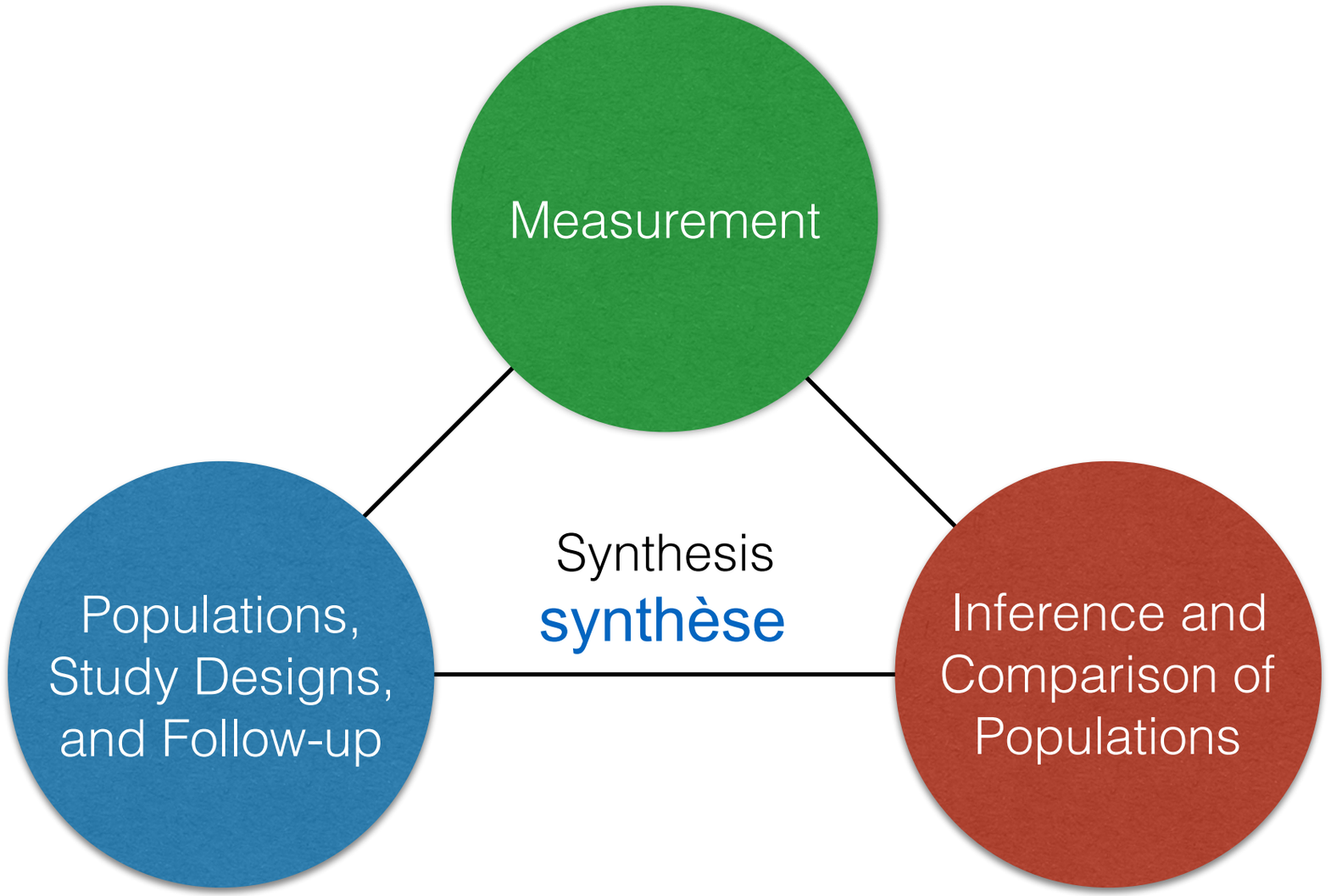
Epidemiological Study Design

Amy Wesolowski
Department of Epidemiology

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

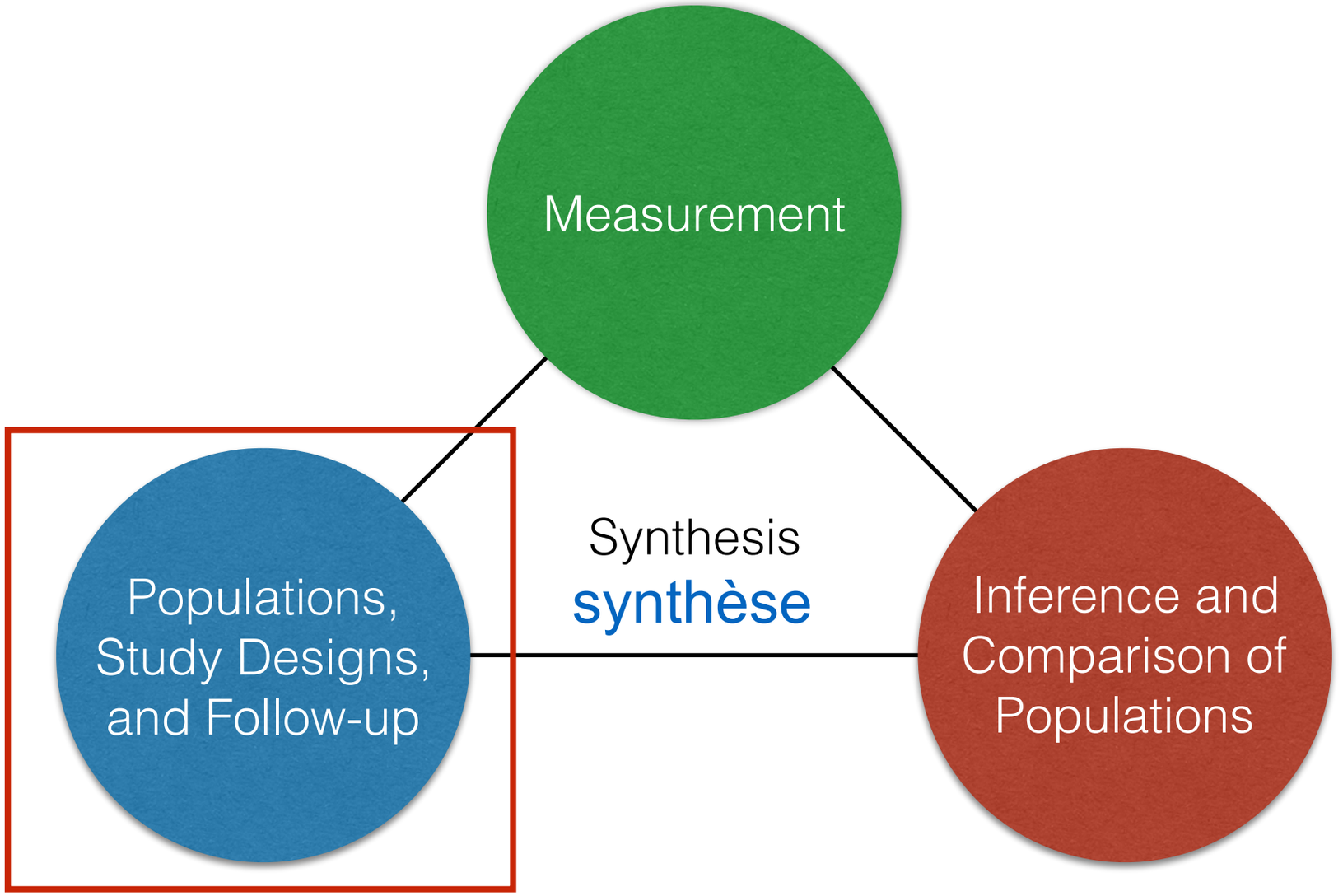
la mesure



populations, plan d'étude et suivi

inférence et comparaison des populations

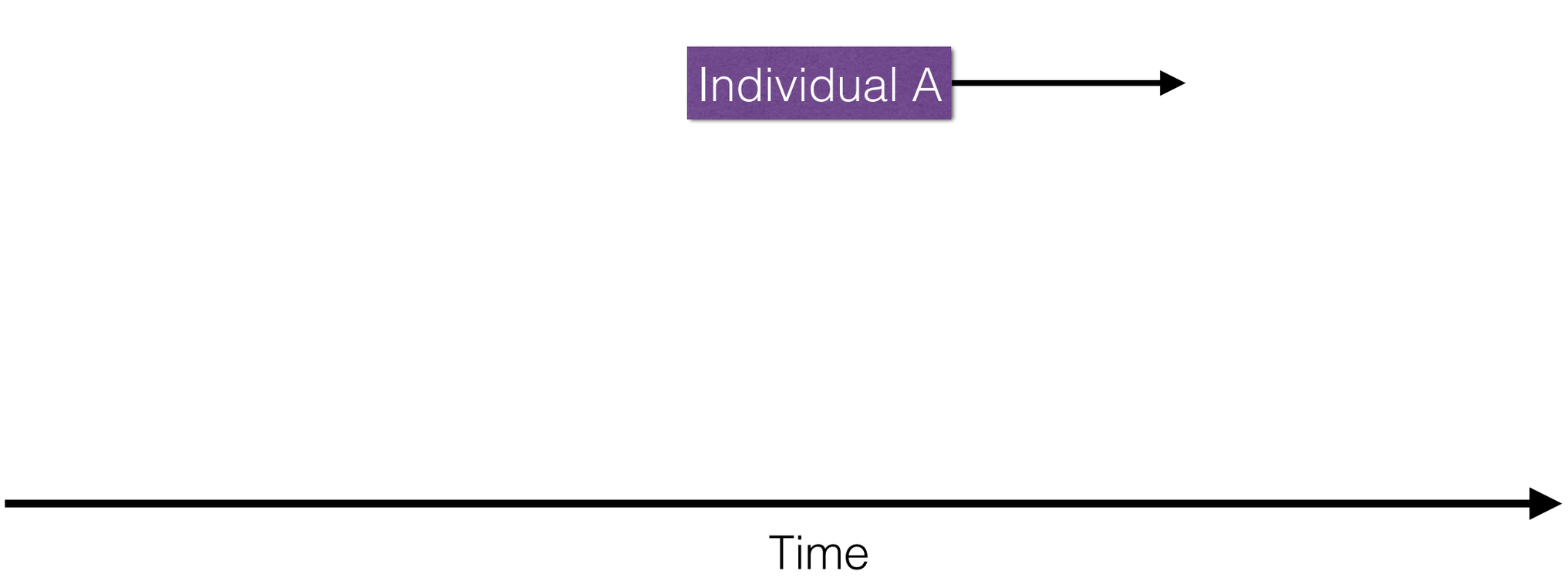
la mesure

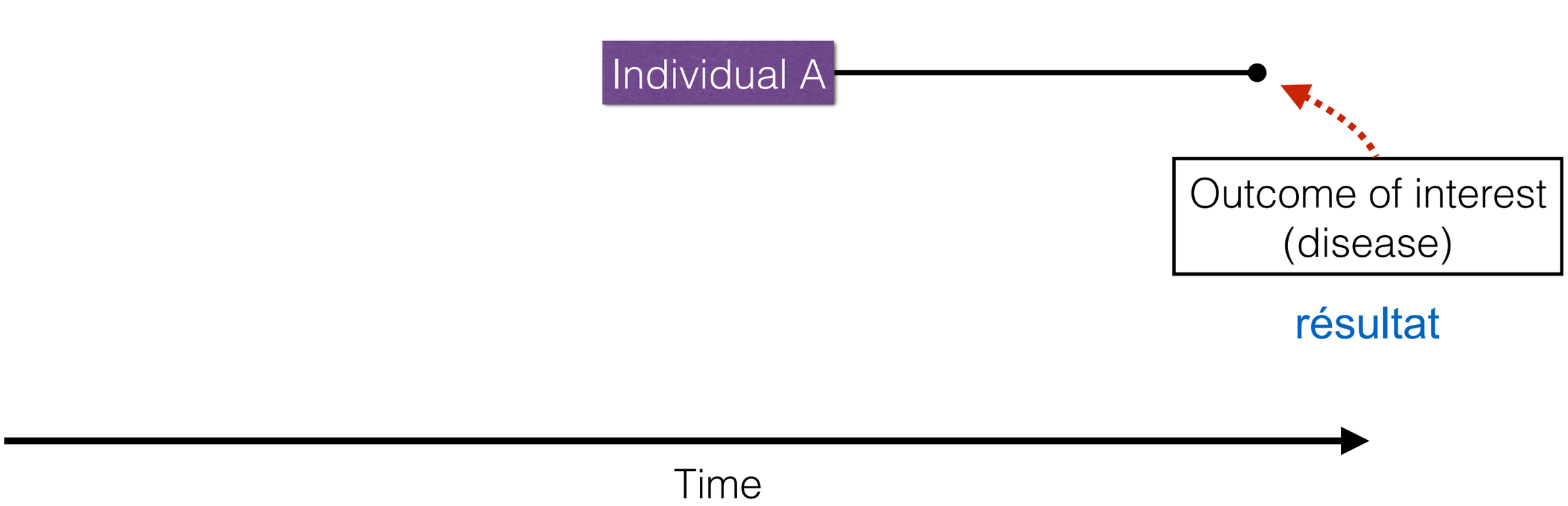


populations, plan d'étude et suivi

inférence et comparaison des populations

Individual A


Individual A
Time


Individual A
Outcome of interest (disease)
résultat
Time

Epidemiological Questions

What exposure causes the outcome?

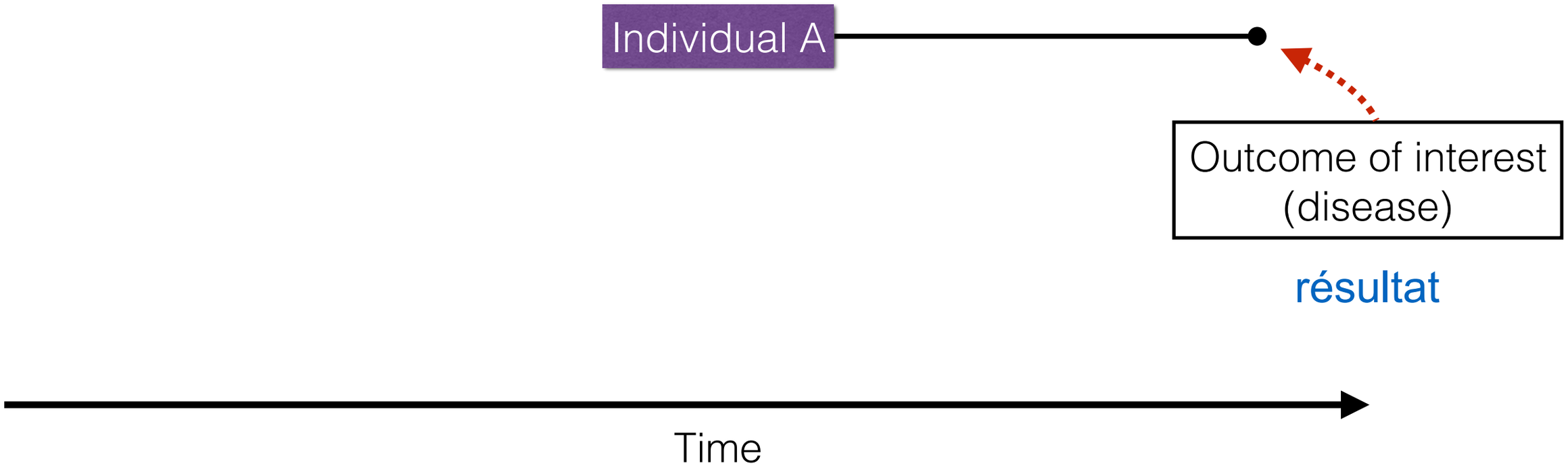
Quelle exposition provoque le résultat?

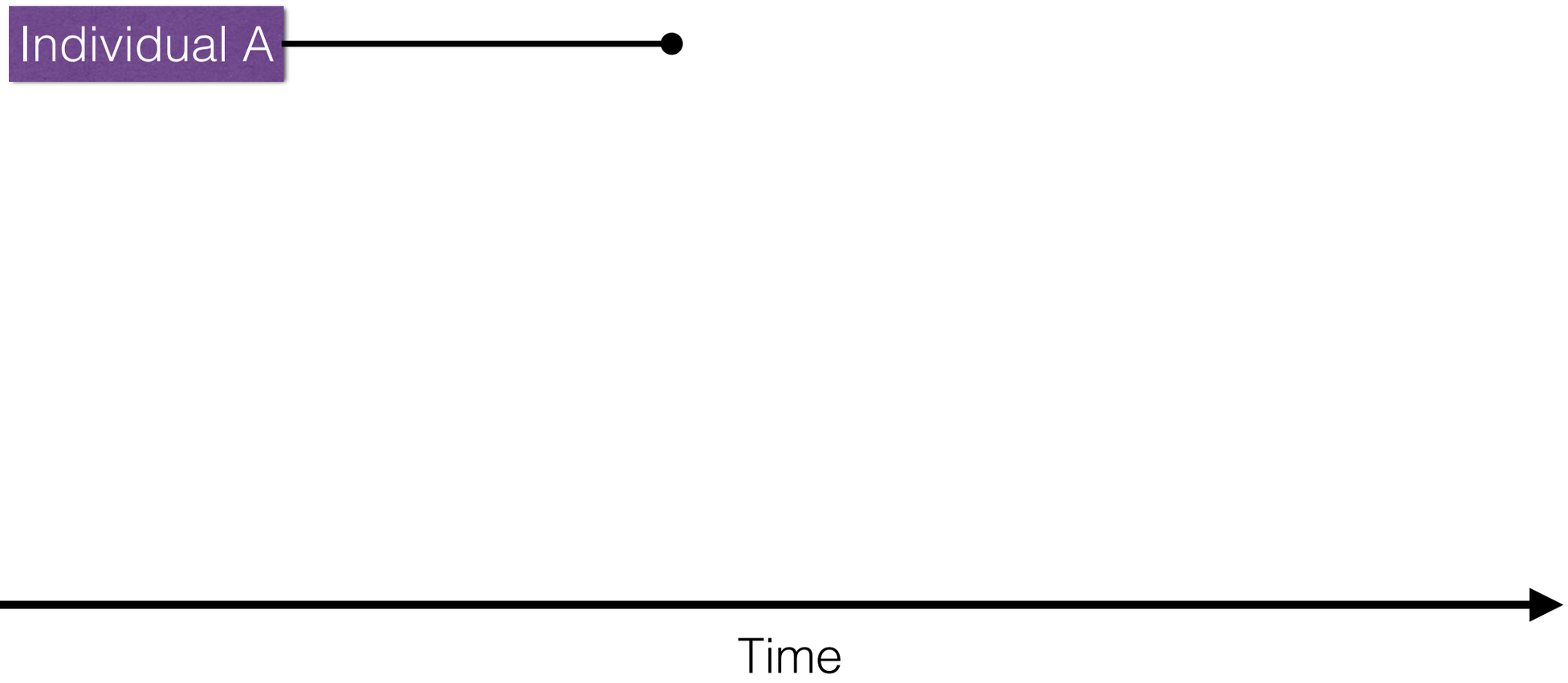
What are risk factors?

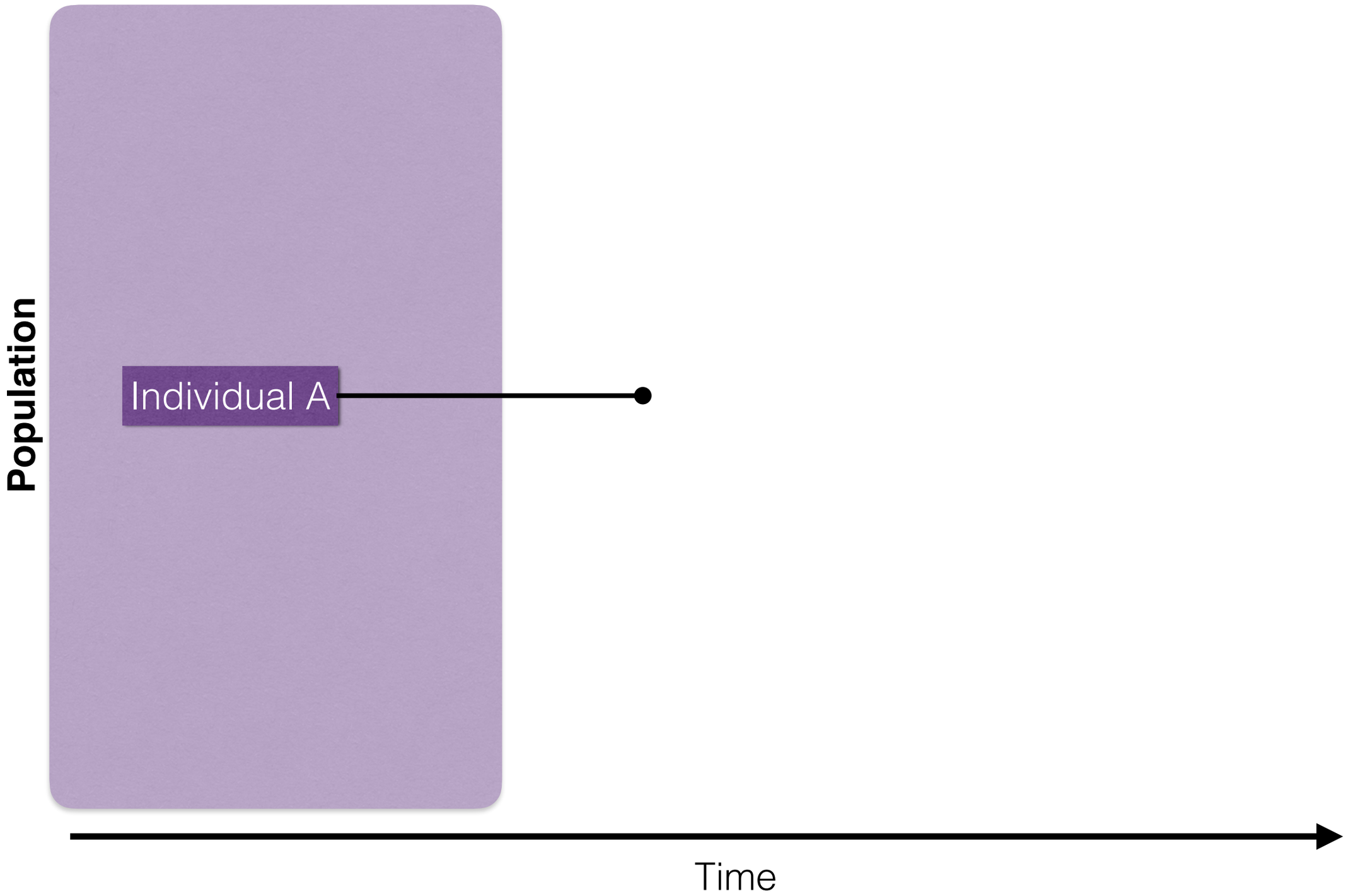
Quels sont les facteurs de risque?

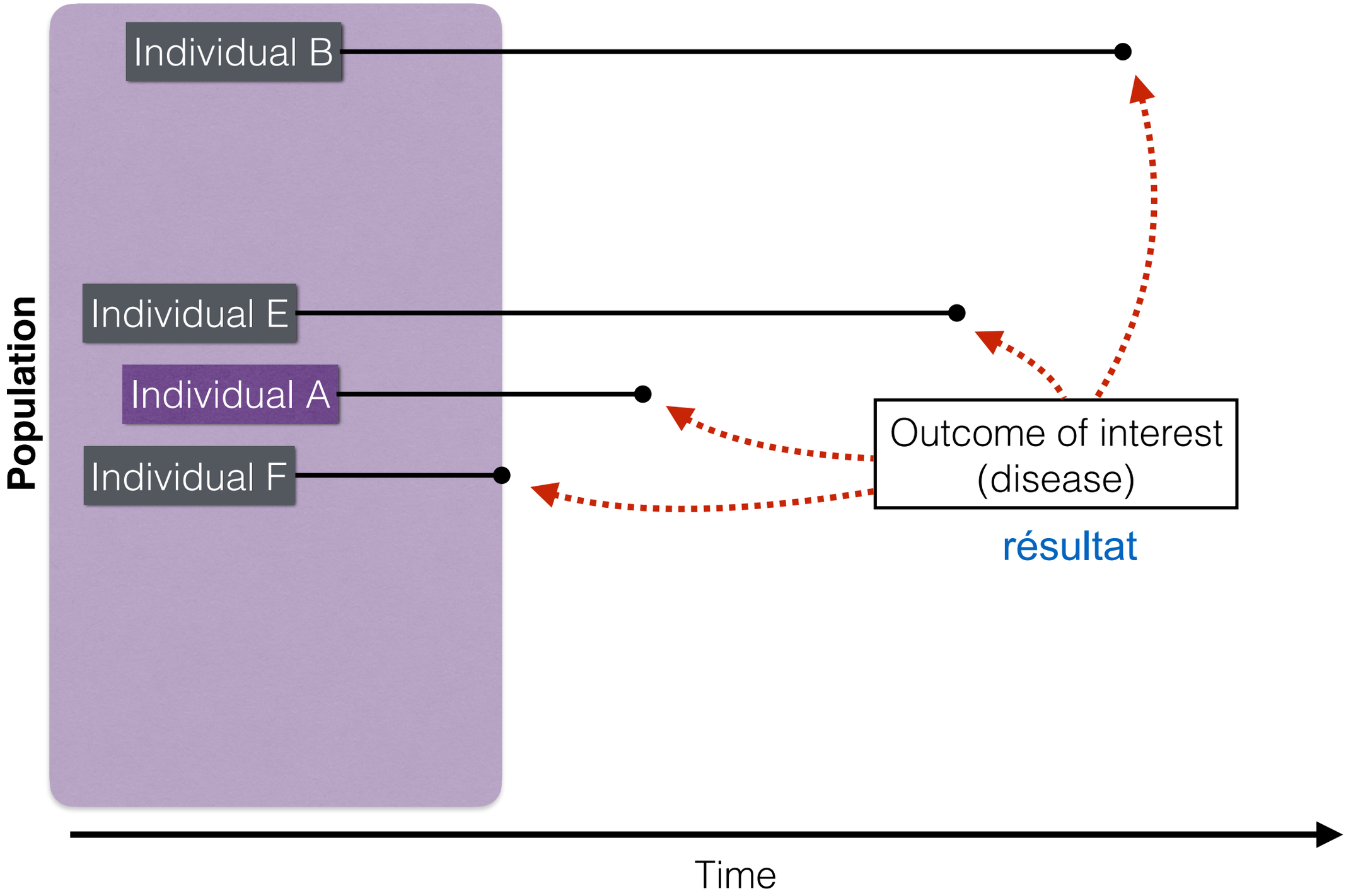
What treatment should be given?

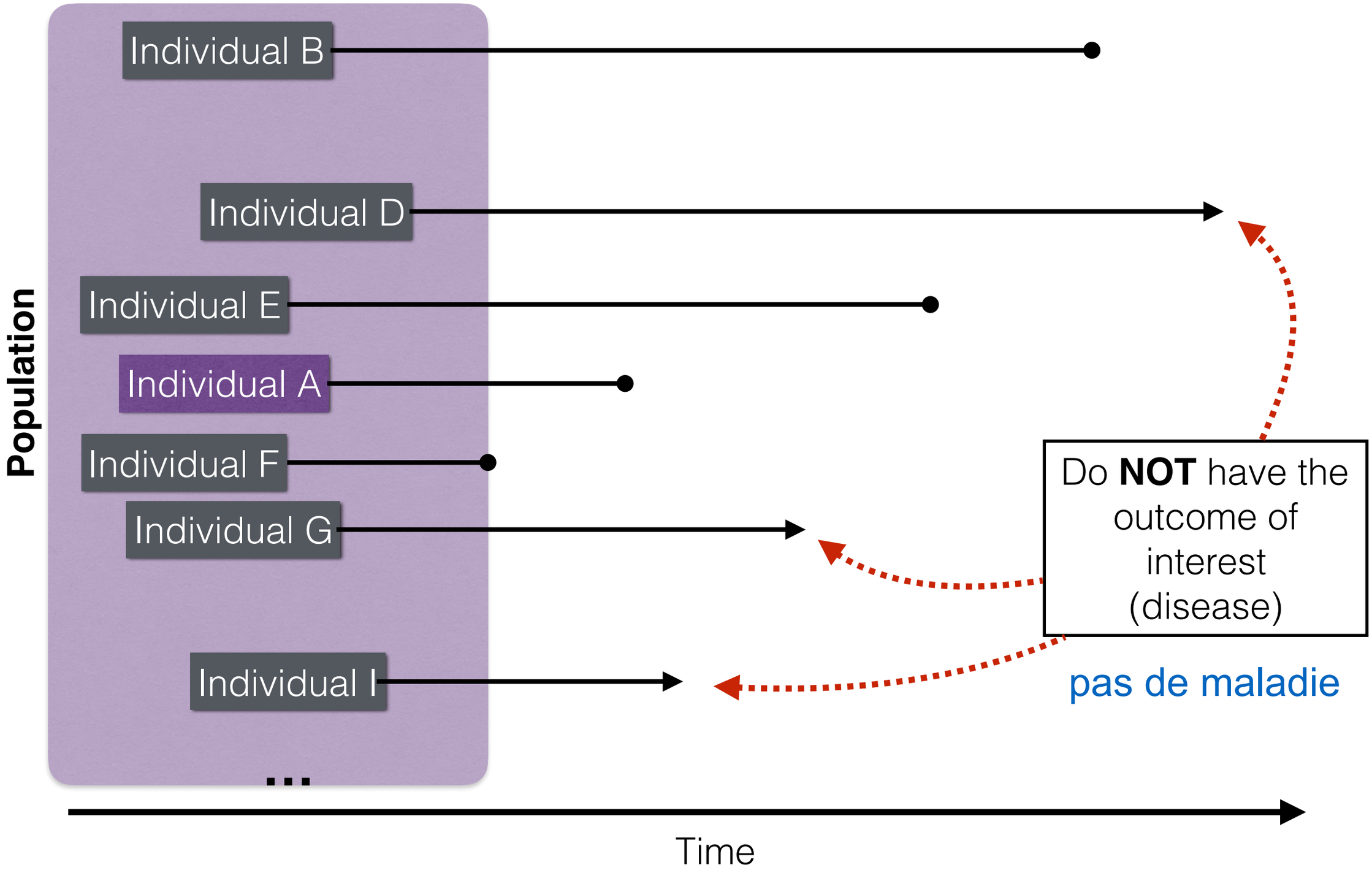
Quel traitement devrait être donné?

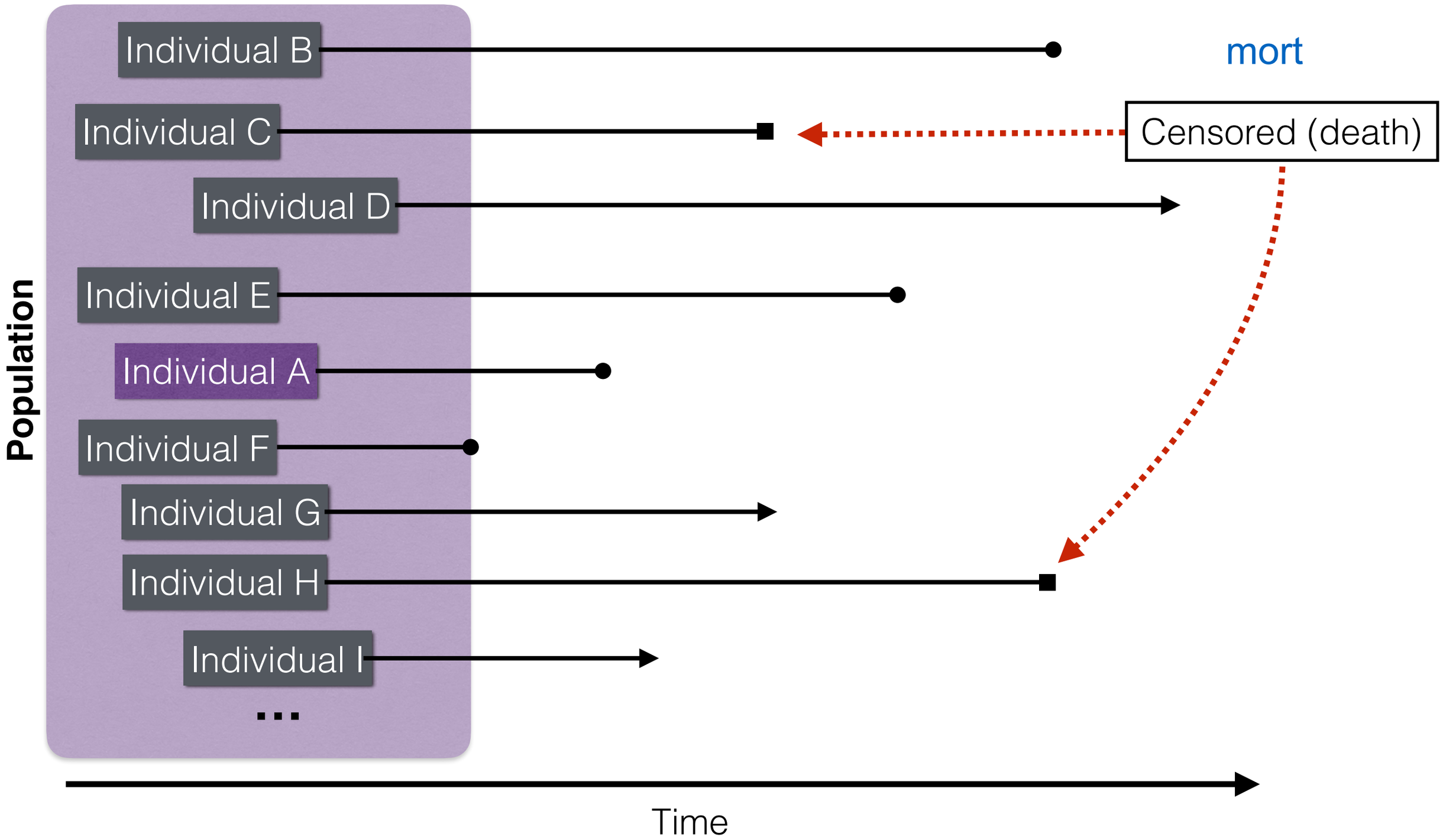


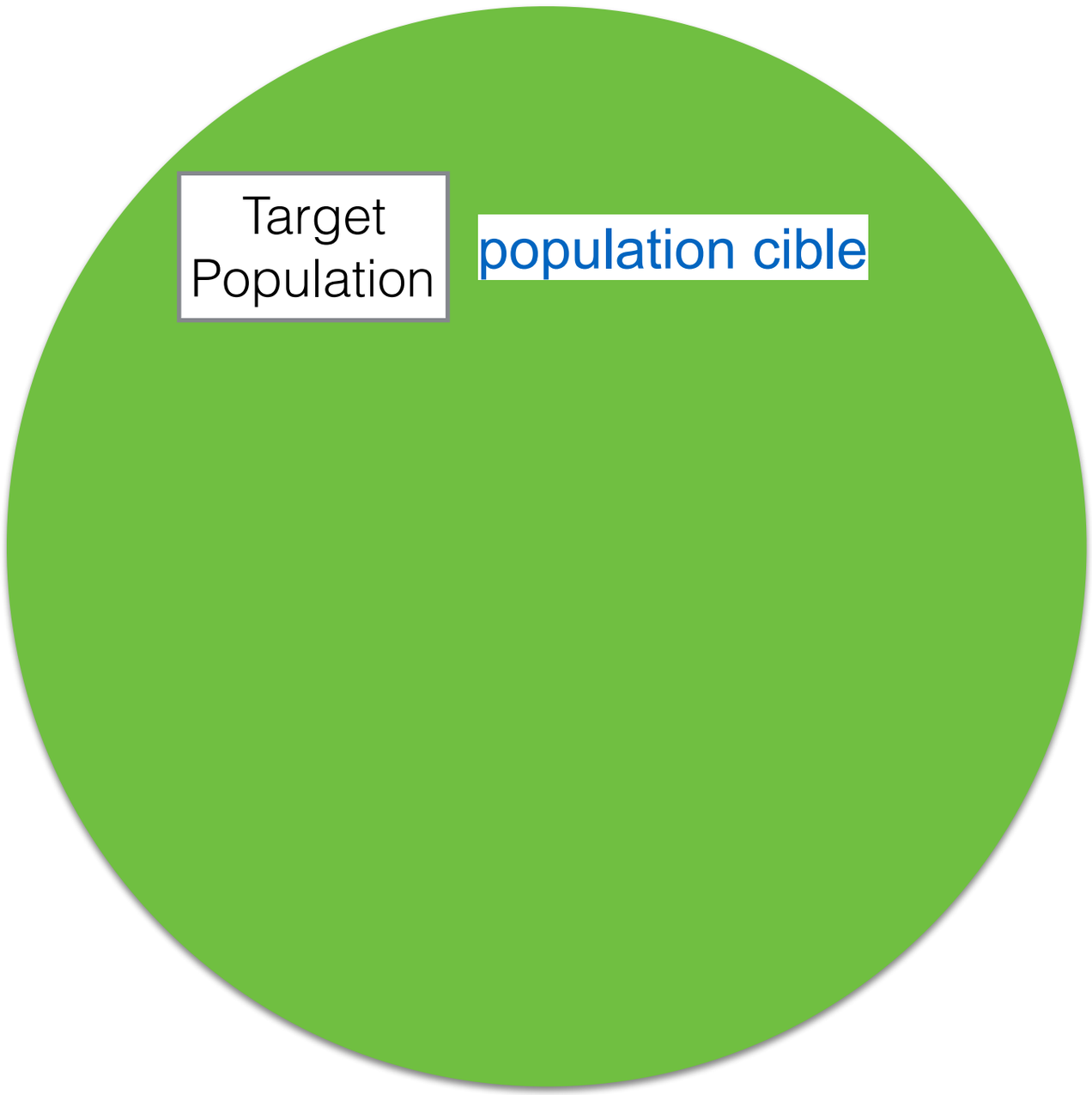

Individual A
Time


Population
Individual A
Time

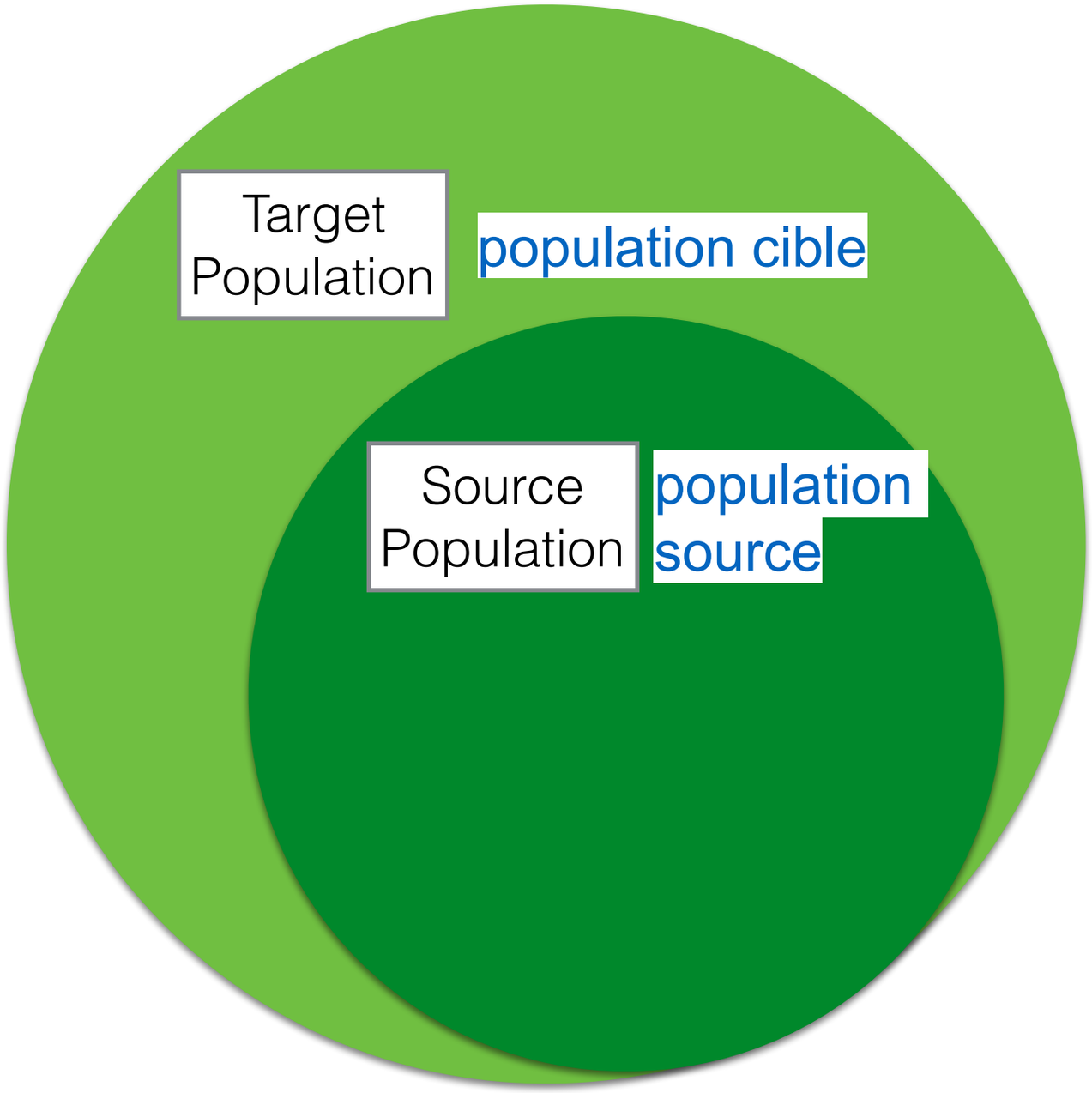

Population
Individual B
Individual E
Individual A
Individual F
Outcome of interest
(disease)
résultat
Time


Population
Individual B
Individual D
Individual E
Individual A
Individual F
Individual G
Individual I
...
Do **NOT** have the outcome of interest (disease)
pas de maladie
Time


Population
Individual B
Individual C
Individual D
Individual E
Individual A
Individual F
Individual G
Individual H
Individual I
...
mort
Censored (death)
Time

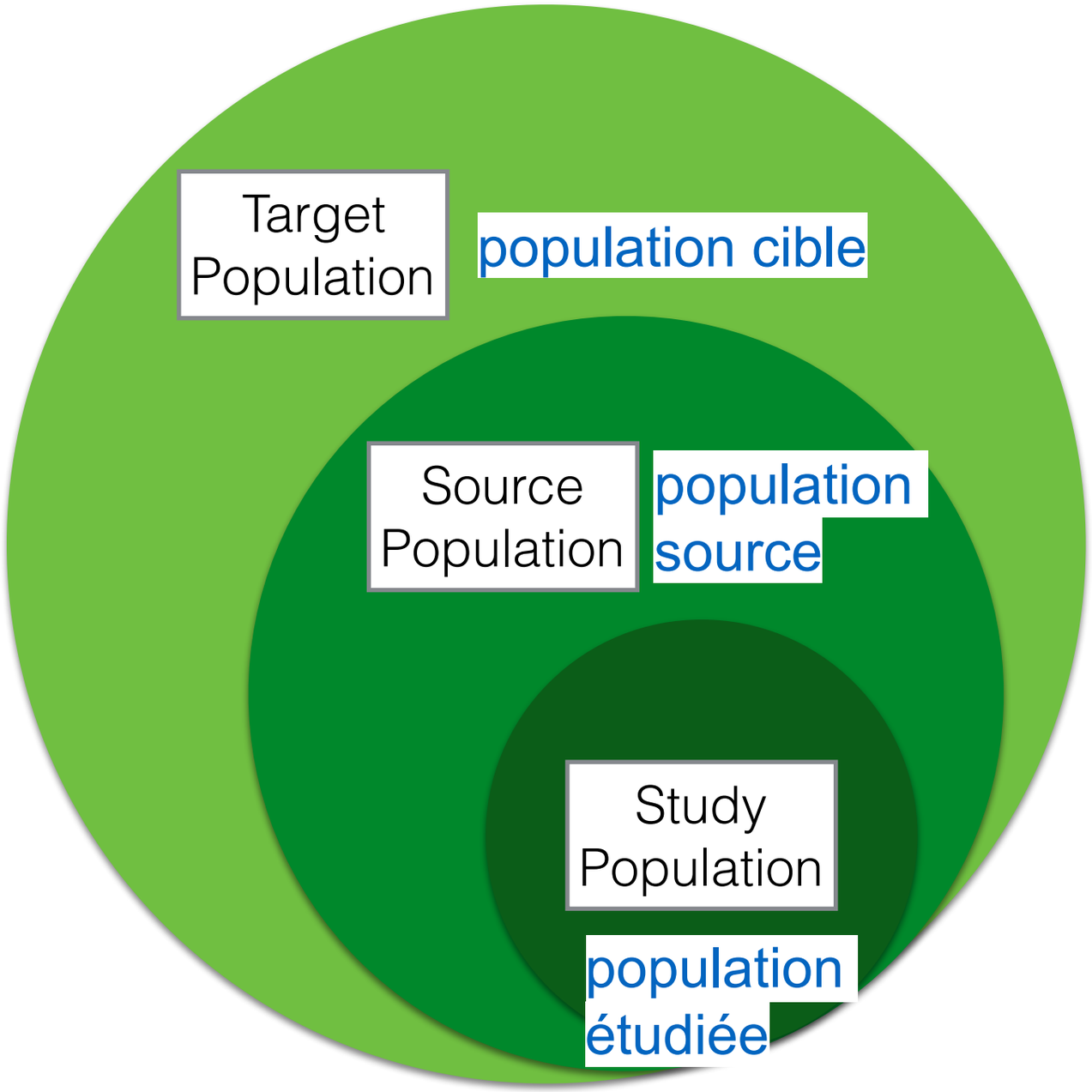


Target Population:
Want to make inferences about



Target Population:
Want to make inferences about

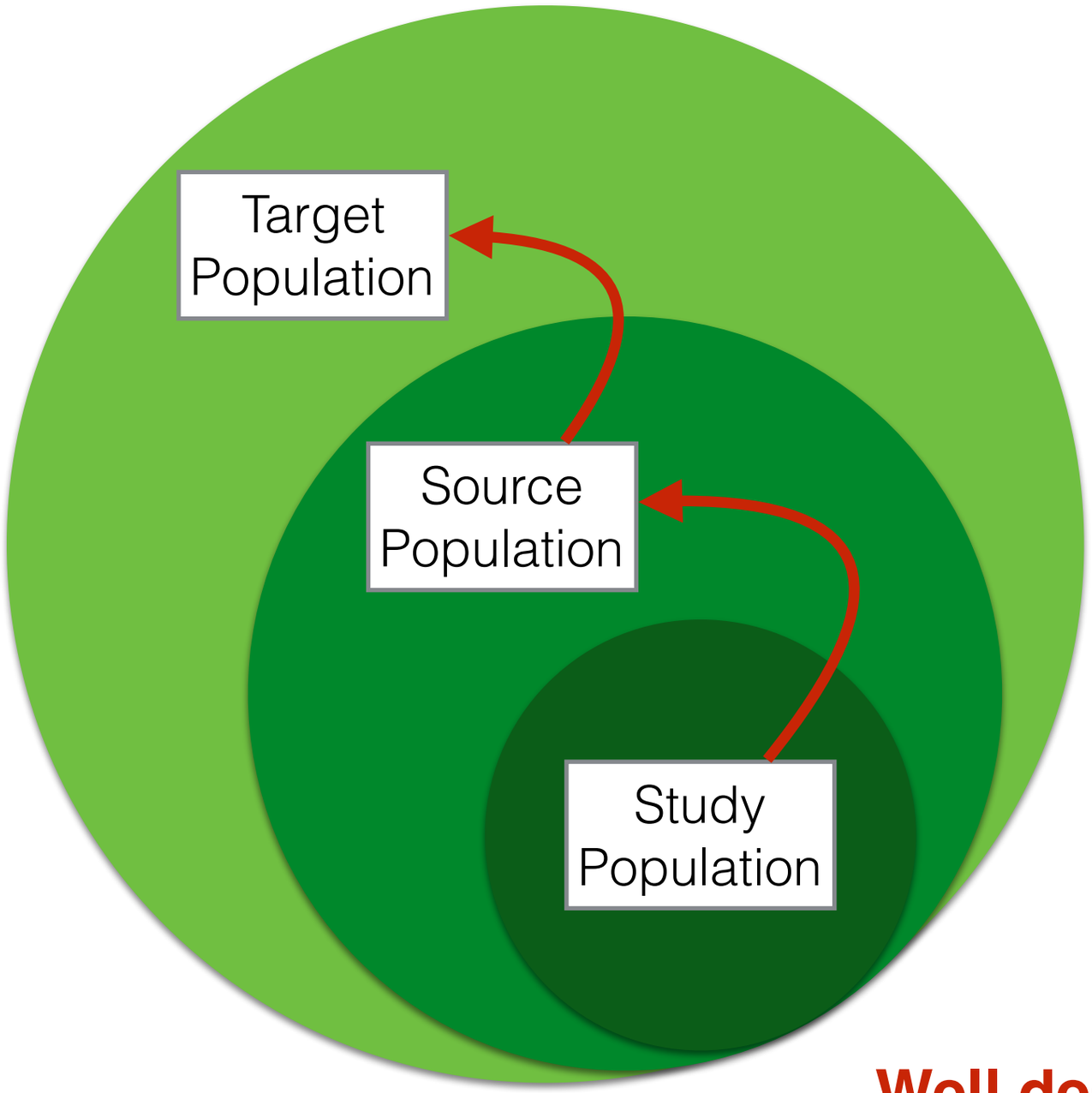
Source Population:
Choosing study population



Target Population:
Want to make inferences about

Source Population:
Choosing study population

Study Population:
Enrolled in study



Target Population:
Want to make inferences about

Source Population:
Choosing study population

Study Population:
Enrolled in study

Well designed studies allow us to make inference about the target population

Overview of different study designs

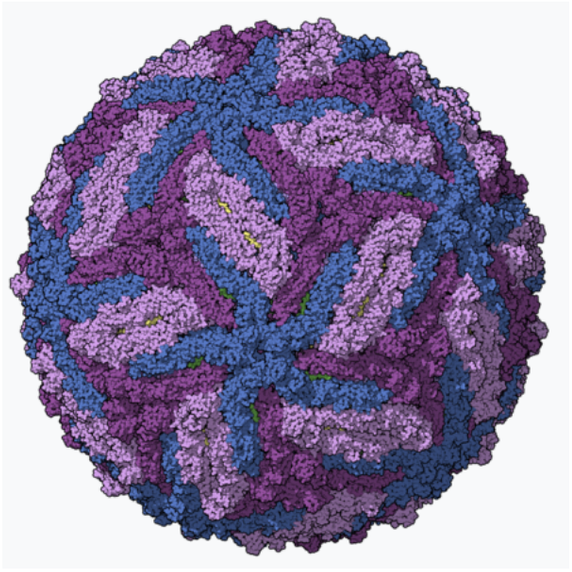
- Observational
  - Case-Control
  - Cross-Sectional
  - Cohort
    - Retrospective
    - Prospective
  - Ecological
- Experimental
  - Randomized Control Trial (RCT)

The Basics

Cross-Sectional ..... Case-Control ..... Cohort ..... RCT

Zika Virus: The Basics

Cross-Sectional ..... Case-Control ..... Cohort ..... RCT



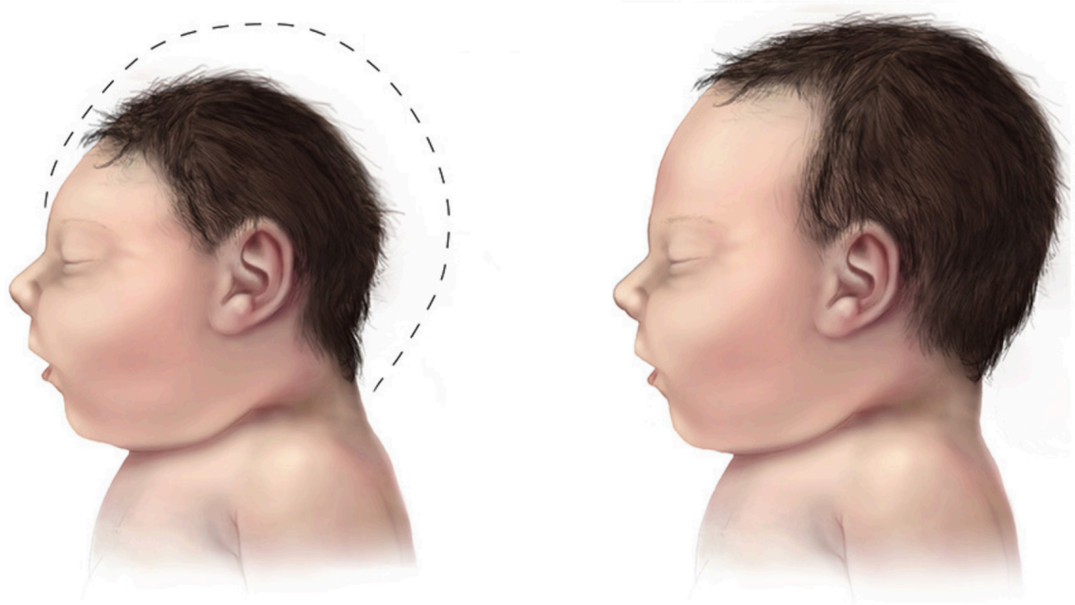
Virus in the family *Flaviviridae* (related to dengue, yellow fever)

Spread by Aedes mosquitoes

Multiple routes of transmission: vector, sexual, vertical

Zika Virus: The Basics

Cross-Sectional ..... Case-Control ..... Cohort ..... RCT

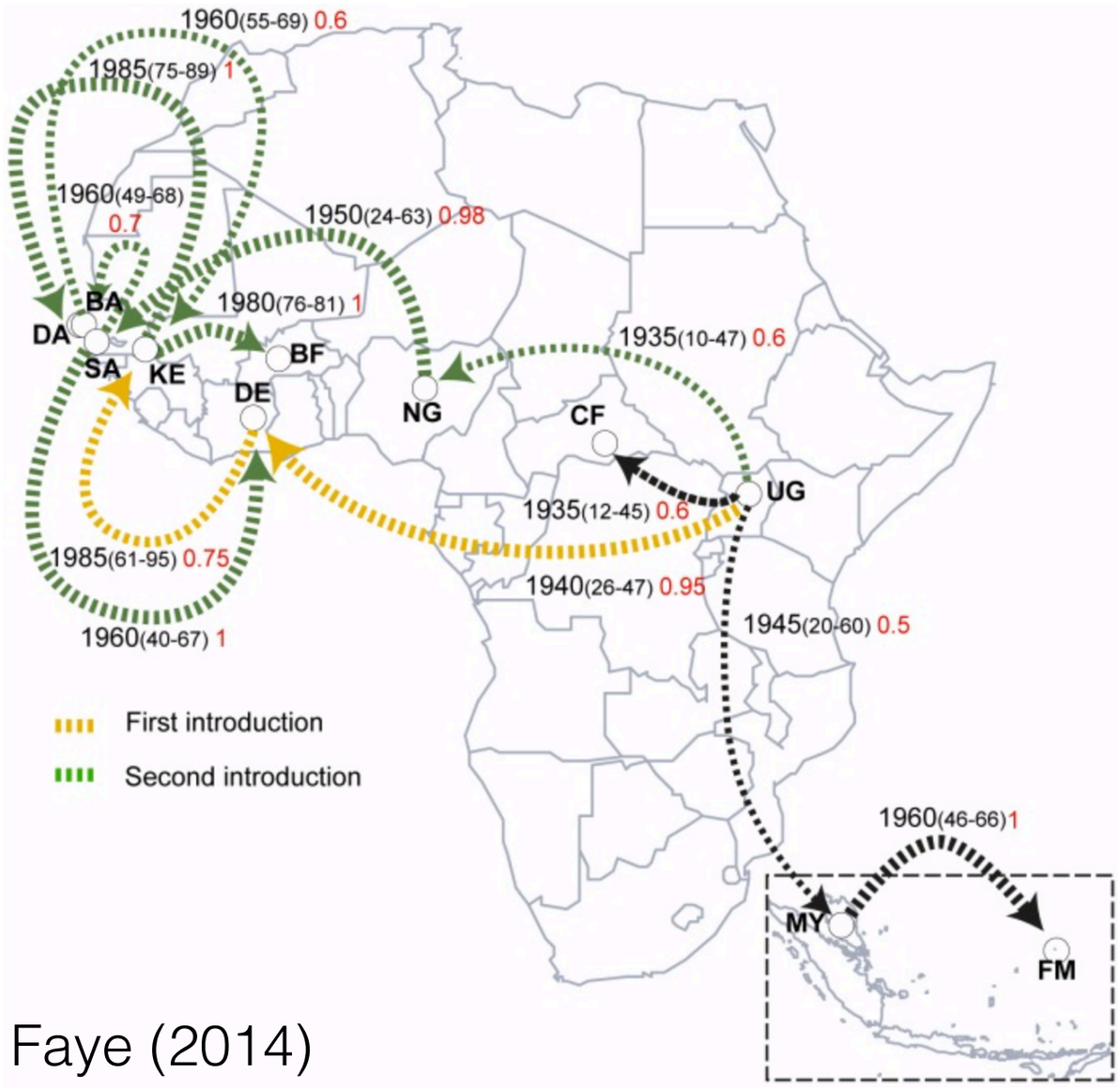


Often causes no or mild symptoms
But can spread from pregnant women to their fetuses and result in microcephaly, severe brain malformations, other birth defects

Currently cannot be prevented by medications or vaccines (current vaccine trials)

Zika Virus: The Basics

Cross-Sectional ..... Case-Control ..... Cohort ..... RCT

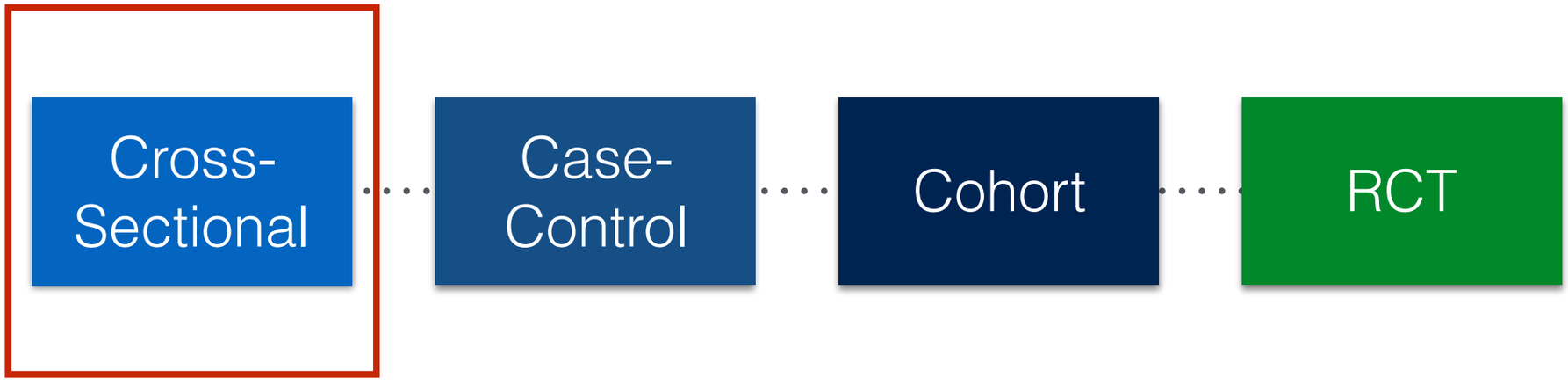


Faye (2014)

Originally isolated in 1947 (Ziika Forest, Uganda)

Sporadic outbreaks in Africa and Asia
Large outbreak in 2015-2016 (Americas, SE Asia, Pacific Islands, Brazil)

General: Study Design



Cross Sectional Study

- Examines the relationship between **diseases** (or other health outcomes) and **other variables of interest** as they exist in a **defined population at one particular point in time**.
  Examine la relation entre les maladies (ou d'autres résultats pour la santé) et d'autres variables d'intérêt, telles qu'elles existent dans une population définie à un moment donné.

- Presence or absence of the disease are determined for each member of the population.

- Commonly used to estimate disease prevalence rather than incidence.

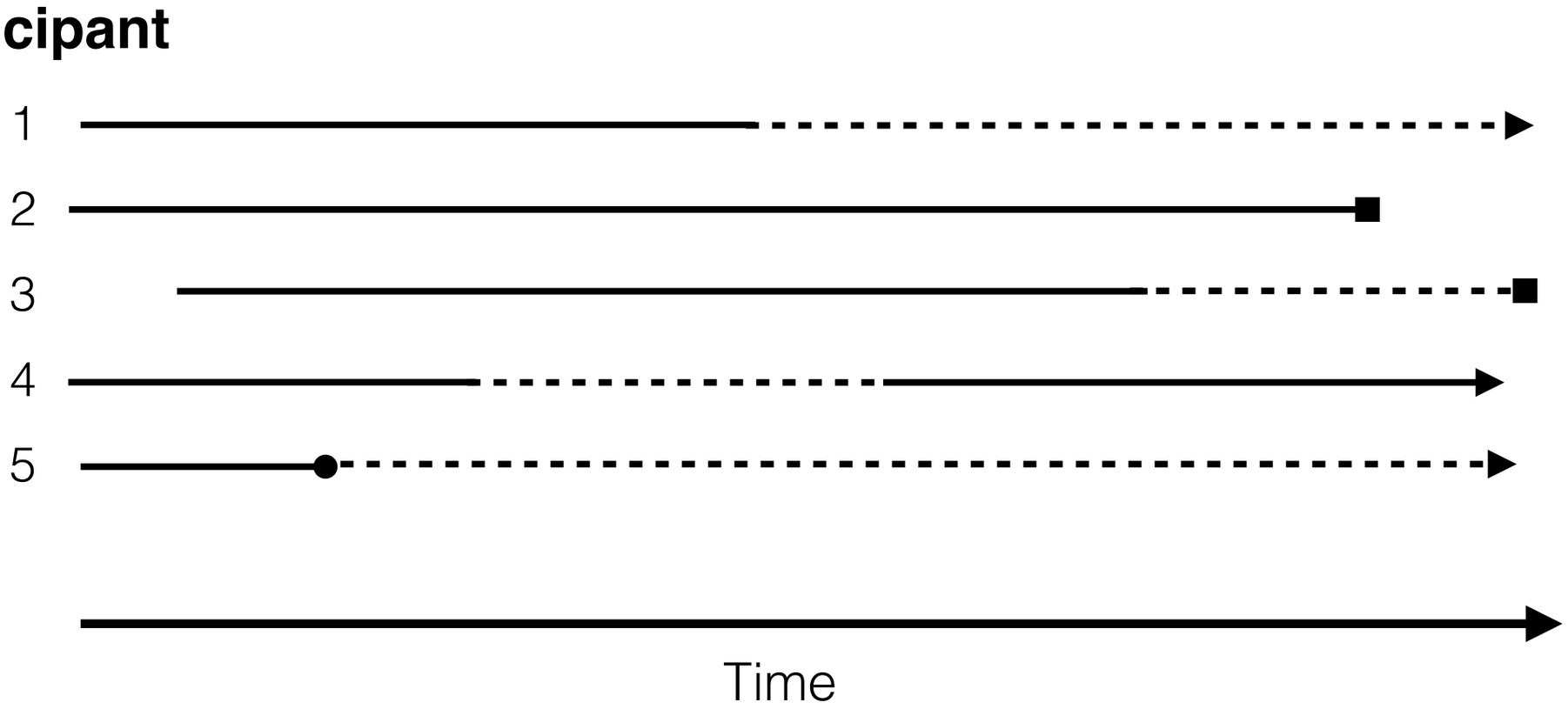
- **Snapshot of the population at a particular time.**
  Instantané de la population à un moment donné.

- **Exposure and outcome are assessed simultaneously.**
  L'exposition et les résultats sont évalués simultanément.

General: Study Design

Cross-Sectional — Case-Control — Cohort — RCT

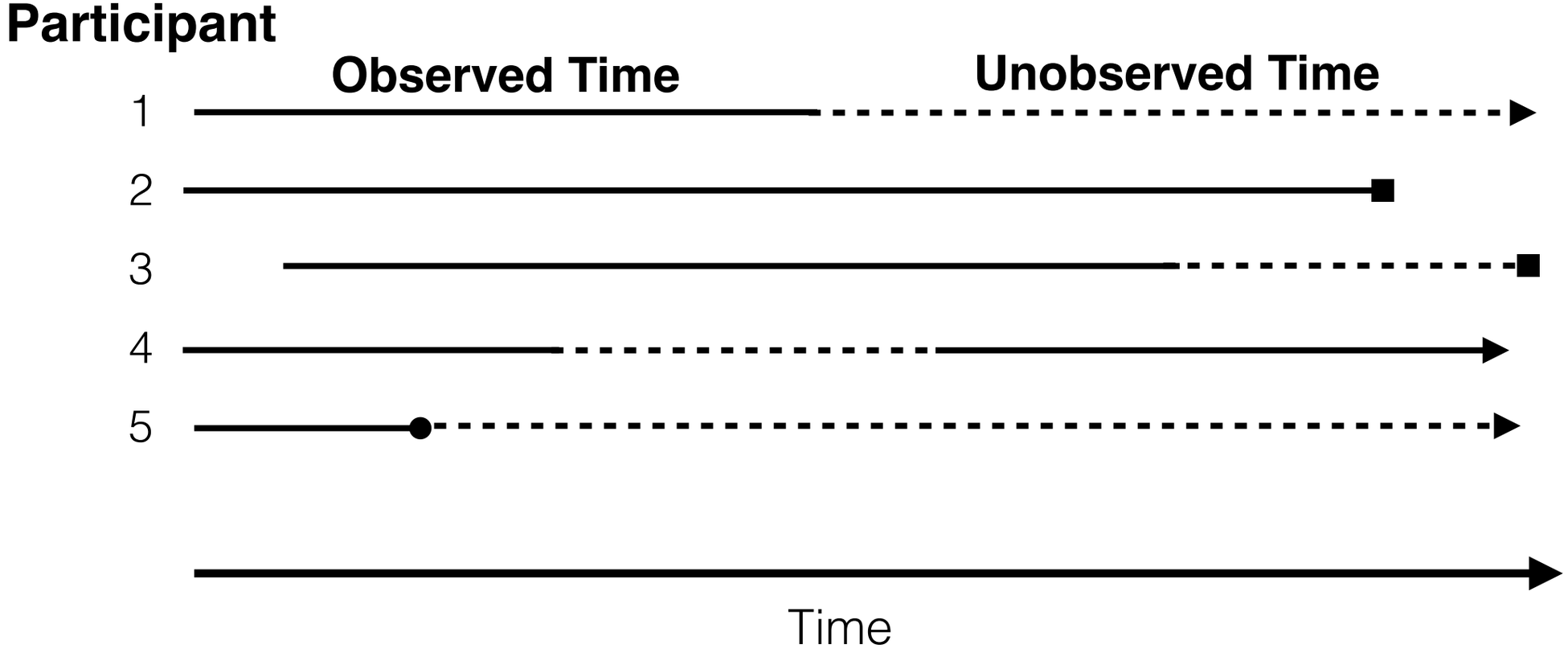
Cross Sectional Study



General: Study Design

Cross-Sectional — Case-Control — Cohort — RCT

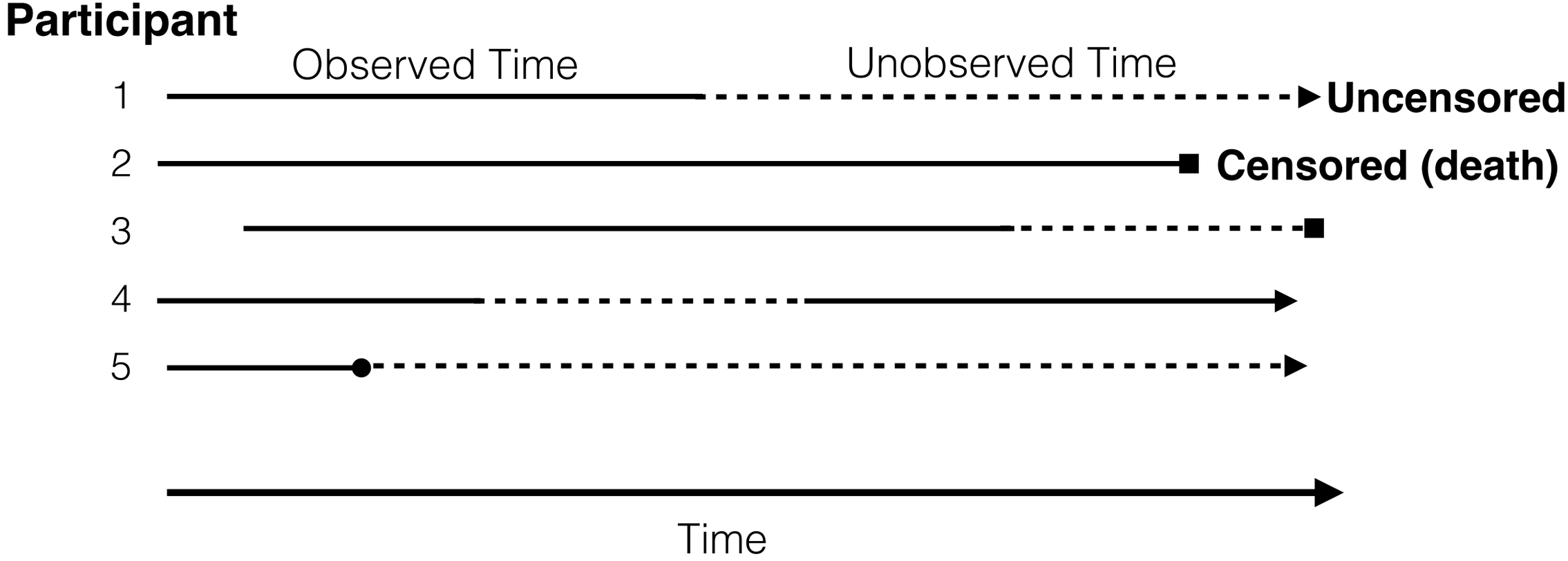
Cross Sectional Study



General: Study Design

Cross-Sectional — Case-Control — Cohort — RCT

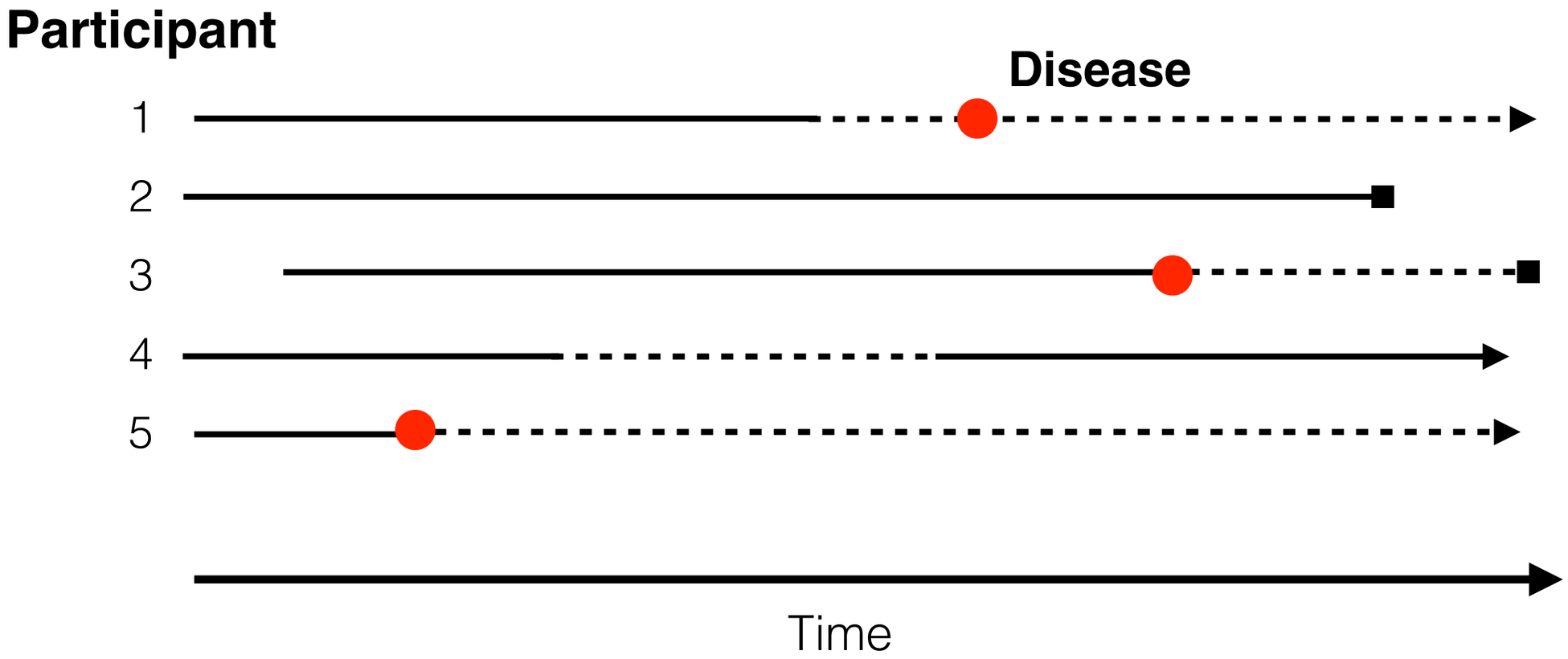
Cross Sectional Study



General: Study Design

Cross-Sectional — Case-Control — Cohort — RCT

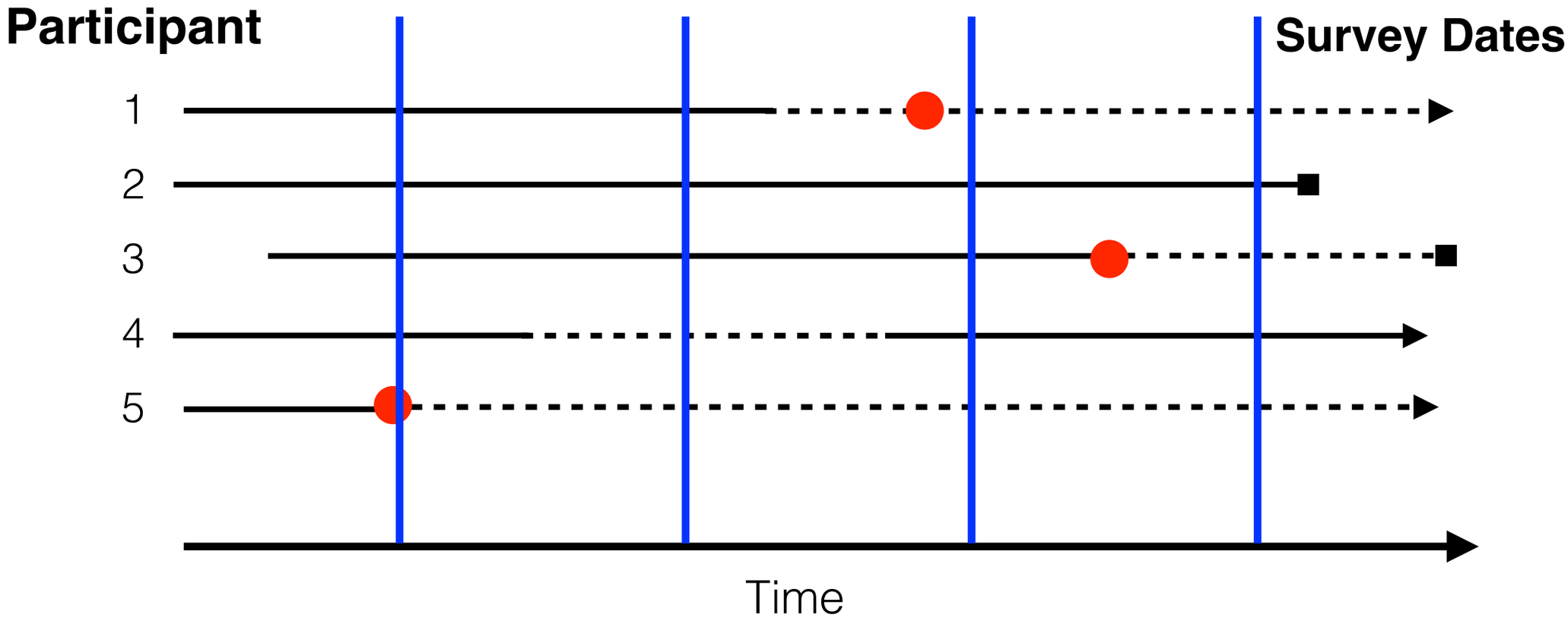
Cross Sectional Study



General: Study Design

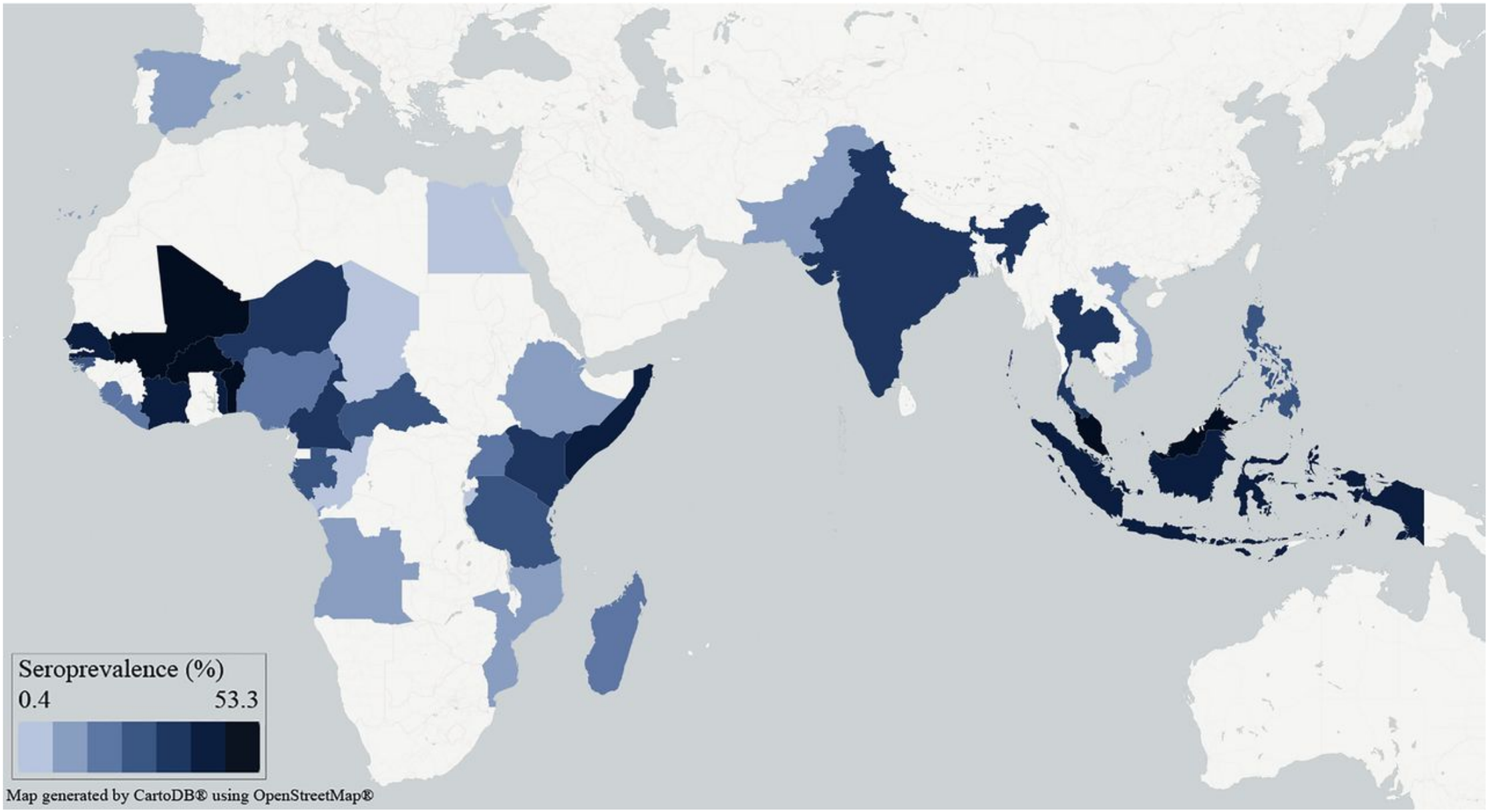
Cross-Sectional — Case-Control — Cohort — RCT

Cross Sectional Study



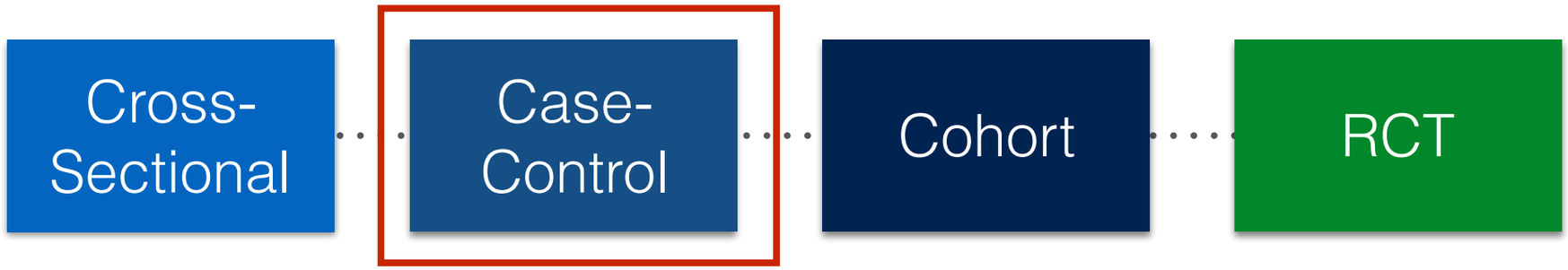
Zika: Study Design

Cross-Sectional ..... Case-Control ..... Cohort ..... RCT



Posen (2016)

General: Study Design



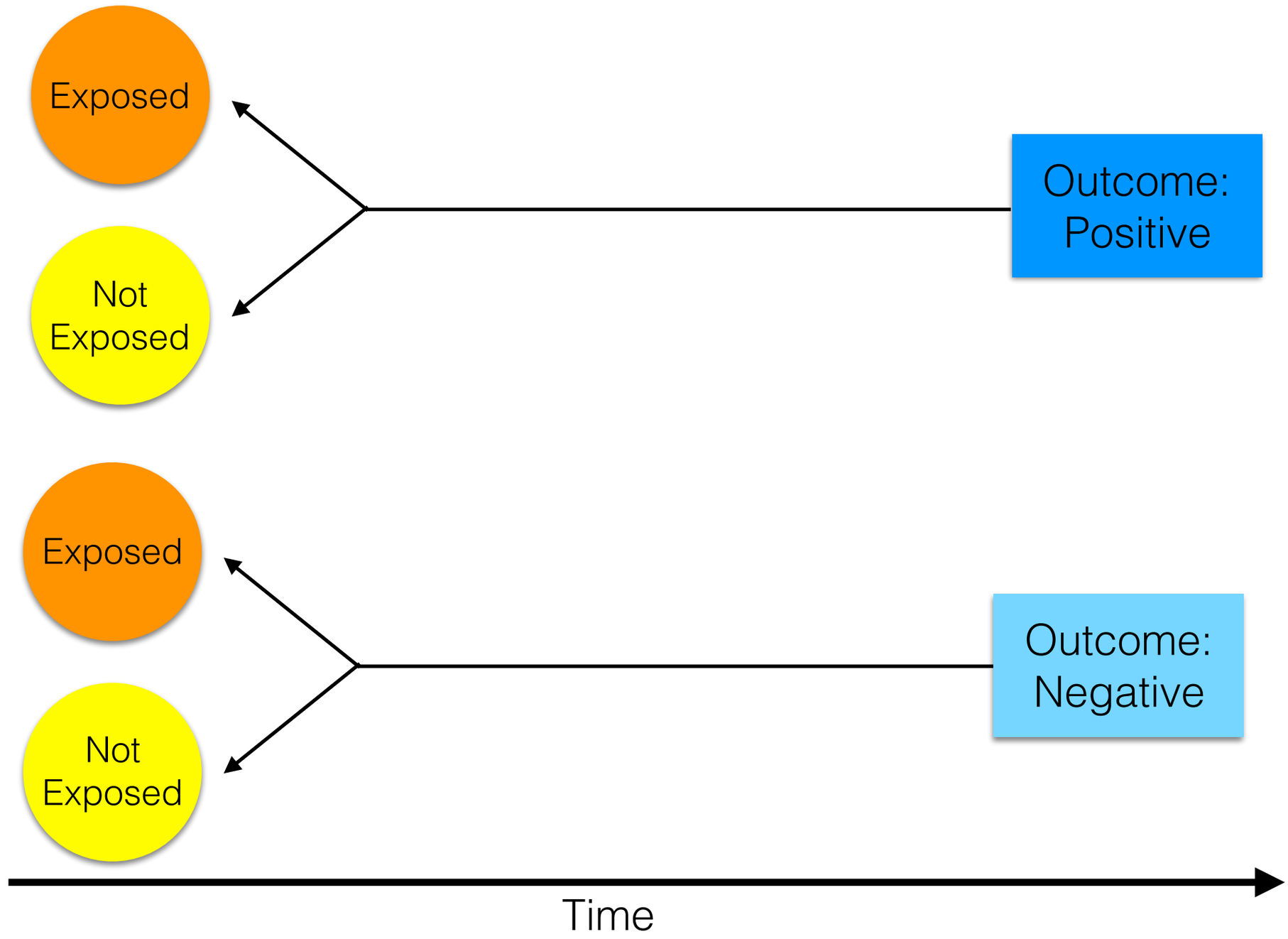
Case-Control Study

- The **observational** epidemiological study of **persons with a disease** (or another outcome variable) of interest and a suitable **control group of persons without the disease**
  Étude épidémiologique d'observation sur des personnes présentant une maladie (ou une autre variable de résultat) d'intérêt et sur un groupe de contrôle approprié de personnes non atteintes de la maladie

- Potential relationship of a suspected risk factor or an attribute to the disease is examined by **comparing the the disease and non-diseased subjects** with regard to how frequently the factor or attribute is present in each of these groups.
  La relation potentielle d'un facteur de risque présumé ou d'un attribut avec la maladie est examinée en comparant la maladie et les sujets non atteints à la fréquence à laquelle le facteur ou l'attribut est présent dans chacun de ces groupes.

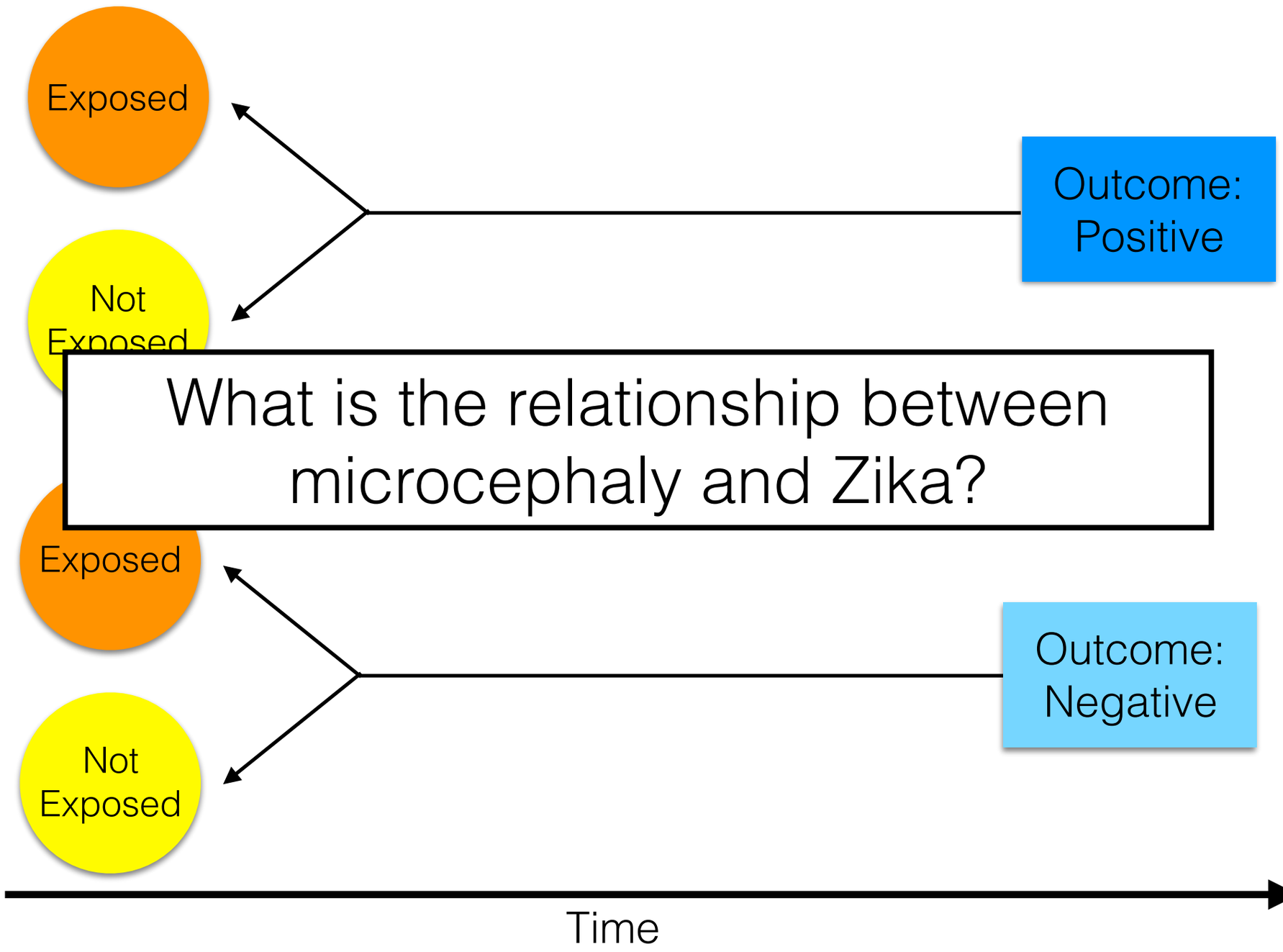
General: Study Design

Cross-Sectional — Case-Control — Cohort — RCT



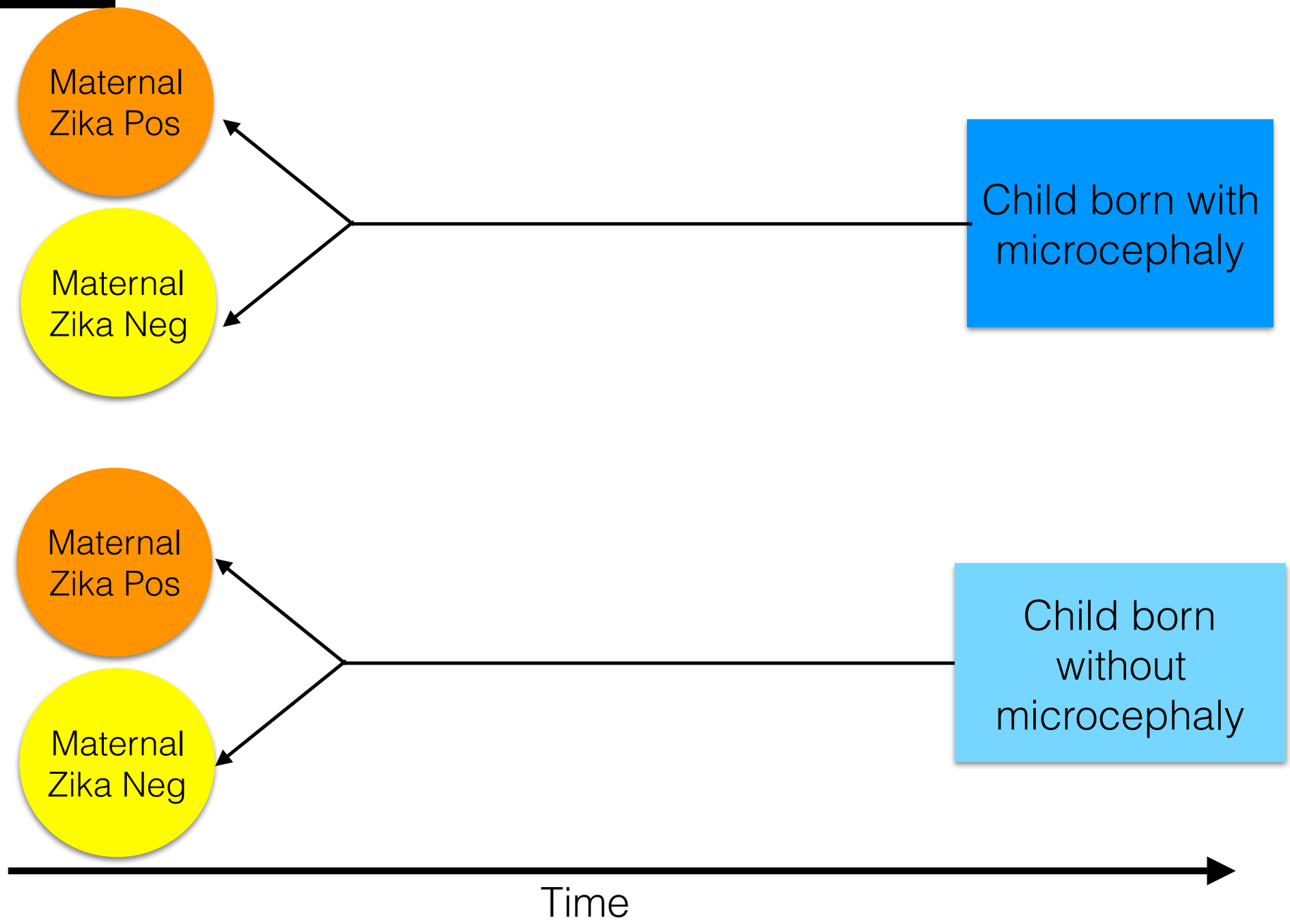
Zika: Study Design

Cross-Sectional · Case-Control · Cohort · RCT



Zika: Study Design

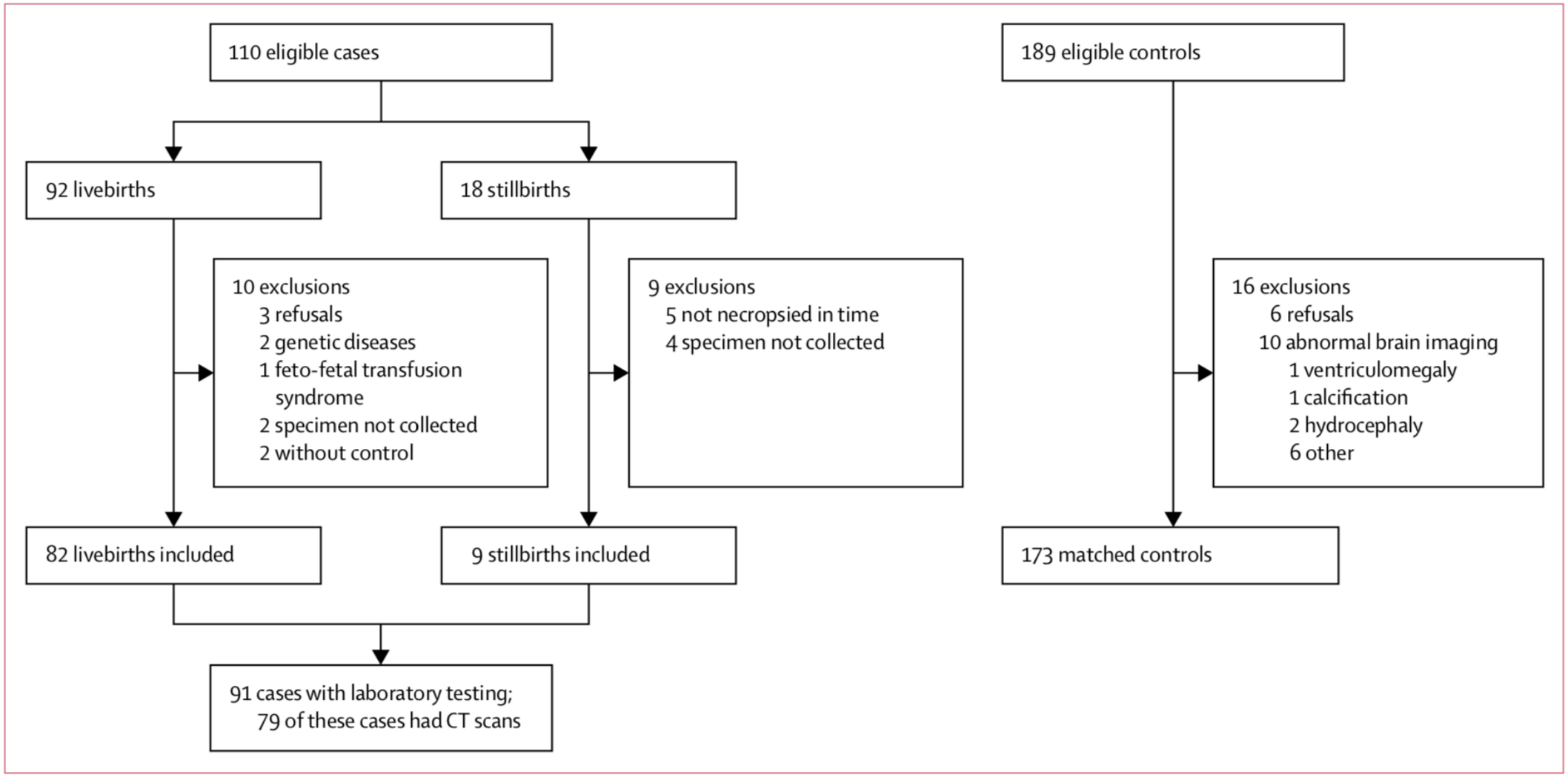
Cross-Sectional · Case-Control · Cohort · RCT



Zika: Study Design

Cross-Sectional . . . Case-Control . . . . Cohort . . . . RCT

Case-Control study - 8 Brazilian hospitals



de Araujo (2018)

Zika: Study Design

Cross-Sectional · · · Case-Control · · · · Cohort · · · · RCT

Case-Control study - 8 Brazilian hospitals

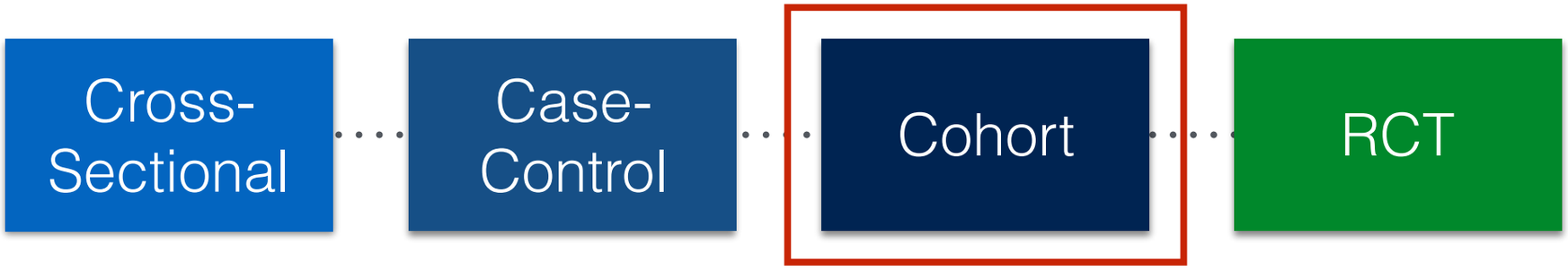
	Cases*	Controls*	Matched odds ratio (95% CI)
Serum, CSF samples, or macerated tissue			
Zika-positive, of total cases or controls	32/91 (35%)	0/173	87·0 (15·6–∞)
Zika-positive, of total cases or controls, adjusted†	..	..	73·1 (13·0–∞)
Cases, categorised by severity of microcephaly‡			
Severe	19/26 (73%)	0/51	52·4 (9·1–∞)
Not severe	13/65 (20%)	0/122	33·7 (5·6–∞)

*Data are the number of all cases or controls who were positive for Zika virus, assessed by qRT-PCR or Zika virus-specific IgM/total number of patients (%). †Odds ratio when adjusted by smoking during pregnancy, maternal vaccination against tetanus, diphtheria, and acellular pertussis during pregnancy, and skin colour. ‡Severe is defined as a head circumference of more than 3 SD smaller than the mean for their sex and gestational age.[10,14] Not severe was defined as a head circumference of 2–3 SD smaller than the mean for their sex and gestational age. Matched odds ratios in this subgroup are crude because of small numbers.

Table 5: **Association between microcephaly and Zika virus infection**

de Araujo (2018)

General: Study Design

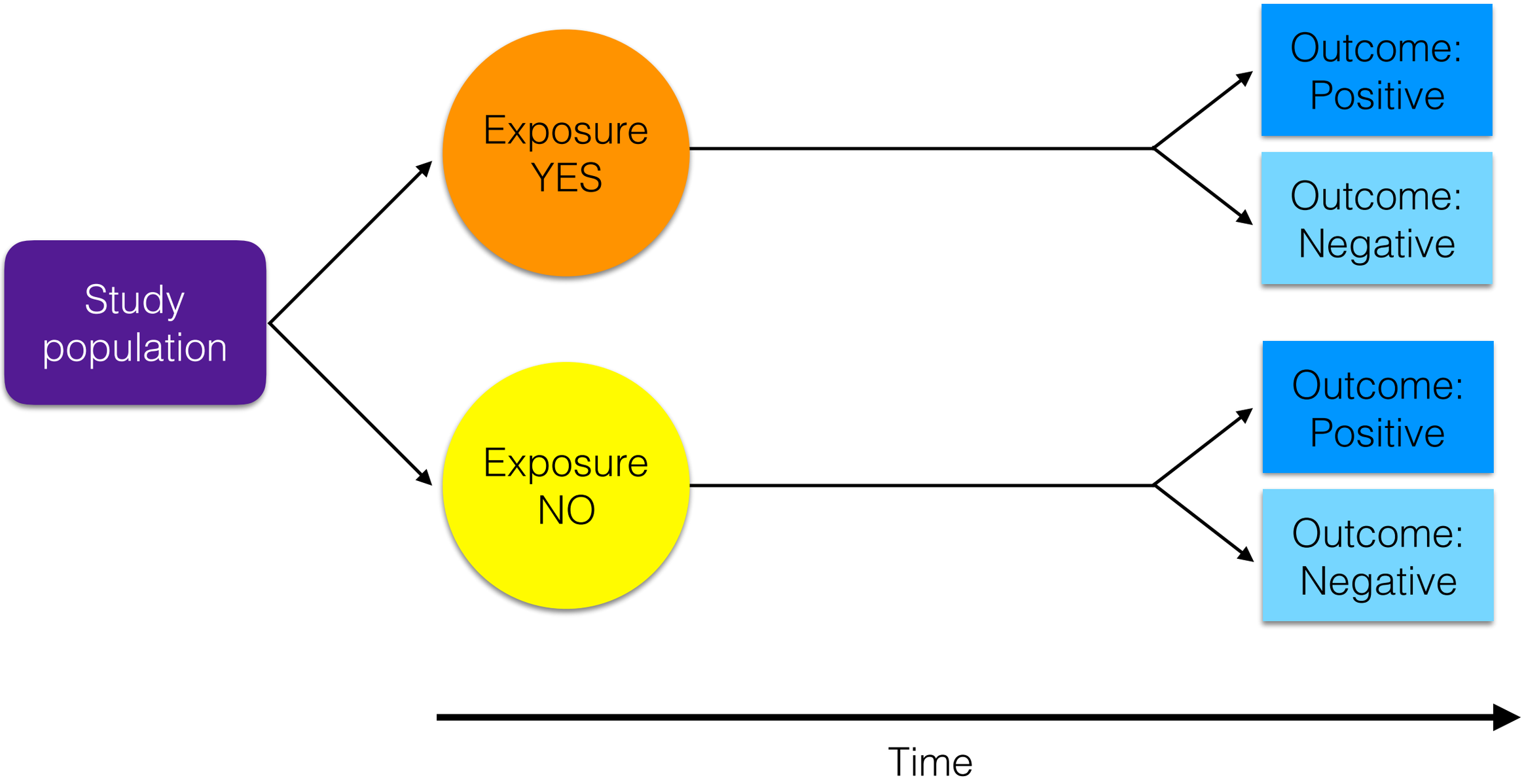


Cohort Study

- **Enrolled group of people** who have a common experience or grouping.
  Groupe de personnes inscrites ayant une expérience commune ou un groupe.
    - Age cohort, risk cohort
- General population sample
- Clinic based
- **Prospective or retrospective**
  Prospective ou rétrospective

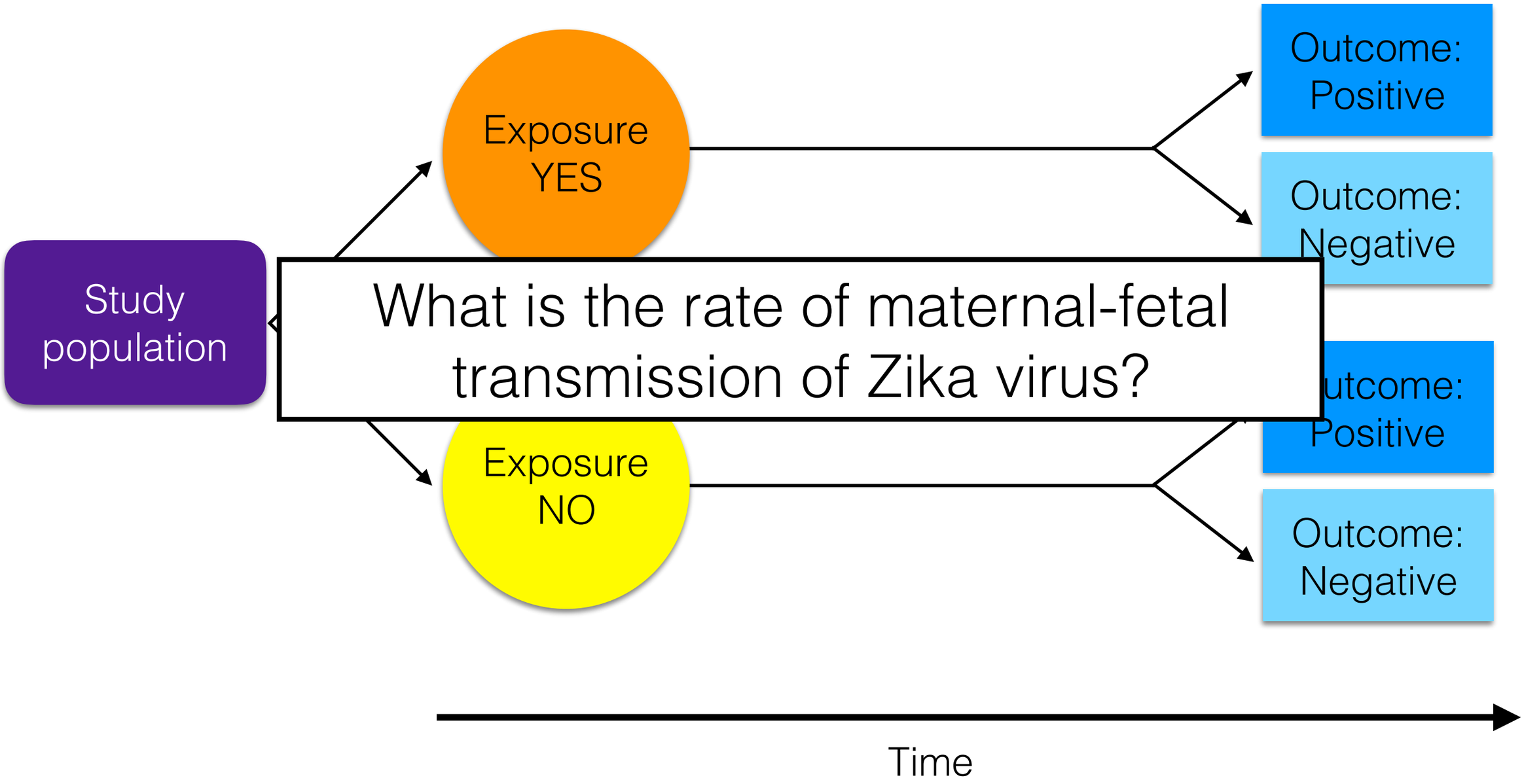
General: Study Design

Cross-Sectional — Case-Control — **Cohort** — RCT



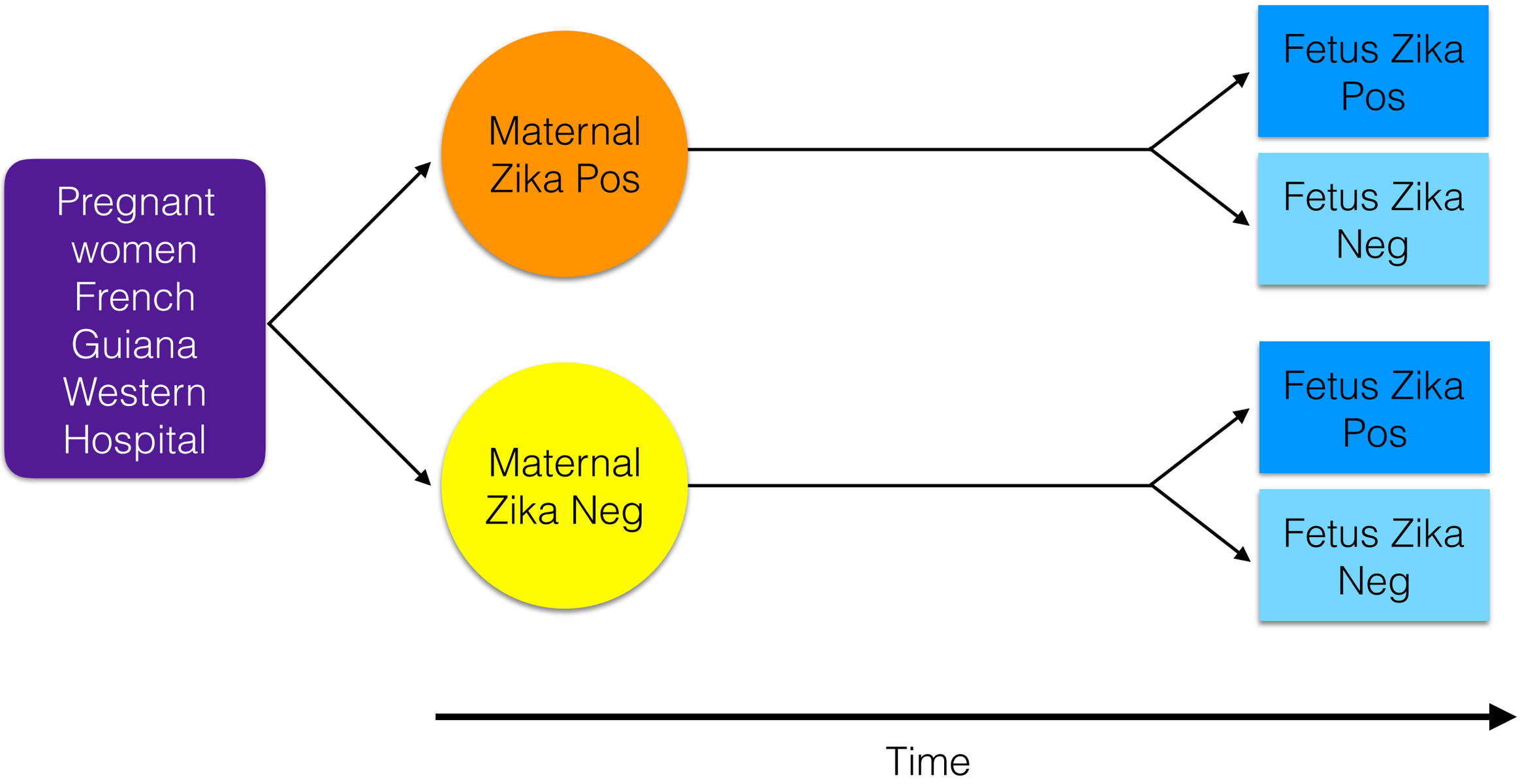
Zika: Study Design

Cross-Sectional | Case-Control | **Cohort** | RCT



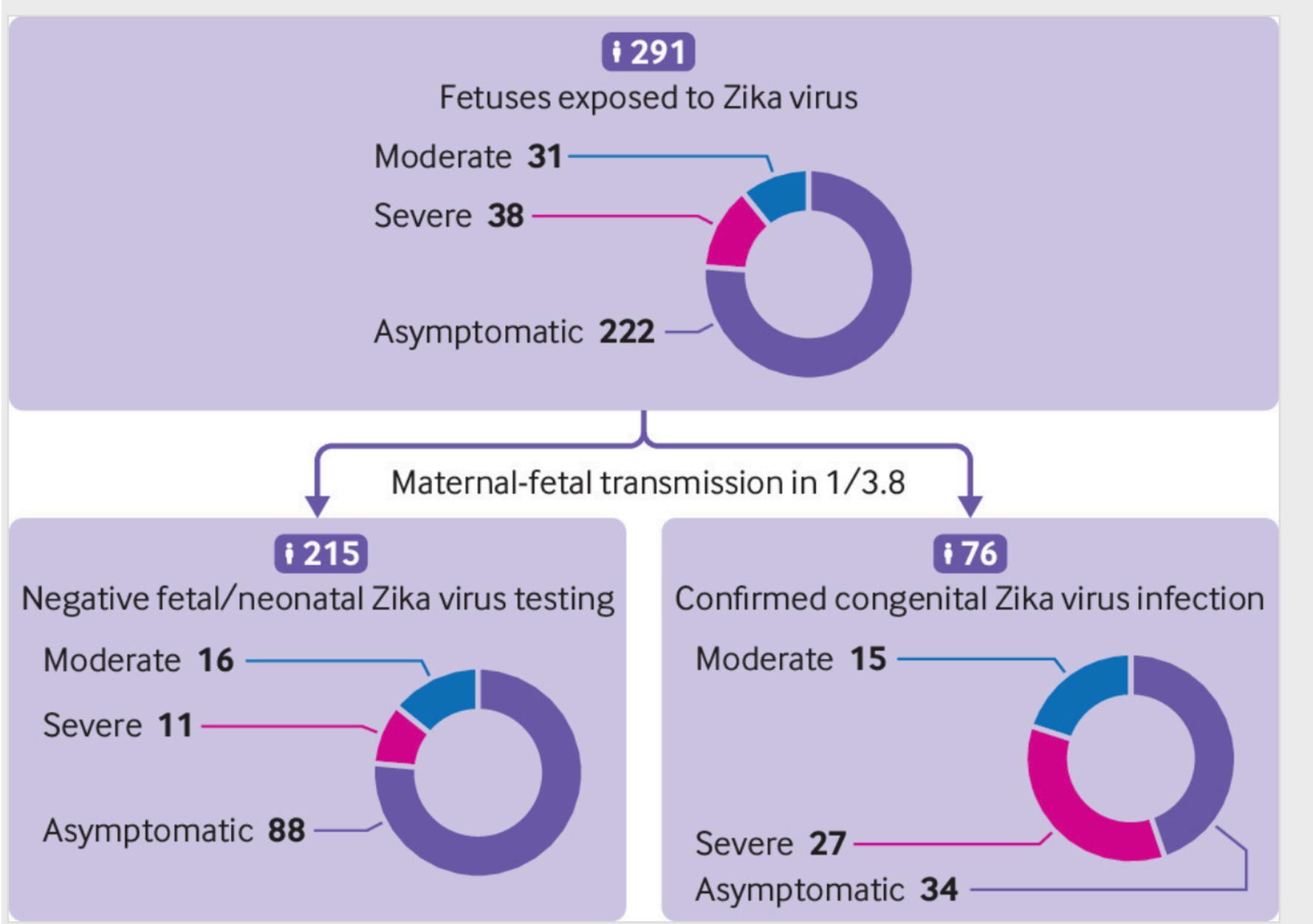
Zika: Study Design

Cross-Sectional — Case-Control — **Cohort** — RCT



Zika: Study Design

Cross-Sectional . . . . Case-Control . . . . **Cohort** . . . . RCT



Pomar (2018)

General: Study Design

Cross-Sectional — Case-Control — Cohort — RCT

Randomized Control Study

- Experiment in which subjects are **randomly allocated into groups** (test and control that are comparable) to receive or not to receive a preventative or a therapeutic procedure or intervention.
  Expérience dans laquelle les sujets sont répartis au hasard dans des groupes (test et contrôle comparables) pour recevoir ou non une procédure ou une intervention préventive ou thérapeutique.

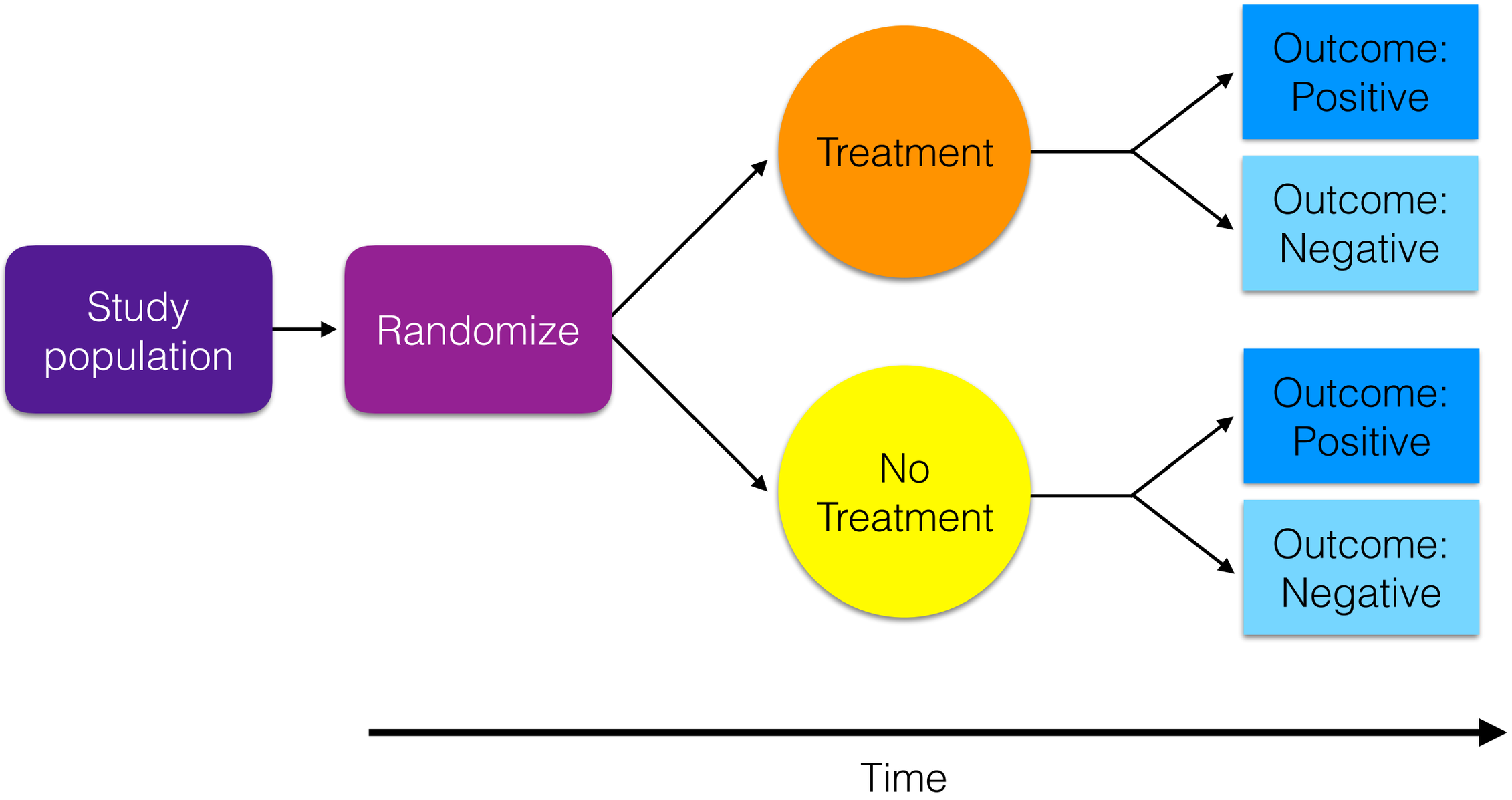
- Results are assessed by comparison of rates of disease, death, recovery, or other outcome in the study groups.

- Generally thought of as the **most rigorous method** of hypothesis testing.
  méthode la plus rigoureuse

- **Randomization should be blinded!**
  La randomisation devrait être aveuglée!

General: Study Design

Cross-Sectional — Case-Control — Cohort — RCT



Zika: Study Design

Cross-Sectional ..... Case-Control ..... Cohort ..... RCT

Study population

Treatment

Outcome: Positive

Outcome: Negative

What is the effectiveness of a Zika Containing Vaccine?

No Treatment

Outcome: Positive

Outcome: Negative

Time

Zika: Study Design

Cross-Sectional — Case-Control — Cohort — **RCT**

28 health volunteers without a history of flavivirus infection

Randomize

- Vaccine
  - Antibodies
  - No antibodies
- No Vaccine
  - Antibodies
  - No antibodies

Durban and Whitehead
First in-human trial
live attenuated Zika vaccine
rZIKV/D4Δ30-713

Time

How does study design happen in practice?

Ohatra: The MAHERY-CRS cross-sectional health survey (2017)

The MAHERY-CRS cross-sectional health survey (2017)

Objectives:

1. Perform a cross-sectional sample to determine the prevalence of several diseases and nutritional deficiencies. (*Aiza no tena misy tazomoka, tsy fahampian-dra, tsy fahampian-tsakafo, sns?*)

2. Survey rural communities (*tanàna kely*) in Madagascar: *Ambanivolobe*

3. Compare across different regions (*faritra*) of Madagascar:
   - Vatovavy-fitovinany, Amoron'i Mania, Atsimo Andrefana

HARVARD UNIVERSITY

Madagascar Health and Environmental Research

CRS

How does study design happen in practice?

Ohatra: The MAHERY-CRS cross-sectional health survey (2017)

How does study design happen in practice?

Ohatra: The MAHERY-CRS cross-sectional health survey (2017)

OR

How to balance scientific and non-scientific factors when designing a study?

How to not be paralyzed (*aza miasa saina*)?

How does study design happen in practice?

Ohatra: The MAHERY-CRS cross-sectional health survey (2017)

OR

How to balance scientific and non-scientific factors when designing a study?

How to not be paralyzed (*aza miasa saina*)?

OR

How does study design place limits on your analysis?

How to do good science with imperfect studies (*'Science' tsara raha tsy lavorary ny 'study'*)?

How does study design happen in practice?

Ohatra: The MAHERY-CRS cross-sectional health survey (2017)

OR

How to balance scientific and non-scientific factors when designing a study?

How to not be paralyzed (*aza miasa saina*)?

OR

How does study design place limits on your analysis?

How to do good science with imperfect studies (*'Science' tsara raha tsy lavorary ny 'study'*)?

The MAHERY-CRS cross-sectional health survey (2017)

The MAHERY-CRS cross-sectional health survey (2017)

HARVARD UNIVERSITY

Madagascar Health and Environmental Research

CRS

Limitations of cross-sectional surveys

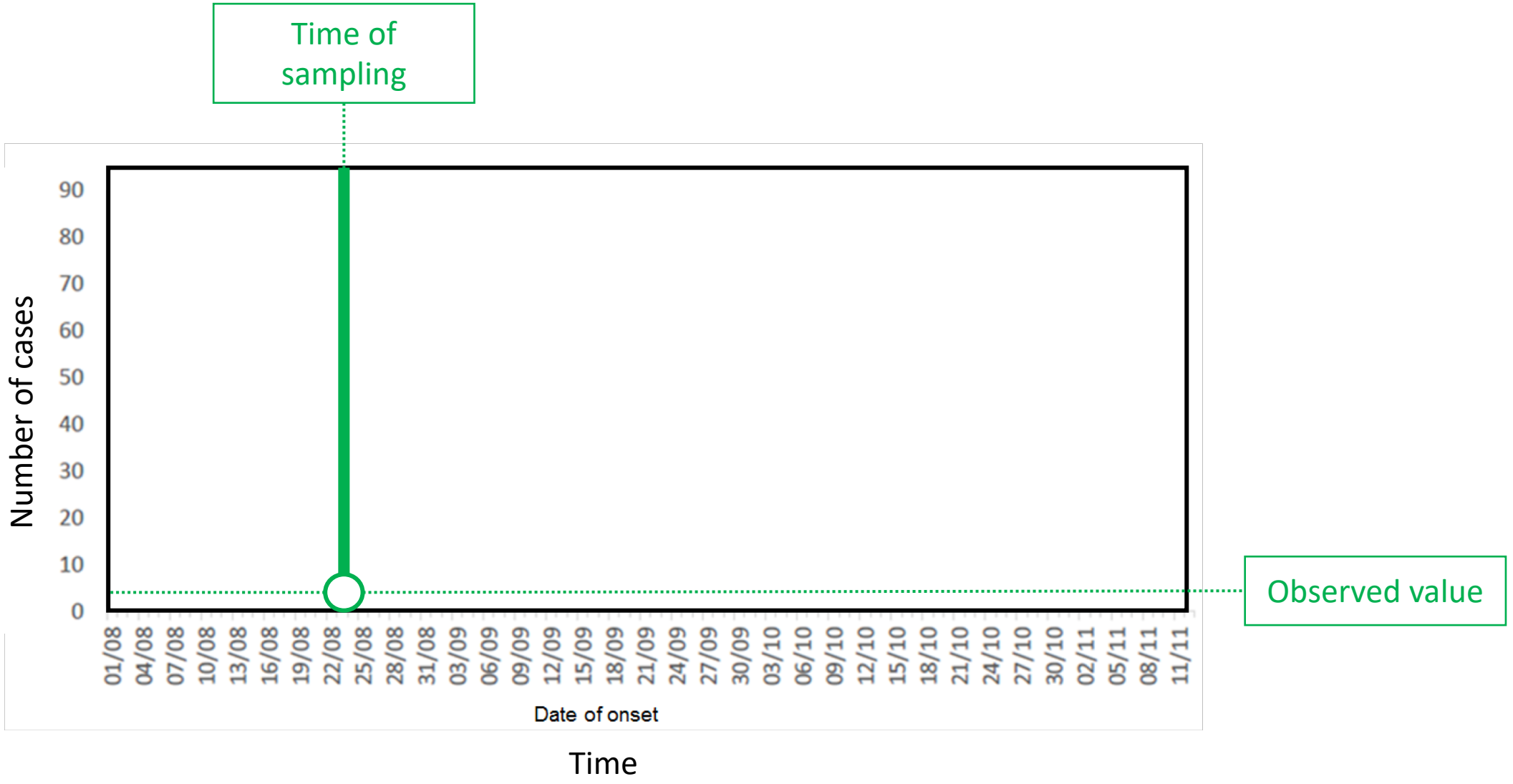
Example: Plague outbreak in Madagascar

What is a major limitation of cross-sectional surveys?

Inona ny olana maventy misakana ny 'cross-sectional surveys'?

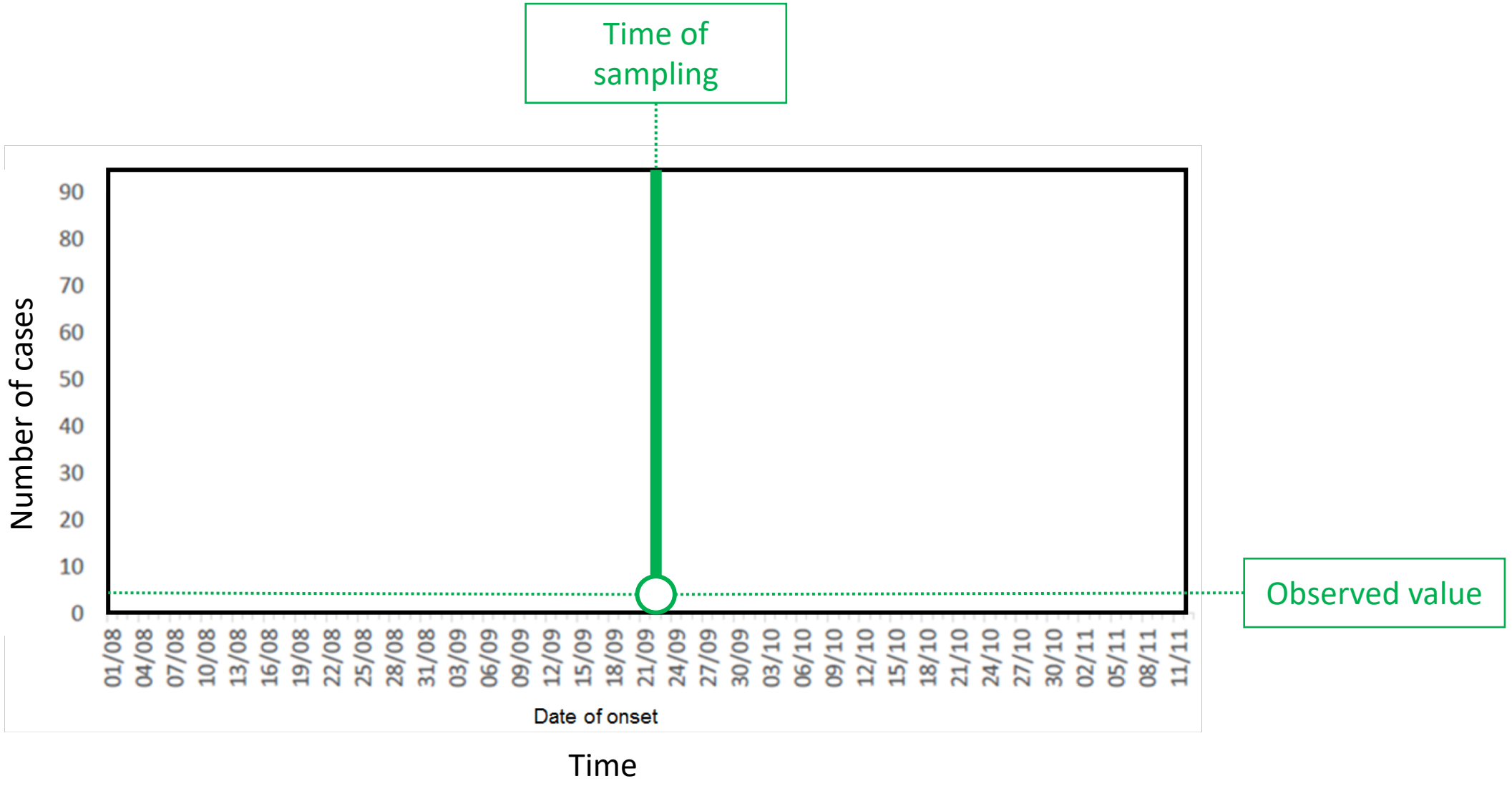
Limitations of cross-sectional surveys

Example: Plague outbreak in Madagascar



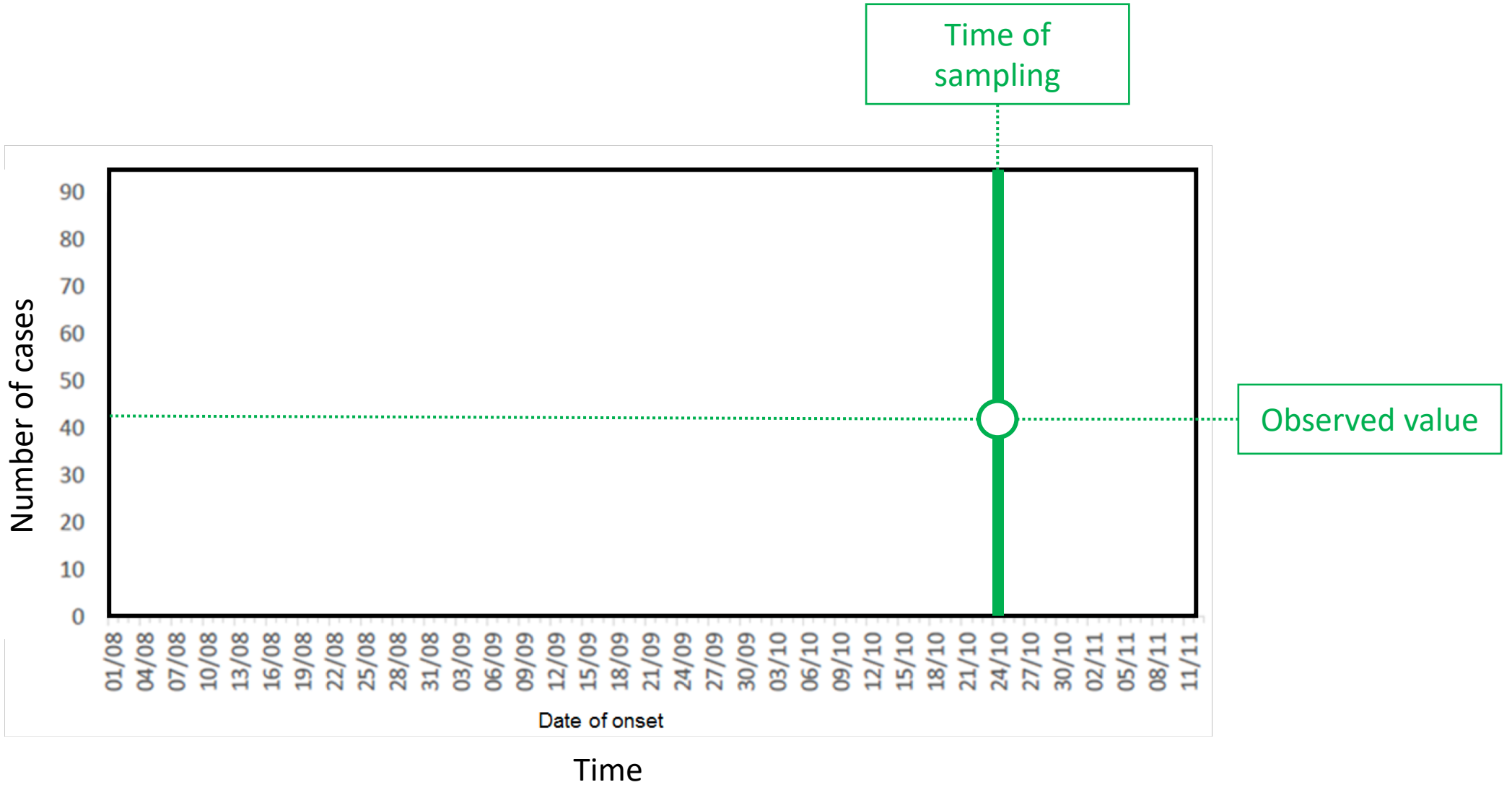
Limitations of cross-sectional surveys

Example: Plague outbreak in Madagascar



Limitations of cross-sectional surveys

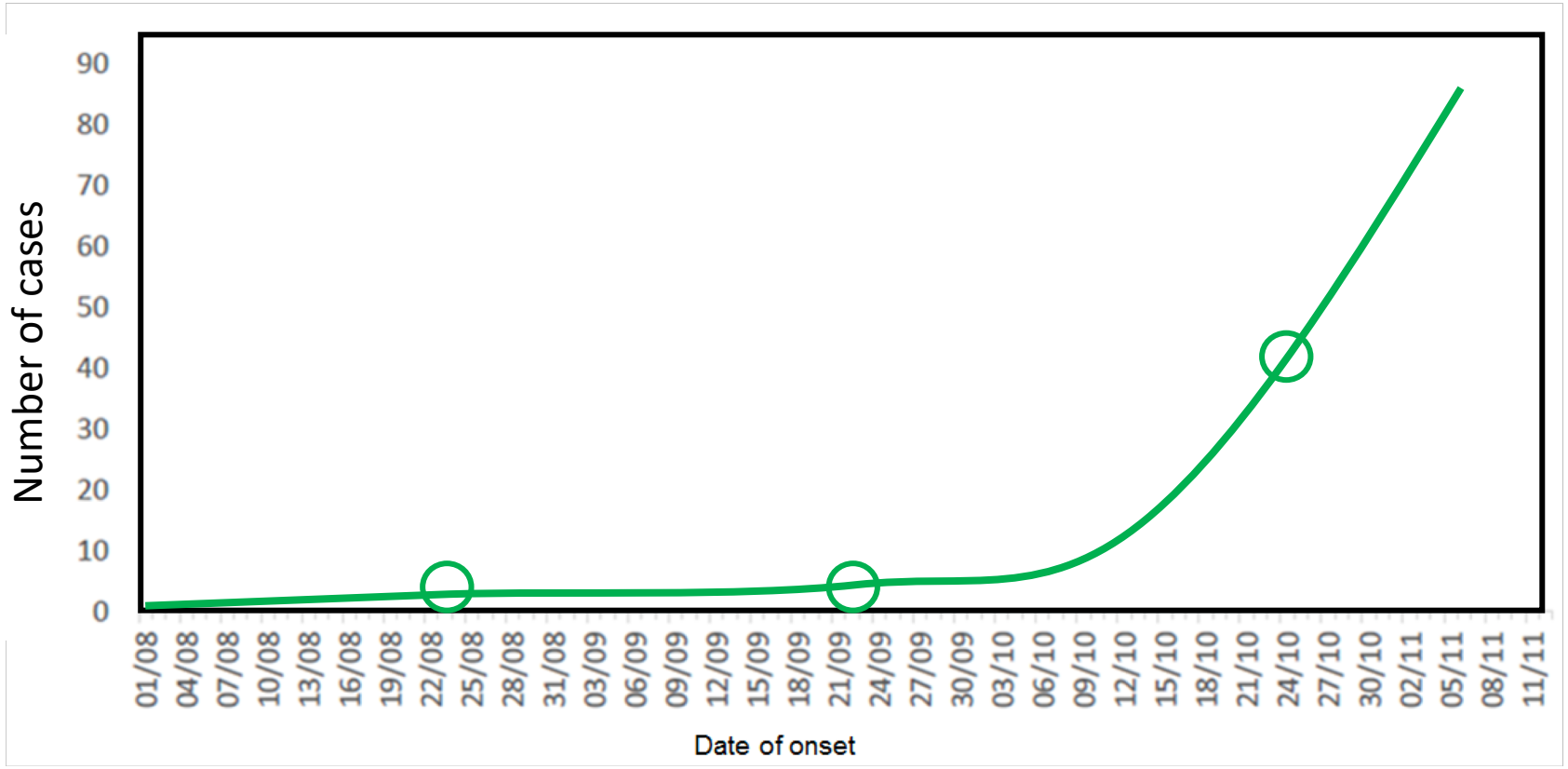
Example: Plague outbreak in Madagascar



Limitations of cross-sectional surveys

Example: Plague outbreak in Madagascar

What we might think is happening if we tried to extrapolate from cross-sectional surveys

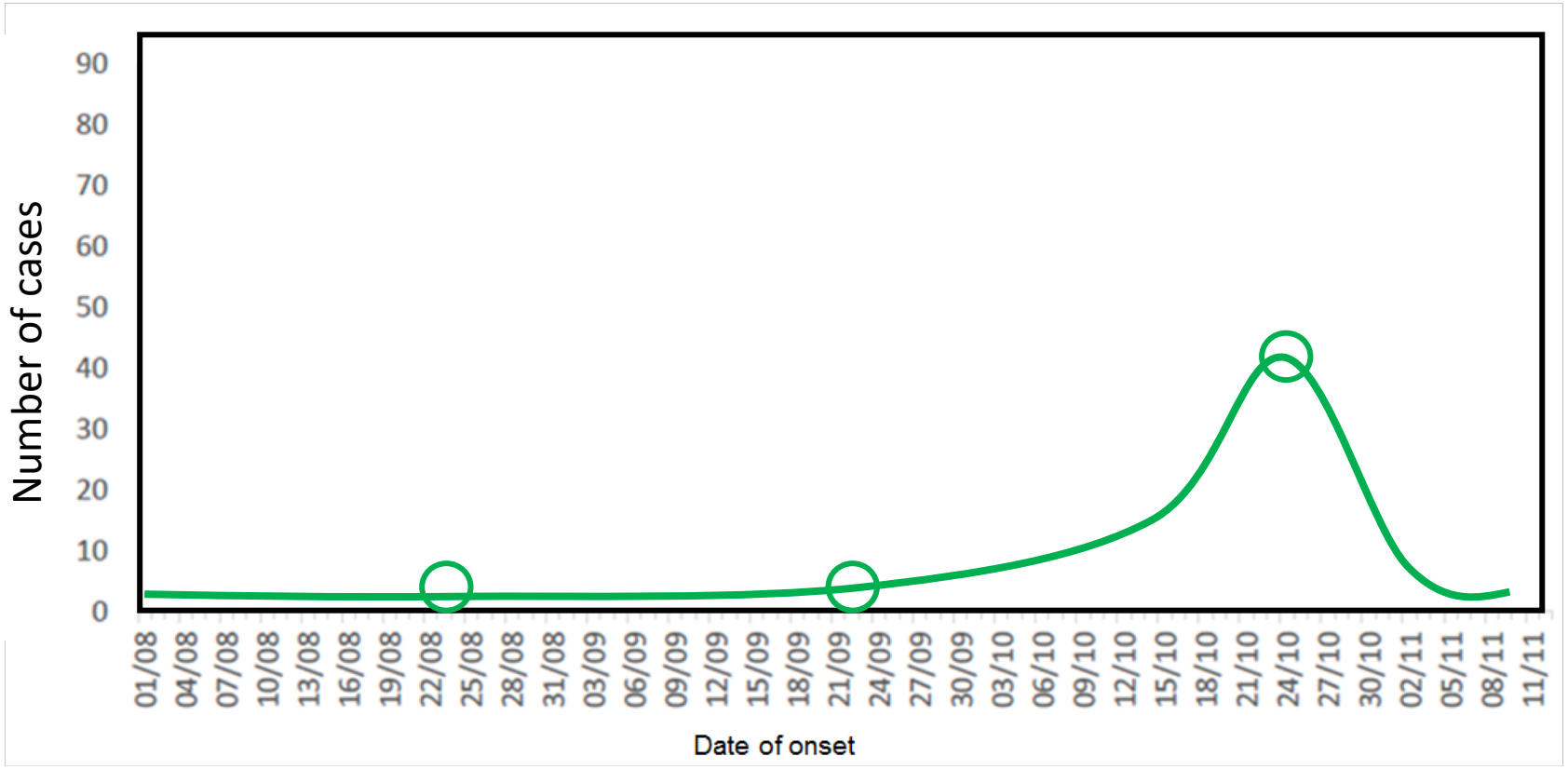


Time

Limitations of cross-sectional surveys

Example: Plague outbreak in Madagascar

What we might think is happening if we tried to extrapolate from cross-sectional surveys

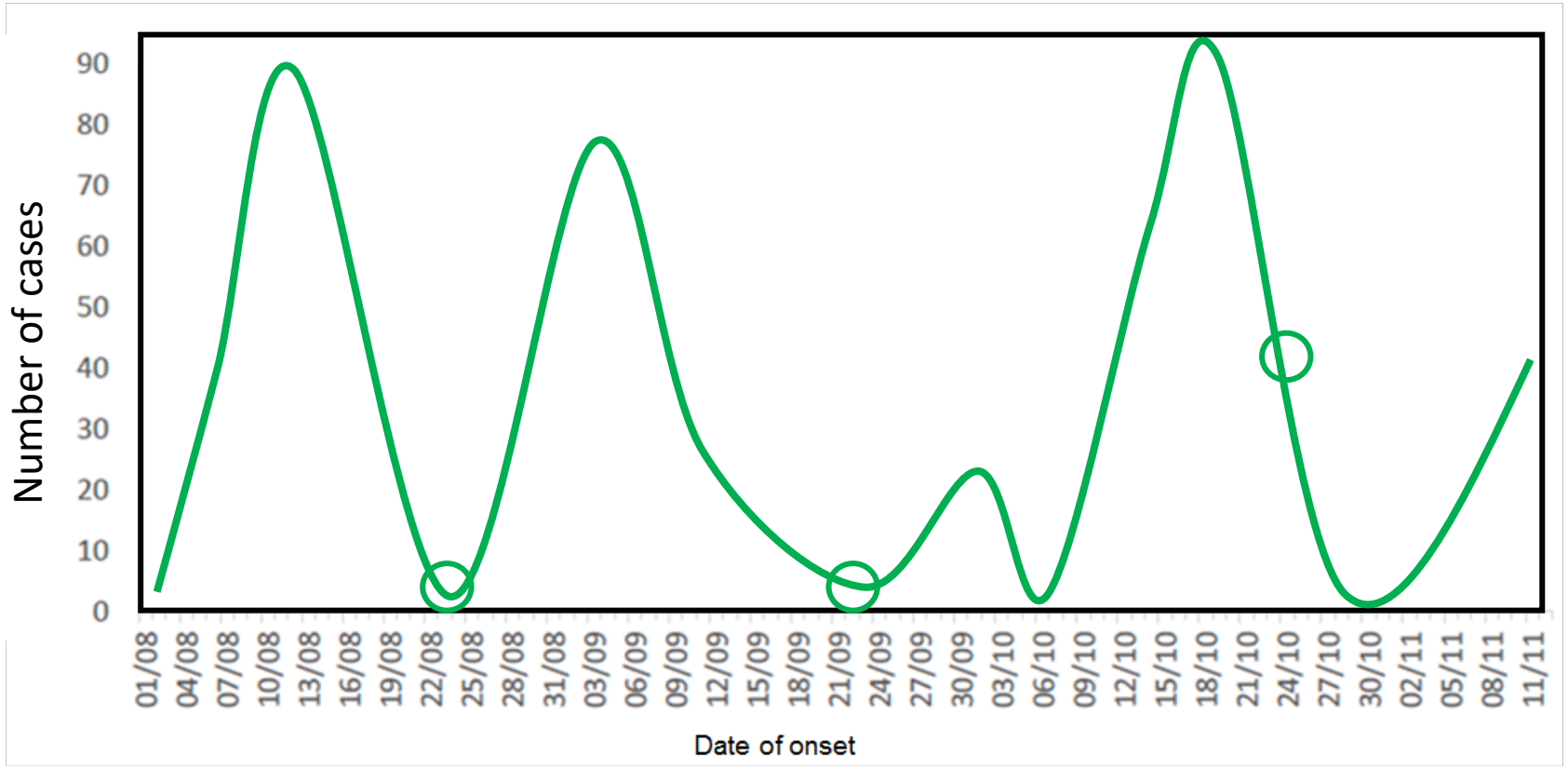


Time

Limitations of cross-sectional surveys

Example: Plague outbreak in Madagascar

What we might think is happening if we tried to extrapolate from cross-sectional surveys

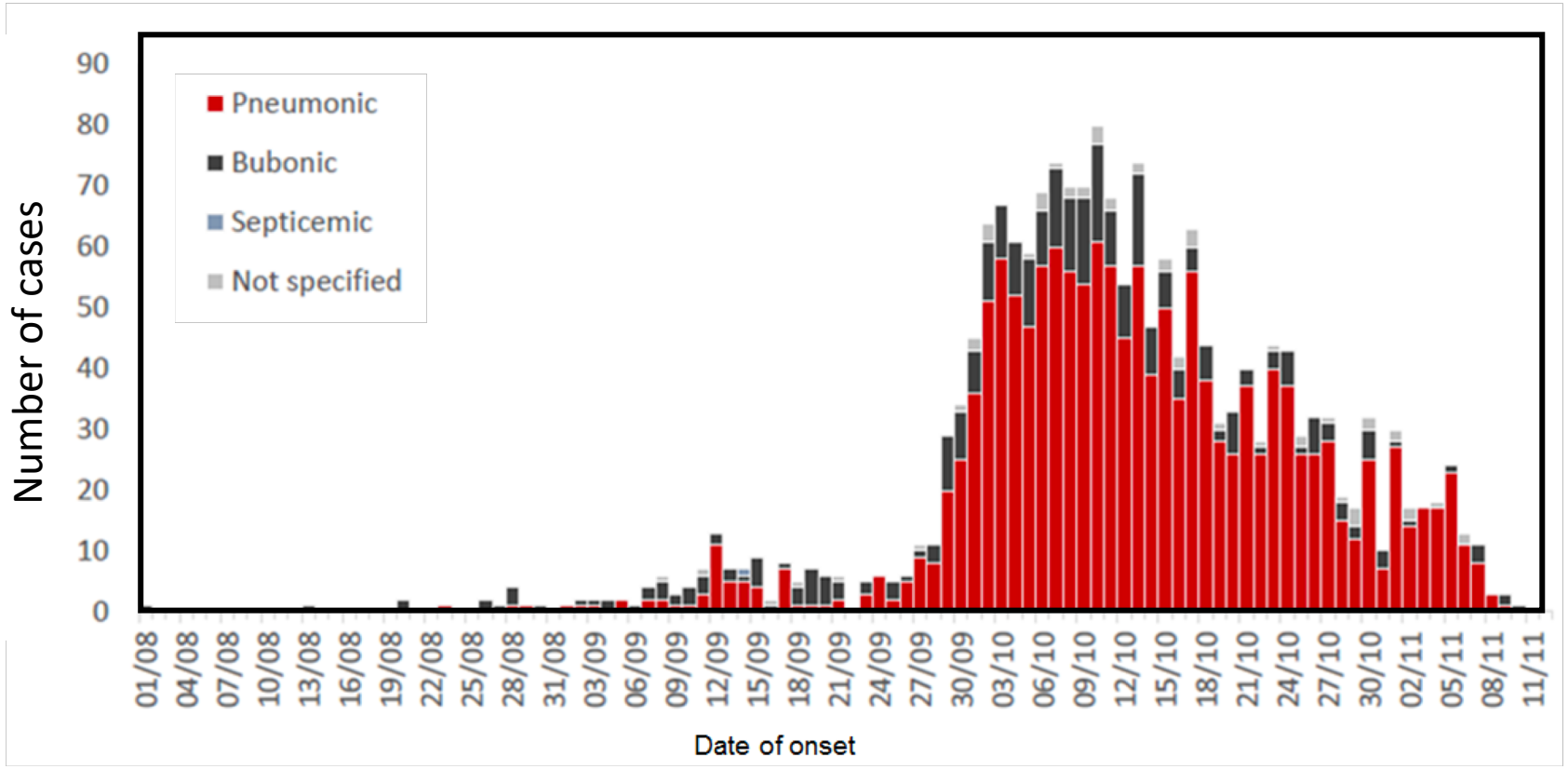


Time

Limitations of cross-sectional surveys

Example: Plague outbreak in Madagascar

What really happened with the plague



Time

Limitations of cross-sectional surveys

Example: Plague outbreak in Madagascar

What is a major limitation of cross-sectional surveys?

Inona ny olana maventy misakana ny 'cross-sectional surveys'?

Limitations of cross-sectional surveys

Example: Plague outbreak in Madagascar

What is a major limitation of cross-sectional surveys?

Inona ny olana maventy misakana ny ‘cross-sectional surveys’?

Difficult to capture changes over time

Sarotra manazava ny fiovana amin'ny fotoana

Limitations of cross-sectional surveys

Example: Plague outbreak in Madagascar

What is a major limitation of cross-sectional surveys?

Inona ny olana maventy no misy ny 'cross-sectional surveys'?

Difficult to capture changes over time

Sarotra manazava ny fiovana amin'ny fotoana

Conclusion: Not all study types can answer all the questions you have

Samy manana ny fanontaniana afaka valiany 'study' tsy iray-iray

How does study design happen in practice?

Ohatra: The MAHERY-CRS cross-sectional health survey (2017)

OR

How to balance scientific and non-scientific factors when designing a study?

How to not be paralyzed (*aza miasa saina*)?

OR

How does study design place limits on your analysis?

How to do good science with imperfect studies (*'Science' tsara raha tsy lavorary ny 'study'*)?

How does study design happen in practice?

Ohatra: The MAHERY-CRS cross-sectional health survey (2017)

OR

How to balance scientific and non-scientific factors when designing a study?

How to not be paralyzed (*aza miasa saina*)?

OR

How does study design place limits on your analysis?

How to do good science with imperfect studies (*'Science' tsara raha tsy lavorary ny 'study'*)?

Limitations of cross-sectional surveys

Example: The MAHERY-CRS cross-sectional health survey (2017)

Problems:

Money is not unlimited (Budget is $175,000 USD)	E.g. transportation, materials, analysis,
Time is not unlimited (Time window: 1 year)	E.g. my advisor starts to miss me
Subject participation is not unlimited	E.g. blood draws are not fun
Need to balance multiple objectives (malaria, viruses, nutrition, etc)	E.g. unknown effect sizes and statistical power

Limitations of cross-sectional surveys

Example: The MAHERY-CRS cross-sectional health survey (2017)

Problems:

Money is not unlimited (Budget is $175,000 USD)	E.g. transportation, materials, analysis,
Time is not unlimited (Time window: 1 year)	E.g. my advisor starts to miss me
Subject participation is not unlimited	E.g. blood draws are not fun
Need to balance multiple objectives (malaria, viruses, nutrition, etc)	E.g. unknown effect sizes and statistical power

Limitations of cross-sectional surveys

Example: The MAHERY-CRS cross-sectional health survey (2017)

Problems:

Money is not unlimited (Budget is $175,000 USD)	E.g. transportation, materials, analysis,
Time is not unlimited (Time window: 1 year)	E.g. my advisor starts to miss me
Subject participation is not unlimited	E.g. blood draws are not fun
Need to balance multiple objectives (malaria, viruses, nutrition, etc)	E.g. unknown effect sizes and statistical power

Limitations of cross-sectional surveys

Example: The MAHERY-CRS cross-sectional health survey (2017)

Problems:

Money is not unlimited (Budget is $175,000 USD)	E.g. transportation, materials, analysis,
Time is not unlimited (Time window: 1 year)	E.g. my advisor starts to miss me
Subject participation is not unlimited	E.g. blood draws are not fun
Need to balance multiple objectives (malaria, viruses, nutrition, etc)	E.g. unknown effect sizes and statistical power

Limitations of cross-sectional surveys

Example: The MAHERY-CRS cross-sectional health survey (2017)

Problems:

Money is not unlimited (Budget is $175,000 USD)	E.g. transportation, materials, analysis,
Time is not unlimited (Time window: 1 year)	E.g. my advisor starts to miss me
Subject participation is not unlimited	E.g. blood draws are not fun
Need to balance multiple objectives (malaria, viruses, nutrition, etc)	E.g. unknown effect sizes and statistical power

Limitations of cross-sectional surveys

Example: The MAHERY-CRS cross-sectional health survey (2017)

Problems:

Money is not unlimited (Budget is $175,000 USD)	E.g. transportation, materials, analysis,
Time is not unlimited (Time window: 1 year)	E.g. my advisor starts to miss me
Subject participation is not unlimited	E.g. blood draws are not fun
Need to balance multiple objectives (malaria, viruses, nutrition, etc)	E.g. unknown effect sizes and statistical power

Limitations of cross-sectional surveys

Example: The MAHERY-CRS cross-sectional health survey (2017)

Problems:

Money is not unlimited (Budget is $175,000 USD)	E.g. transportation, materials, analysis,
Time is not unlimited (Time window: 1 year)	E.g. my advisor starts to miss me
Subject participation is not unlimited	E.g. blood draws are not fun
Need to balance multiple objectives (malaria, viruses, nutrition, etc)	E.g. unknown effect sizes and statistical power

How does study design happen in practice?

Ohatra: The MAHERY-CRS cross-sectional health survey (2017)

OR

How to balance scientific and non-scientific factors when designing a study?

How to not be paralyzed (*aza miasa saina*)?

OR

How does study design place limits on your analysis?

How to do good science with imperfect studies (*'Science' tsara raha tsy lavorary ny 'study'*)?

How does study design happen in practice?

Ohatra: The MAHERY-CRS cross-sectional health survey (2017)

OR

How to balance scientific and non-scientific factors when designing a study?

How to not be paralyzed (*aza miasa saina*)?

OR

How does study design place limits on your analysis?

How to do good science with imperfect studies (*'Science' tsara raha tsy lavorary ny 'study'*)?

How can R help at every stage from day zero to the day of publication?

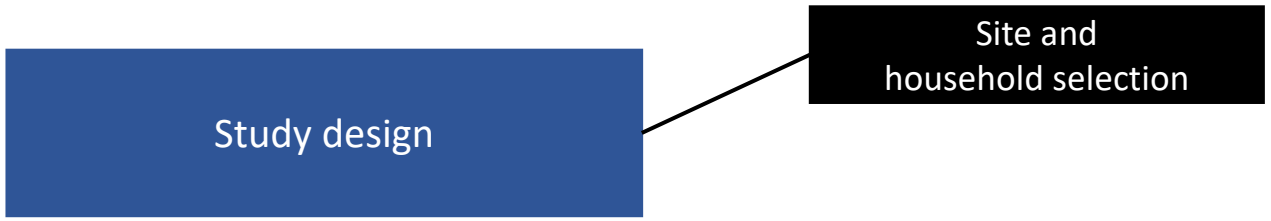
(*Amin'ny 'study', manomboka andro voalohany hatramin'ny andron'ny 'publication':*

Mora mora kokoa ve raha miasa amin'ny R?)

Using R in the MAHERY-CRS Project

Study design

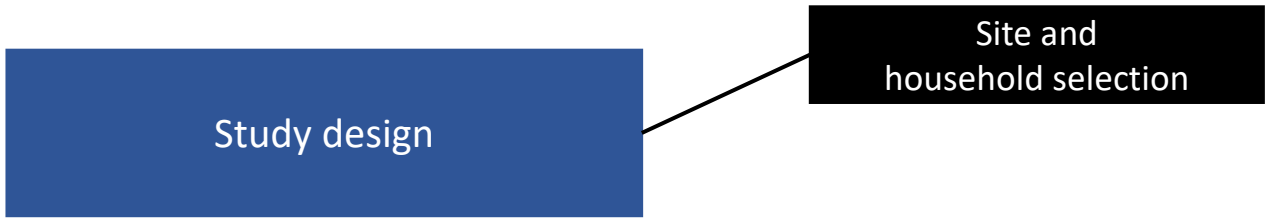
Using R in the MAHERY-CRS Project




First, create a list of all *tanana kely* in the study region


Second, create a list of all
tokantrano in the *tanana kely*

Using R in the MAHERY-CRS Project



Using R in the MAHERY-CRS Project

Study design

Site and household selection

```
##############################################################################
## Randomly selecting study sites
##############################################################################

# Import a table of study sites
study_site_table <- read.csv("study_sites.csv", stringsAsFactors = FALSE)

# Load the tidyverse package to access some useful functions
library(tidyverse)

# Randommly select 6 sites within each region
## use the group_by() and random() function
selected_sites_table <- group_by(study_site_table, region_id) %>%
  mutate(site_random_number = sample(1:length(region_id), length(region_id))) %>%
  mutate(site_selected_y_n =
           ifelse(site_random_number <= 6, 1, 0))

selected_sites_table <- selected_sites_table[
  selected_sites_table$site_selected_y_n == 1, ]
```

R code to randomly select sites and households: tidyverse package + group_by() function

Using R in the MAHERY-CRS Project

Study design

Site and household selection

```
##############################################################################
## Randomly selecting study sites
##############################################################################

# Import a table of study sites
study_site_table <- read.csv("study_sites.csv", stringsAsFactors = FALSE)

# Load the tidyverse package to access some useful functions
library(tidyverse)

# Randommly select 6 sites within each region
## use the group_by() and random() function
selected_sites_table <- group_by(study_site_table, region_id) %>%
  mutate(site_random_number = sample(1:length(region_id), length(region_id))) %>%
  mutate(site_selected_y_n =
           ifelse(site_random_number <= 6, 1, 0))

selected_sites_table <- selected_sites_table[
  selected_sites_table$site_selected_y_n == 1, ]
```

R code to randomly select sites and households: tidyverse package + group_by() function

```
> selected_sites_table
# A tibble: 24 x 4
# Groups:   region_id [4]
   region_id site_id site_random_number site_selected_y_n
   <fct>       <int>              <int>              <dbl>
 1 R1            101                  5                  1
 2 R1            102                  1                  1
 3 R1            103                  6                  1
 4 R1            104                  3                  1
 5 R1            109                  2                  1
 6 R1            110                  4                  1
 7 R2            201                  4                  1
 8 R2            202                  6                  1
```

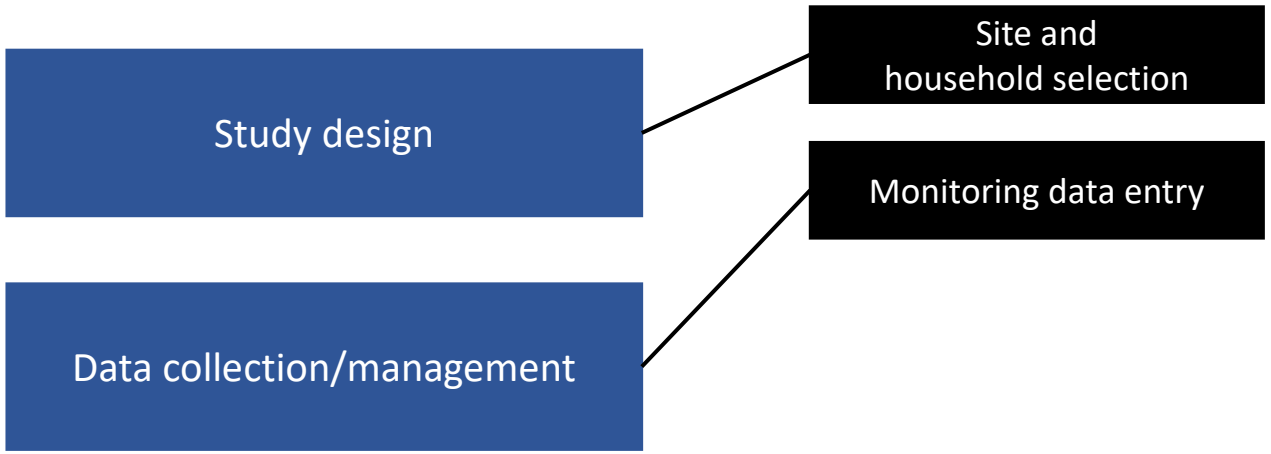
Using R in the MAHERY-CRS Project

Study design

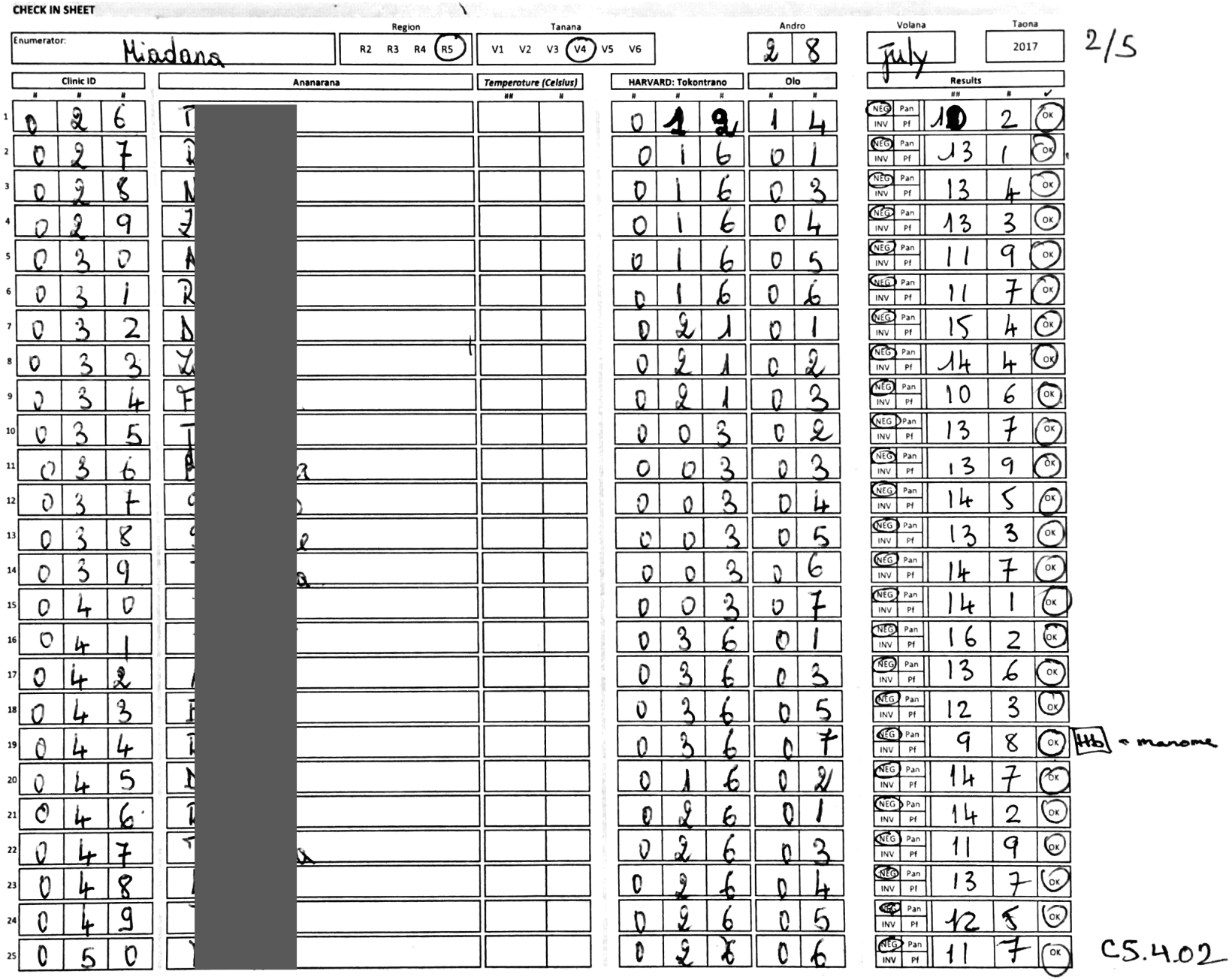
Site and household selection

Data collection/management

Using R in the MAHERY-CRS Project



Data Sheet 1



CHECK IN SHEET

Enumerator: Miadana

Region: R2 R3 R4 (R5)

Tanana: V1 V2 V3 (V4) V5 V6

Andro: 2 8

Volana: July

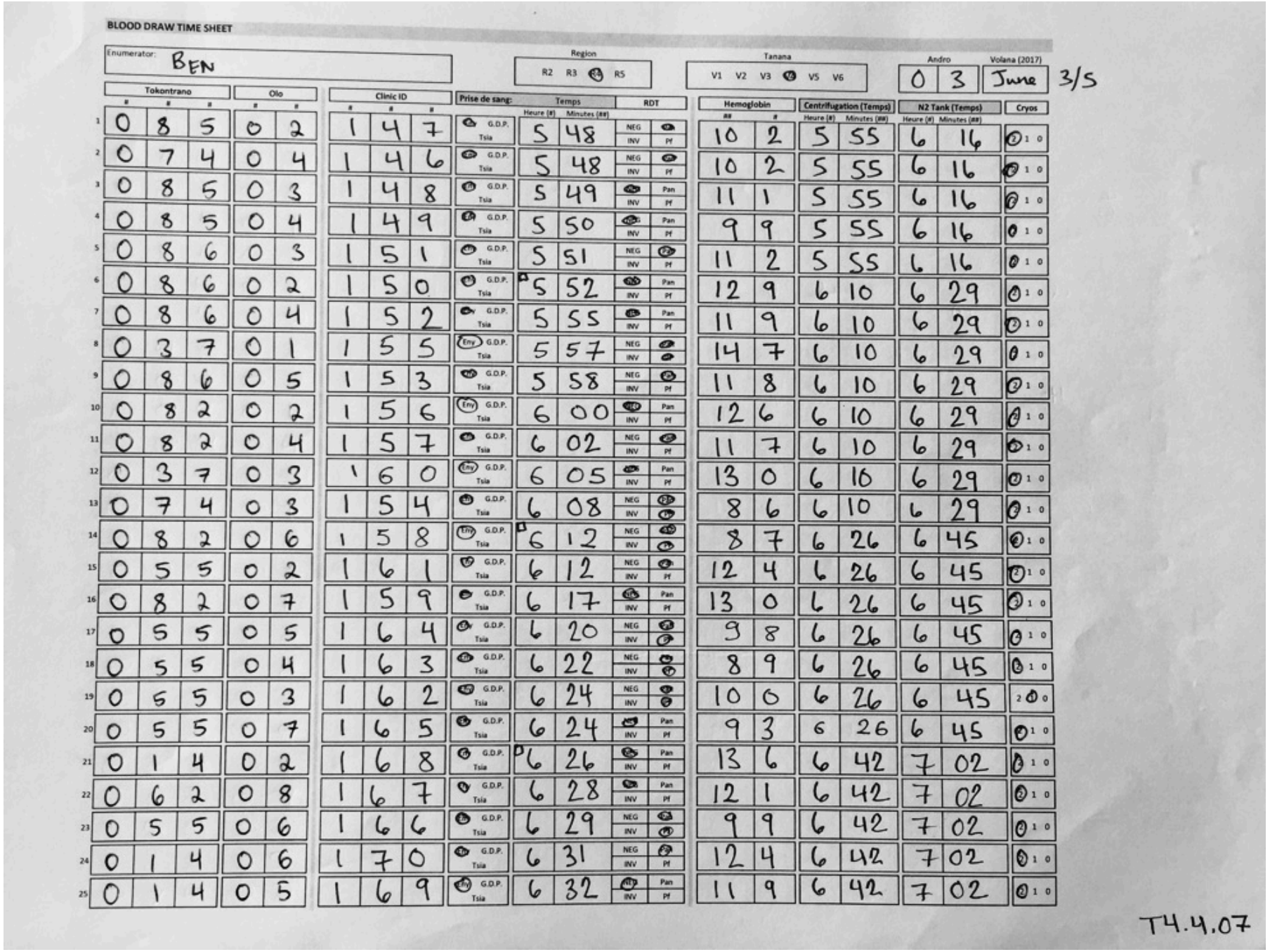
Taona: 2017

2/5

	Clinic ID			Ananarana	Temperature (Celsius)		HARVARD: Tokontrano			Olo			Results			
	#	#	#		##	#	#	#	#	#	#		##	#	✓	
1	0	2	6	[redacted]			0	1	2	1	4	NEG Pan INV Pf	10	2	OK	
2	0	2	7				0	1	6	0	1	NEG Pan INV Pf	13	1	OK	
3	0	2	8				0	1	6	0	3	NEG Pan INV Pf	13	4	OK	
4	0	2	9				0	1	6	0	4	NEG Pan INV Pf	13	3	OK	
5	0	3	0				0	1	6	0	5	NEG Pan INV Pf	11	9	OK	
6	0	3	1				0	1	6	0	6	NEG Pan INV Pf	11	7	OK	
7	0	3	2				0	2	1	0	1	NEG Pan INV Pf	15	4	OK	
8	0	3	3				0	2	1	0	2	NEG Pan INV Pf	14	4	OK	
9	0	3	4				0	2	1	0	3	NEG Pan INV Pf	10	6	OK	
10	0	3	5				0	0	3	0	2	NEG Pan INV Pf	13	7	OK	
11	0	3	6				0	0	3	0	3	NEG Pan INV Pf	13	9	OK	
12	0	3	7				0	0	3	0	4	NEG Pan INV Pf	14	5	OK	
13	0	3	8				0	0	3	0	5	NEG Pan INV Pf	13	3	OK	
14	0	3	9				0	0	3	0	6	NEG Pan INV Pf	14	7	OK	
15	0	4	0				0	0	3	0	7	NEG Pan INV Pf	14	1	OK	
16	0	4	1				0	3	6	0	1	NEG Pan INV Pf	16	2	OK	
17	0	4	2				0	3	6	0	3	NEG Pan INV Pf	13	6	OK	
18	0	4	3				0	3	6	0	5	NEG Pan INV Pf	12	3	OK	
19	0	4	4				0	3	6	0	7	NEG Pan INV Pf	9	8	OK	Hb = marome
20	0	4	5				0	1	6	0	2	NEG Pan INV Pf	14	7	OK	
21	0	4	6				0	2	6	0	1	NEG Pan INV Pf	14	2	OK	
22	0	4	7				0	2	6	0	3	NEG Pan INV Pf	11	9	OK	
23	0	4	8				0	2	6	0	4	NEG Pan INV Pf	13	7	OK	
24	0	4	9				0	2	6	0	5	NEG Pan INV Pf	12	5	OK	
25	0	5	0				0	2	8	0	6	NEG Pan INV Pf	11	7	OK	

C5.4.02

Data Sheet 2



BLOOD DRAW TIME SHEET

Enumerator: BEN

Region: R2 R3 (R4) R5

Tanana: V1 V2 V3 (V4) V5 V6

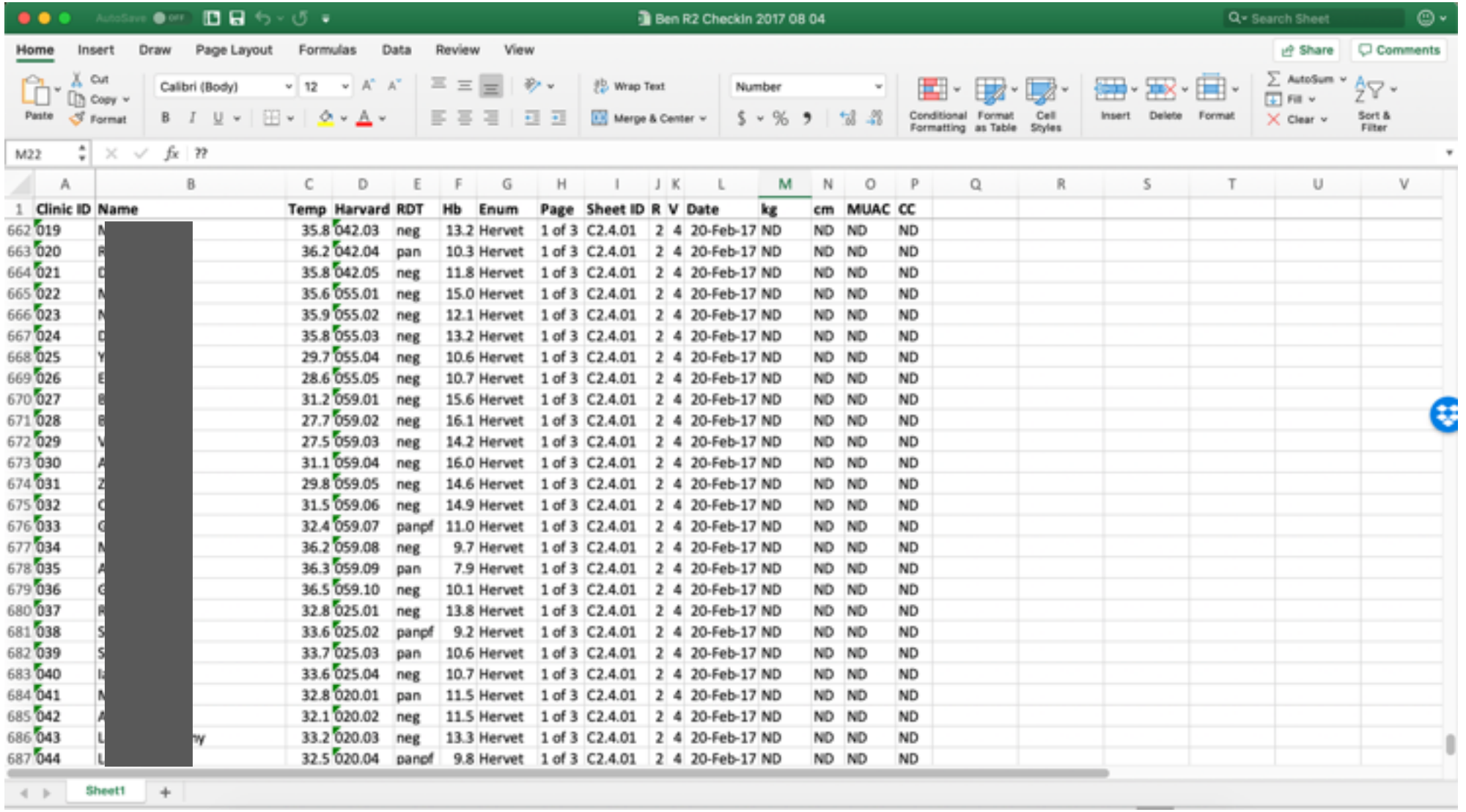
Andro: 0 3

Volana (2017): June 3/5

	Tokontrano			Olo		Clinic ID			Prise de sang:	Temps Heure (#)	Temps Minutes (##)	RDT		Hemoglobin		Centrifugation (Temps) Heure (#)	Centrifugation (Temps) Minutes (##)	N2 Tank (Temps) Heure (#)	N2 Tank (Temps) Minutes (##)	Cryos
	#	#	#	#	#	#	#	#						##	#						
1	0	8	5	0	2	1	4	7	G.D.P. Tsia	5	48	NEG INV	(4) Pf	10	2	5	55	6	16	0 1 0
2	0	7	4	0	4	1	4	6	G.D.P. Tsia	5	48	NEG INV	(4) Pf	10	2	5	55	6	16	0 1 0
3	0	8	5	0	3	1	4	8	G.D.P. Tsia	5	49	(4) INV	Pan Pf	11	1	5	55	6	16	0 1 0
4	0	8	5	0	4	1	4	9	G.D.P. Tsia	5	50	(4) INV	Pan Pf	9	9	5	55	6	16	0 1 0
5	0	8	6	0	3	1	5	1	G.D.P. Tsia	5	51	NEG INV	(4) Pf	11	2	5	55	6	16	0 1 0
6	0	8	6	0	2	1	5	0	G.D.P. Tsia	5	52	(4) INV	Pan Pf	12	9	6	10	6	29	0 1 0
7	0	8	6	0	4	1	5	2	G.D.P. Tsia	5	55	(4) INV	Pan Pf	11	9	6	10	6	29	0 1 0
8	0	3	7	0	1	1	5	5	(Env) G.D.P. Tsia	5	57	NEG INV	(4)	14	7	6	10	6	29	0 1 0
9	0	8	6	0	5	1	5	3	G.D.P. Tsia	5	58	NEG INV	(4) Pf	11	8	6	10	6	29	0 1 0
10	0	8	2	0	2	1	5	6	(Env) G.D.P. Tsia	6	00	(4) INV	Pan Pf	12	6	6	10	6	29	0 1 0
11	0	8	2	0	4	1	5	7	G.D.P. Tsia	6	02	NEG INV	(4) Pf	11	7	6	10	6	29	0 1 0
12	0	3	7	0	3	1	6	0	(Env) G.D.P. Tsia	6	05	(4) INV	Pan Pf	13	0	6	10	6	29	0 1 0
13	0	7	4	0	3	1	5	4	G.D.P. Tsia	6	08	NEG INV	(4)	8	6	6	10	6	29	0 1 0
14	0	8	2	0	6	1	5	8	(Env) G.D.P. Tsia	6	12	NEG INV	(4)	8	7	6	26	6	45	0 1 0
15	0	5	5	0	2	1	6	1	G.D.P. Tsia	6	12	NEG INV	(4) Pf	12	4	6	26	6	45	0 1 0
16	0	8	2	0	7	1	5	9	G.D.P. Tsia	6	17	(4) INV	Pan Pf	13	0	6	26	6	45	0 1 0
17	0	5	5	0	5	1	6	4	G.D.P. Tsia	6	20	NEG INV	(4)	9	8	6	26	6	45	0 1 0
18	0	5	5	0	4	1	6	3	G.D.P. Tsia	6	22	NEG INV	(4)	8	9	6	26	6	45	0 1 0
19	0	5	5	0	3	1	6	2	G.D.P. Tsia	6	24	NEG INV	(4)	10	0	6	26	6	45	2 0 0
20	0	5	5	0	7	1	6	5	G.D.P. Tsia	6	24	(4) INV	Pan Pf	9	3	6	26	6	45	0 1 0
21	0	1	4	0	2	1	6	8	G.D.P. Tsia	6	26	(4) INV	Pan Pf	13	6	6	42	7	02	0 1 0
22	0	6	2	0	8	1	6	7	G.D.P. Tsia	6	28	(4) INV	Pan Pf	12	1	6	42	7	02	0 1 0
23	0	5	5	0	6	1	6	6	G.D.P. Tsia	6	29	NEG INV	(4)	9	9	6	42	7	02	0 1 0
24	0	1	4	0	6	1	7	0	G.D.P. Tsia	6	31	NEG INV	(4) Pf	12	4	6	42	7	02	0 1 0
25	0	1	4	0	5	1	6	9	(Env) G.D.P. Tsia	6	32	(4) INV	Pan Pf	11	9	6	42	7	02	0 1 0

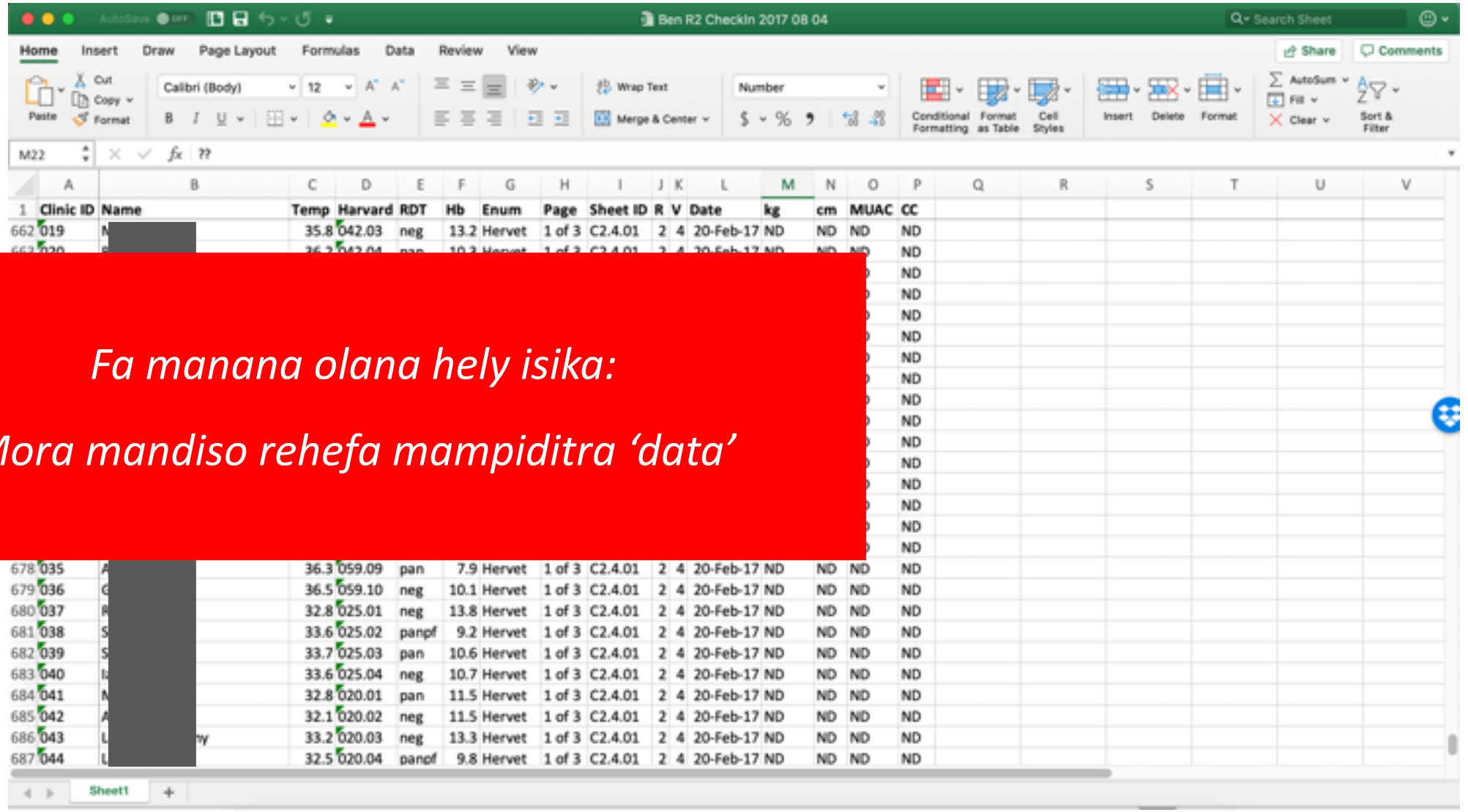
T4.4.07

Data entry into excel



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Clinic ID	Name	Temp	Harvard	RDT	Hb	Enum	Page	Sheet ID	R	V	Date	kg	cm	MUAC	CC
662	019	[redacted]	35.8	042.03	neg	13.2	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
663	020		36.2	042.04	pan	10.3	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
664	021		35.8	042.05	neg	11.8	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
665	022		35.6	055.01	neg	15.0	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
666	023		35.9	055.02	neg	12.1	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
667	024		35.8	055.03	neg	13.2	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
668	025		29.7	055.04	neg	10.6	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
669	026		28.6	055.05	neg	10.7	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
670	027		31.2	059.01	neg	15.6	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
671	028		27.7	059.02	neg	16.1	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
672	029		27.5	059.03	neg	14.2	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
673	030		31.1	059.04	neg	16.0	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
674	031		29.8	059.05	neg	14.6	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
675	032		31.5	059.06	neg	14.9	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
676	033		32.4	059.07	panpf	11.0	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
677	034		36.2	059.08	neg	9.7	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
678	035		36.3	059.09	pan	7.9	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
679	036		36.5	059.10	neg	10.1	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
680	037		32.8	025.01	neg	13.8	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
681	038		33.6	025.02	panpf	9.2	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
682	039		33.7	025.03	pan	10.6	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
683	040		33.6	025.04	neg	10.7	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
684	041		32.8	020.01	pan	11.5	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
685	042		32.1	020.02	neg	11.5	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
686	043		33.2	020.03	neg	13.3	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND
687	044		32.5	020.04	panpf	9.8	Hervet	1 of 3	C2.4.01	2	4	20-Feb-17	ND	ND	ND	ND

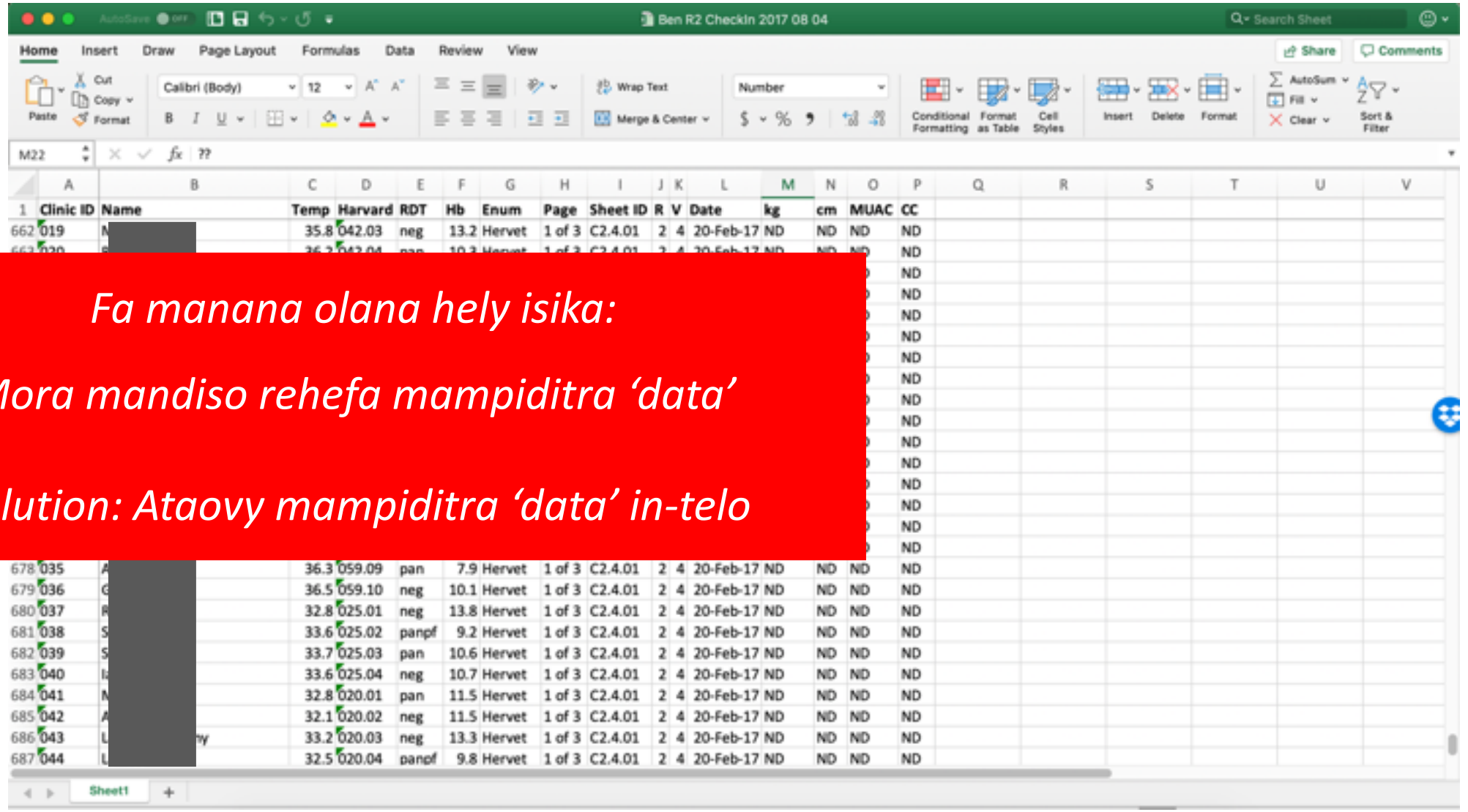
Data entry into excel



Fa manana olana hely isika:

Mora mandiso rehefa mampiditra 'data'

Data entry into excel

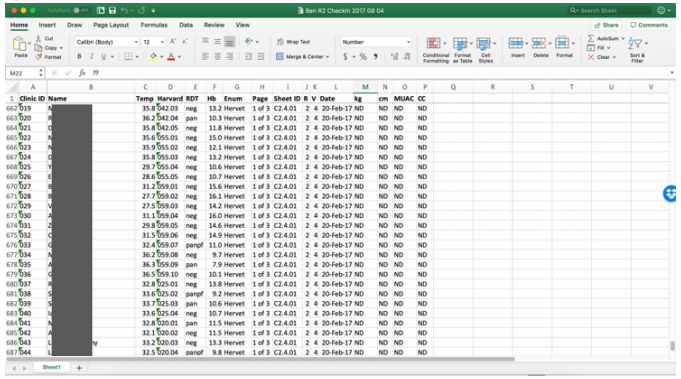


Fa manana olana hely isika:

Mora mandiso rehefa mampiditra 'data'

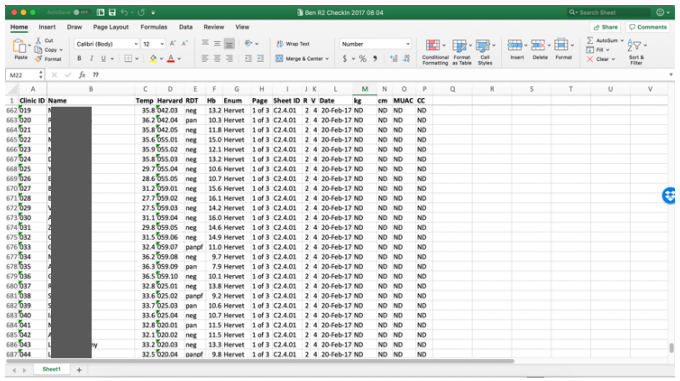
Solution: Ataovy mampiditra 'data' in-telo

Data entry into excel
(Repeat to minimize errors)



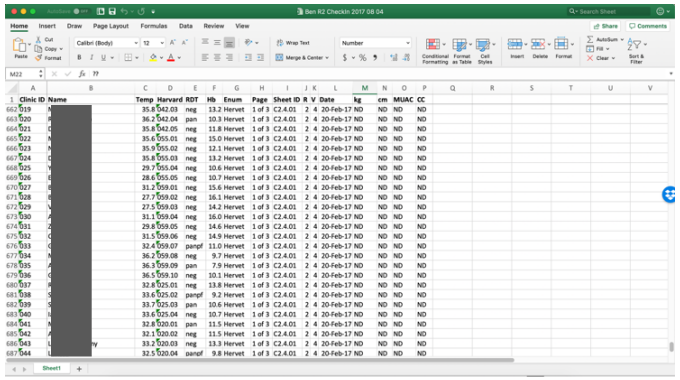
Replicate 1

+



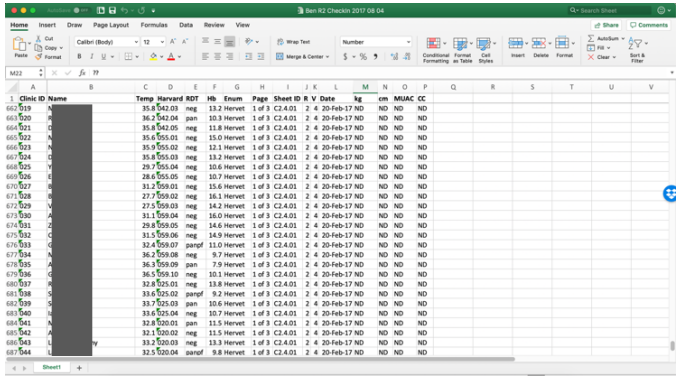
Replicate 2

+



Replicate 3

Data entry into excel
(Repeat to minimize errors)



Replicate 1

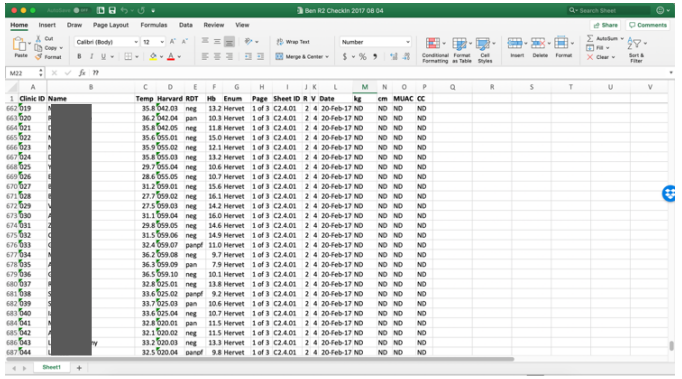
Fa hitondra olana hely hafa:

Mila mihaza ny ‘errors’ / Mila mizaha ny ‘errors’

Solution: R code

eplicate 2

+



Replicate 3

Using R in the MAHERY-CRS Project

Study design

Data collection/management

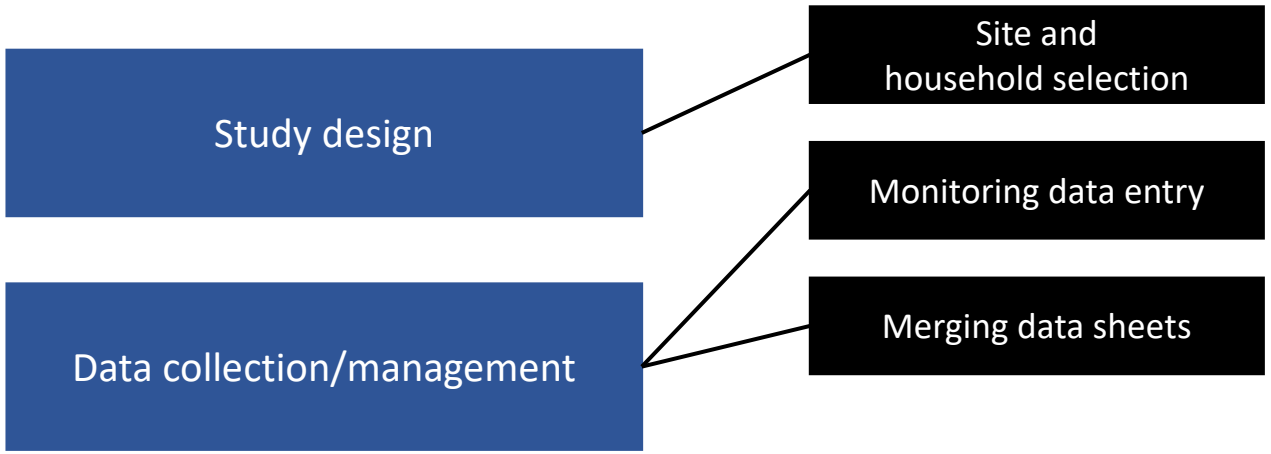
Site and household selection

Monitoring data entry

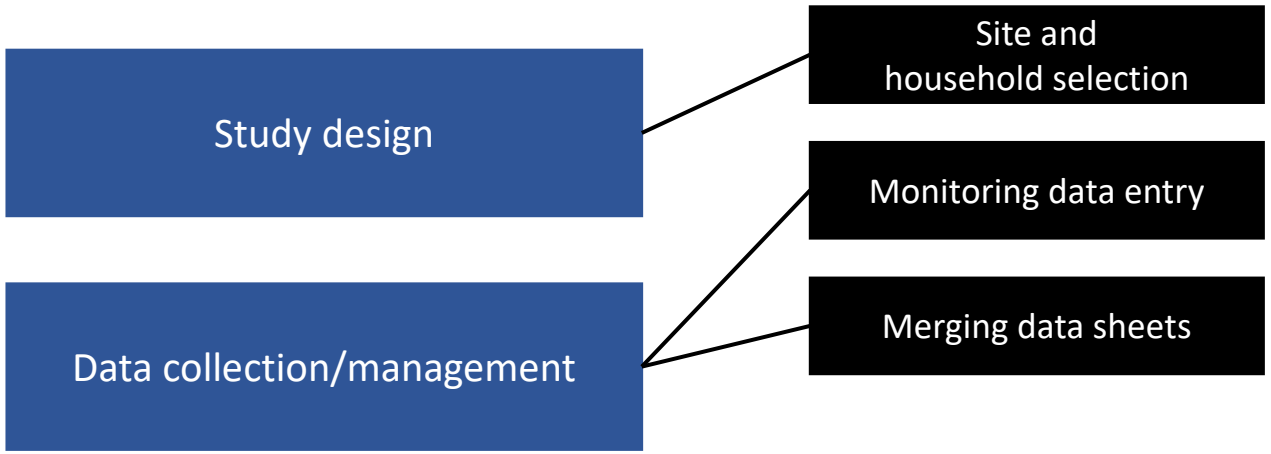
```r
##############################################################################
## Monitoring data entry
##############################################################################

# Concatenate data fields in a row into one string using a function and a for loop
concatenate.function <- function(dataframe){
  df <- dataframe
  concatenated_df <- rep(0, length(df1[,1]))
  for(i in 1:length(df1[,1])){
    concatenated_df[i] <- paste(
      df[i, 1], df[i, 2], df[i, 3], df[i, 4], df[i, 5], sep = ".")
  }
  return(concatenated_df)
}

df1.concatenated <- concatenate.function(df1)
df2.concatenated <- concatenate.function(df2)

# compare the repeated data entries to see if entries are the same
compare.function <- function(dataframe1, dataframe2){
  rows_with_errors <- rep(NA, length(dataframe1))
  for(i in 1:length(dataframe1)){
    rows_with_errors[i] <- ifelse(dataframe1[i] == dataframe2[i], 0, 1)
  }
  return(rows_with_errors)
}

compare.function(df1.concatenated, df2.concatenated)
```

R code to compare data tables:
for loops + ifelse() function

Using R in the MAHERY-CRS Project



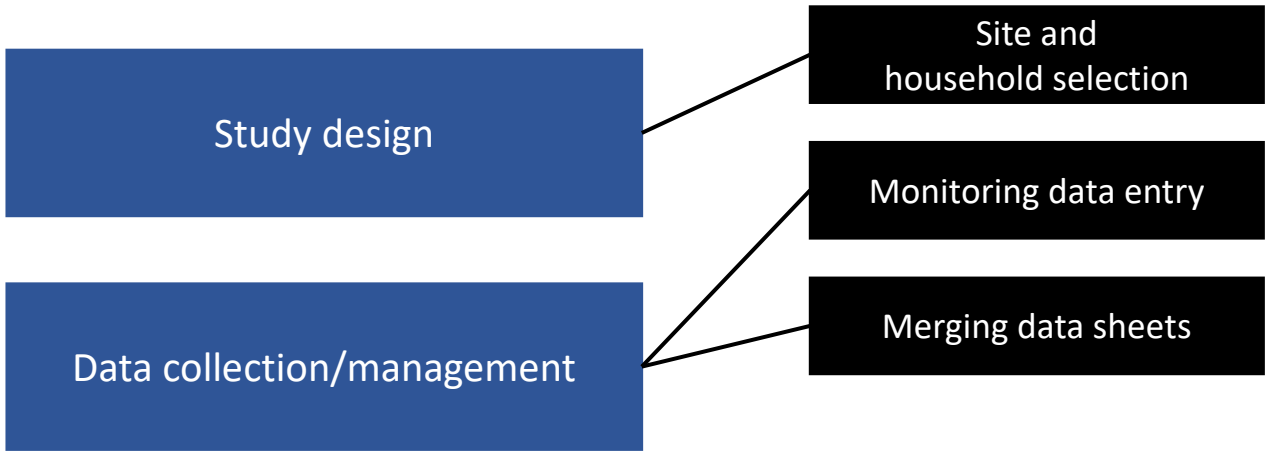
Using R in the MAHERY-CRS Project



Olana: Mila manambatra ‘data files’ roa

Clinical data + Survey data = Data to analyze

Using R in the MAHERY-CRS Project



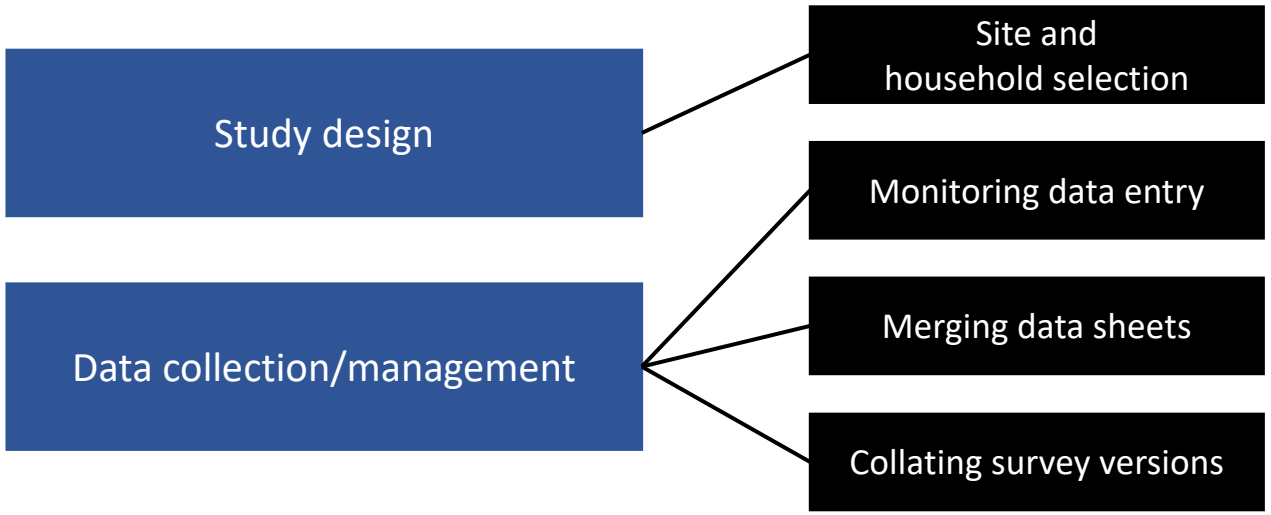
Olana: Mila manambatra ‘data files’ roa

Clinical data + Survey data = Data to analyze

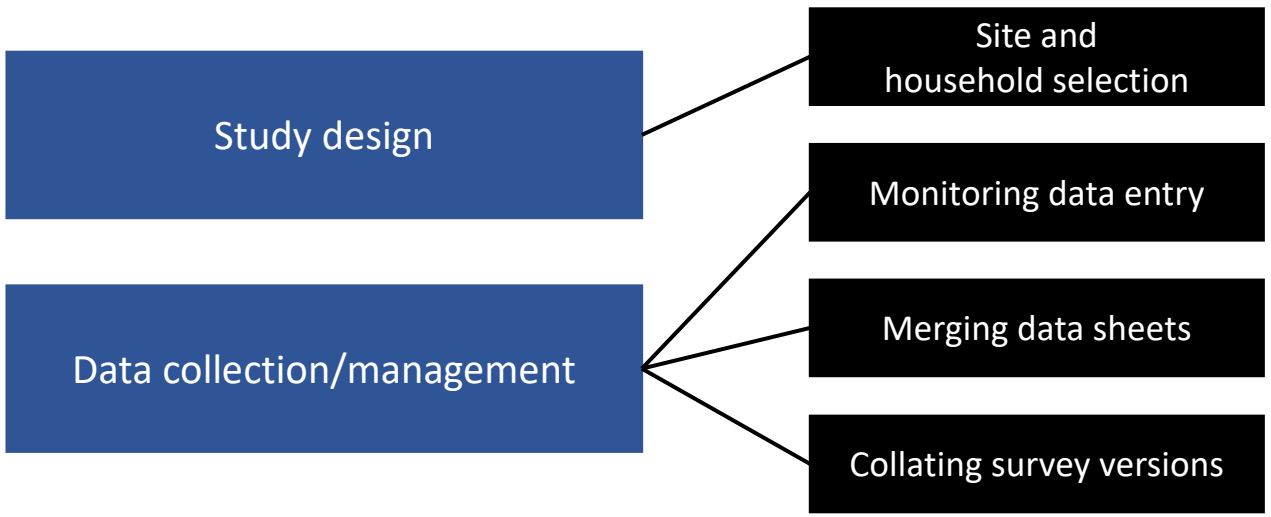
```
##############################################################################
## Merging data sheets
##############################################################################

# Import clinical data and survey data
clinical_data <- read.csv("clinical_data.csv", stringsAsFactors = FALSE)
survey_data <- read.csv("survey_data.csv", stringsAsFactors = FALSE)

# Subset the clinical data to keep the relevant variables
rdt_subset_variables <- c("unique_ind_id", "rdt_result")
clinical_data_trimmed <-clinical_data[rdt_subset_variables]

# Merge data frames using the full_join() function

# First, need to check for duplicate IDs using anyDuplicated()
# If the output of anyDuplicated > 0, then there are duplicate IDs
anyDuplicated(clinical_data_trimmed$unique_ind_id)
anyDuplicated(survey_data$unique_ind_id)

# Merge dataframes using the full_join() function and store as data_joined
data_joined <- full_join(clinical_data_trimmed, survey_data, by = "unique_ind_id")

# Check merged sheet
head(data_joined)
str(data_joined)
```

Check for duplicate IDs, merge using full_join()

Using R in the MAHERY-CRS Project



Using R in the MAHERY-CRS Project

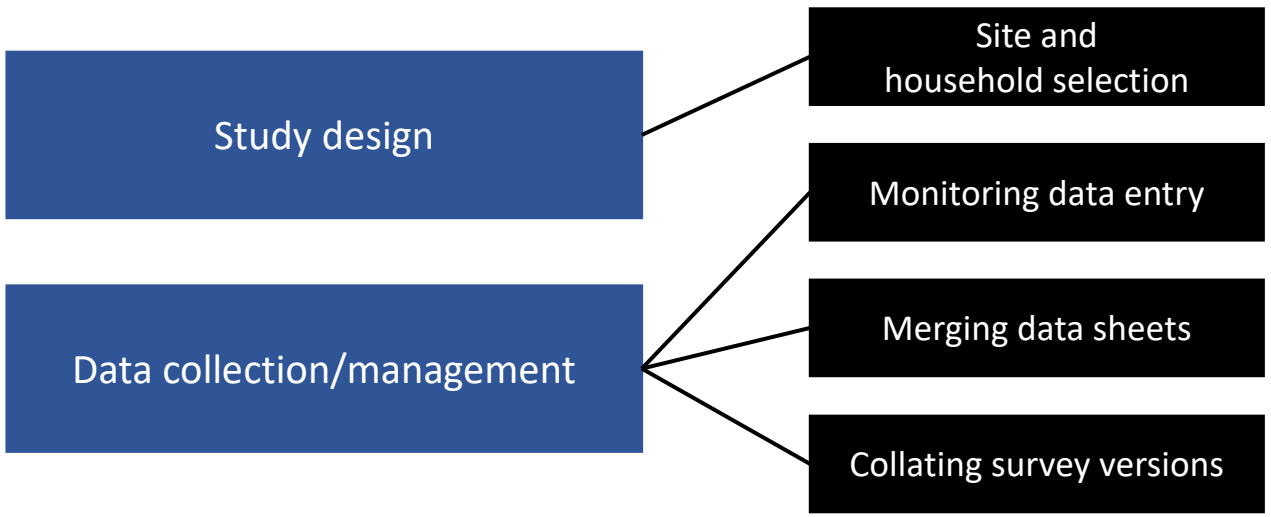


Olana be vata: Mila manambatra ‘versions’ efatra

Version 1 + Version 2 + Version 3 + Version 4

English question text	Malagasy question text	name_r5v	label_r5v	name_R3V2	label_R3V2	name_R3V1	label_R3V1	name_r2v	label_r2v	na
Section 00: Standard Initial Section	**Section 00: Standard premiere question**	_06_01_001	**Section 01: Standard pr**	_06_01_001	**Section 01: Standard**	_06_01_001	**Section 01: Standard**	_06_01_001	**Section 01: Standard**	_0
Site ID (standardized ID code for sites: Region Code + Site Code)	Site ID	_06_01_001_01	Site ID	_06_01_001_01	Site ID	_06_01_001_01	Site ID	_06_01_001_01	Site ID	_0
Region code (2 = Mananjary, 3 = Toliara, 4 = Morombe, 5 = Amoroni Mania)	Region	_06_01_001_02	Region	_06_01_001_02	Region	_06_01_001_02	Region	_06_01_001_02	Region	_0
Village code (1-6)	Village code (1-6)	_06_01_001_03	Village code (1-6)	_06_01_001_03	Village code (1-6)	_06_01_001_03	Village code (1-6)	_06_01_001_03	Village code (1-6)	_0
Unique Household ID (Region Code + Site Code + HH Code)	Household ID	_06_01_001_04	Household ID	_06_01_001_04	Household ID	_06_01_001_04	Household ID	_06_01_001_04	Household ID	_0
Unique Individual ID (Region Code + Site Code + HH Code + Ind Code)	Individual ID	_06_01_001_05	Individual ID	_06_01_001_05	Individual ID	_06_01_001_05	Individual ID	_06_01_001_05	Individual ID	_0
Prompt: This first section is for the enumerator to respond (not a question)	Ho an'ny mpanadiady manokana ity section 1 ity	_06_01_002	Ho an'ny mpanadiady r	_06_01_002	Ho an'ny mpanadiad	_06_01_002	Ho an'ny mpanadiady	_06_01_002	Ho an'ny mpanadiac	_0
Enumerator Name	Mpanadiady	_06_01_003	Mpanadiady	_06_01_003	Mpanadiady	_06_01_003	Mpanadiady	_06_01_003	Mpanadiady	_0
District	Disctrict	_06_01_004	Disctrict	_06_01_004	Disctrict	_06_01_004	Disctrict	_06_01_004	Disctrict	_0
Village Name (Mananjary District)	Village Mananjary	_06_01_005	Village Mananjary	_06_01_005	Village Mananjary	_06_01_005	Village Mananjary	_06_01_005	Village Mananjary	_0
Village Name (Toliara II District)	Village Toliara II	_06_01_006	Village Toliara II	_06_01_006	Village Toliara II	_06_01_006	Village Toliara II	_06_01_006	Village Toliara II	_0
Village Name (Morombe District)	Village Morombe	_06_01_007	Village Morombe	_06_01_007	Village Morombe	_06_01_007	Village Morombe	_06_01_007	Village Morombe	_0
Village (Amoron'i Mania Faritra)	Village Amoron'i Mania	_06_01_008	Village Amoron'i Mania	_06_01_008	Village Amoron'i Ma	_06_01_008	Village Amoron'i Ma	_06_01_008	Village Amoron`i Ma	_0
Enumerator: Are you ready to begin the survey?	Vonona hanomboka?	_06_01_009	Vonona hanomboka?	_06_01_009	Vonona hanomboka?	_06_01_009	Vonona hanomboka?	_06_01_009	Vonona hanomboka`	_0
Enumerator: Does the interviewee agree to do the questionnaire and do they understand the questionnaire?	Manaiky ny hamaly ny fanontaniana ve ireo olona voakasik'ialy fanadiadiana?	_06_01_010	Manaiky ny hamaly ny	_06_01_010	Manaiky ny hamaly	_06_01_010	Manaiky ny hamaly r	_06_01_010	Manaiky ny hamaly	_0
Text of summary of consent/assent form to remind enumerators in case there are questions	Text of summary of consent/assent form to remind enumerators in case there are questions (See to the right)	_06_01_0121	Fanazavana mikasika il	_06_01_0121	Fanazavana mikasika	_06_01_0121	Fanazavana mikasika	_06_01_0121	Fanazavana mikasik	_0
Section 01: ID and gender	**Section 01: ID des individus et genre**	_06_01_0131	**Section 02: ID des indiv**	_06_01_0131	**Section 02: ID des in**	_06_01_0131	**Section 02: ID des in**	_06_01_0131	**Section 1.1: ID des i**	_0
Gender	Genre	_06_01_013	Genre	_06_01_013	Genre	_06_01_013	Genre	_06_01_013	Genre	_0
Harvard Household ID	Harvard Menage ID (###)	_06_01_014	Harvard Menage ID (##	_06_01_014	Harvard Menage ID	_06_01_014	Harvard Menage ID (	_06_01_014	Harvard Menage ID	_0
Harvard Individual	Harvard Individu ID (##)	_06_01_015	Harvard Individu ID (##	_06_01_015	Harvard Individu ID (	_06_01_015	Harvard Individu ID (a	_06_01_015	Harvard Individu ID	_0
DOB: Year	Daty nahaterahana: Taona (####)	_06_01_016	Daty nahaterahana: Ta	_06_01_016	Daty nahaterahana:	_06_01_016	Daty nahaterahana: `	_06_01_016	Daty nahaterahana:	_0
DOB: Month	Daty nahaterahana: Volana	_06_01_017	Daty nahaterahana: Vo	_06_01_017	Daty nahaterahana:	_06_01_017	Daty nahaterahana: '	_06_01_017	Daty nahaterahana:	_0
DOB: Day	Daty nahaterahana: Andro (##)	_06_01_018	Daty nahaterahana: An	_06_01_018	Daty nahaterahana:	_06_01_018	Daty nahaterahana: ,	_06_01_018	Daty nahaterahana:	_0
Actual or approximate DOB?	TENA Taona na VERS Taona?	_06_01_0181	TENA Taona na VERS T	_06_01_0181	TENA Taona na VERS	_06_01_0181	TENA Taona na VERS	Taona?		
Calculated age (2017 minus year of birth)	Taonan'ilay olona	_06_01_019	Taonan'ilay olona	_06_01_019	Taonan'ilay olona	_06_01_019	Taonan'ilay olona	_06_01_019	Taonan'ilay olona	_0
Section 02: Anthropometry and direct measures (recorded in individuals survey for R3V only)	**Section 02: Lanja: Anthropometry and direct measures (R3V only)**	NA	NA	NA	NA	_06_02_001	**Section 02: Antropon**	NA	NA	NA
Height (cm)	Taille (##.# or ###.#)	NA	NA	NA	NA	_06_02_002	Taille (##.# or ###.#)	NA	NA	NA
Weight (kg)	Poids (#.# or ##.# or ###.#)	NA	NA	NA	NA	_06_02_003	Poids (#.# or ##.# or #	NA	NA	NA
MUAC (cm)	MUAC (##.#)	NA	NA	NA	NA	_06_02_004	MUAC (##.#)	NA	NA	NA
Cranial circumference (cm)	Circonference de la tete (CDLT) (##.#)	NA	NA	NA	NA	_06_02_005	Circonference de la te	NA	NA	NA
Section 03: Development and disability	**Section 03: Développement et handicap**	_06_023_001	**Section 03: Développen**	_06_023_001	**Section 03: Développ**	_06_023_001	**Section 03: Développ**	_06_02_001	**Section 02: Dévelop**	_0
Does this individual have a physical disability that prevents them from moving like others their own age?	Manana fahasembanana ara-batana ve io olona io izay manjary lasa sakana amin'ny fihetsiny raha hoarina amin'ny ireo olona mitovy taona aminy?	_06_023_002	Manana fahasembanan	_06_023_002	Manana fahasemban	_06_023_002	Manana fahasembana	_06_02_002	Manana fahasembar	_0
Please describe the disability.	Inona ilay fahasembanana ara-batana?	_06_023_003	Inona ilay fahasembana	_06_023_003	Inona ilay fahasemba	_06_023_003	Inona ilay fahasemba	_06_02_003	Inona ilay fahasemb	_0

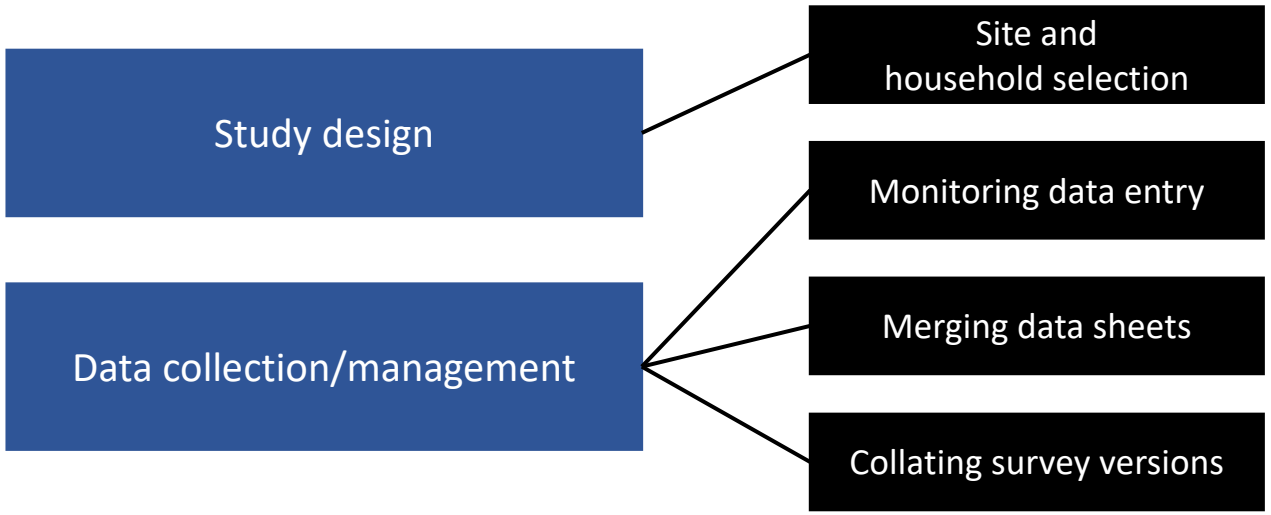
Using R in the MAHERY-CRS Project



Olana be vata: Mila manambatra ‘versions’ efatra

Version 1 + Version 2 + Version 3 + Version 4

Using R in the MAHERY-CRS Project



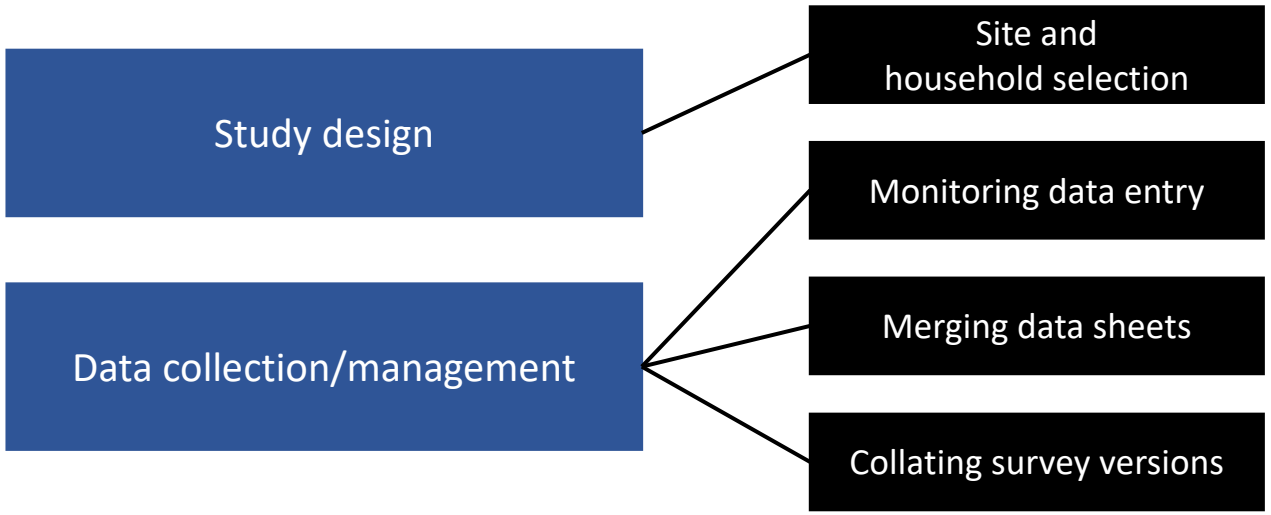
Olana be vata: Mila manambatra ‘versions’ efatra

Version 1 + Version 2 + Version 3 + Version 4

My strategy:

Step 1)

Using R in the MAHERY-CRS Project



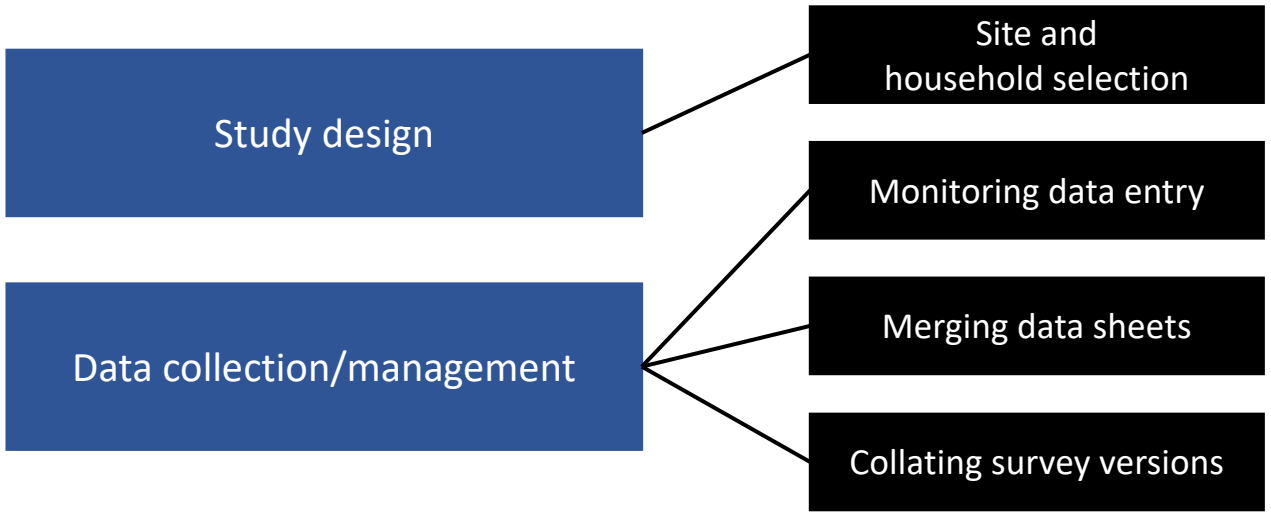
Olana be vata: Mila manambatra 'versions' efatra

Version 1 + Version 2 + Version 3 + Version 4

My strategy:

Step 1) Mangataka fanampiana amin'ny Amy Wesolowski sy Amy Winter ary Jessica Metcalf (satria manan-tsaina be izy telo)

Using R in the MAHERY-CRS Project



Olana be vata: Mila manambatra 'versions' efatra

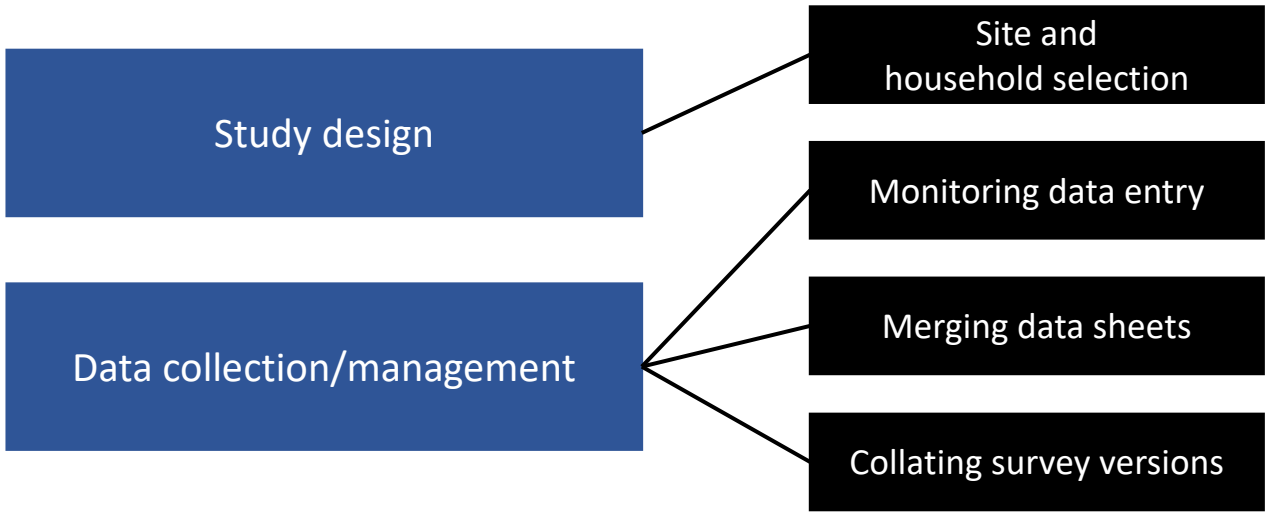
Version 1 + Version 2 + Version 3 + Version 4

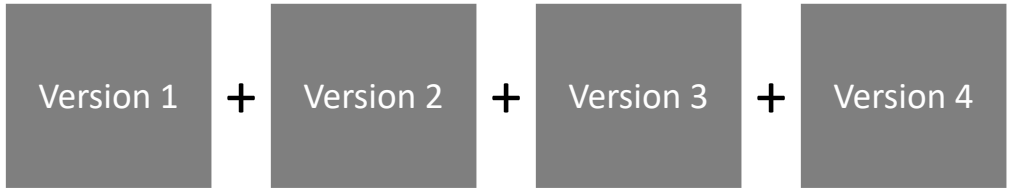
My strategy:

Step 1) Mangataka fanampiana amin'ny Amy Wesolowski sy Amy Winter ary Jessica Metcalf (satria manan-tsaina be izy telo)

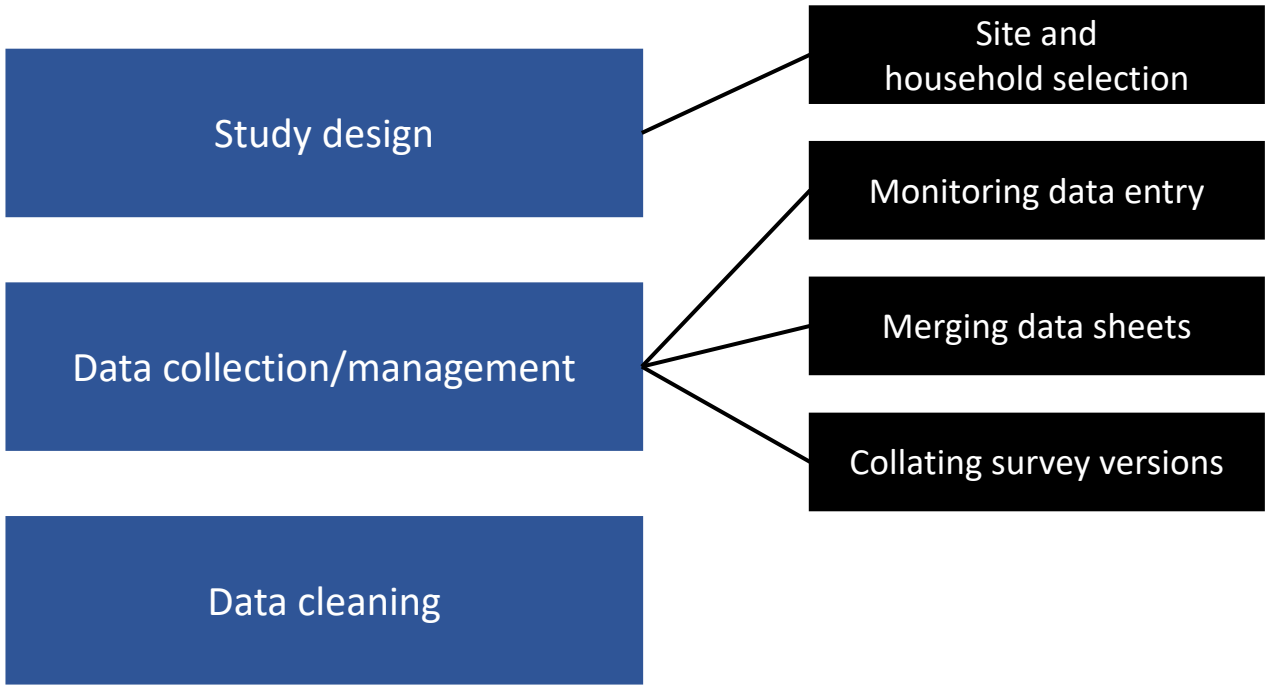
Step 2) "Misaotra betsaka"

Using R in the MAHERY-CRS Project

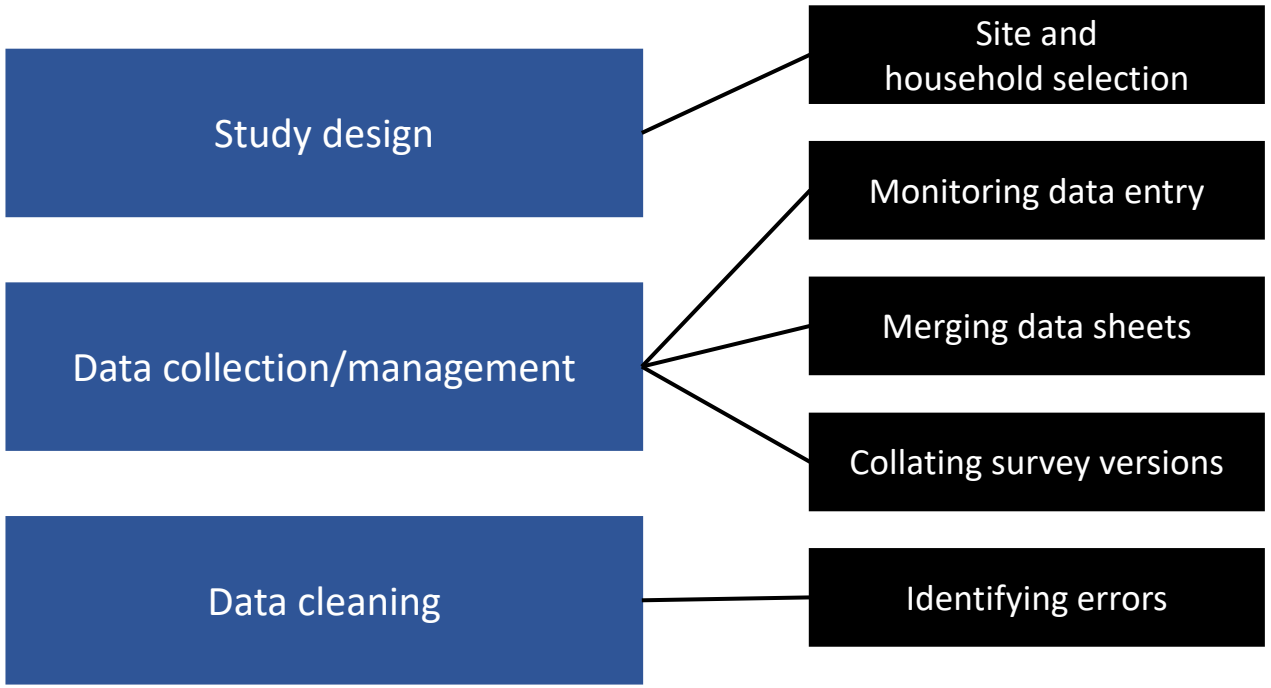




Using R in the MAHERY-CRS Project



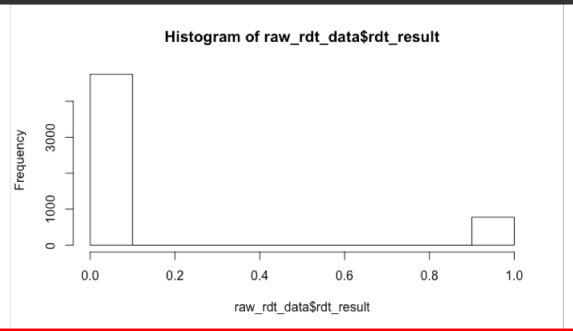
Using R in the MAHERY-CRS Project



Using R in the MAHERY-CRS Project

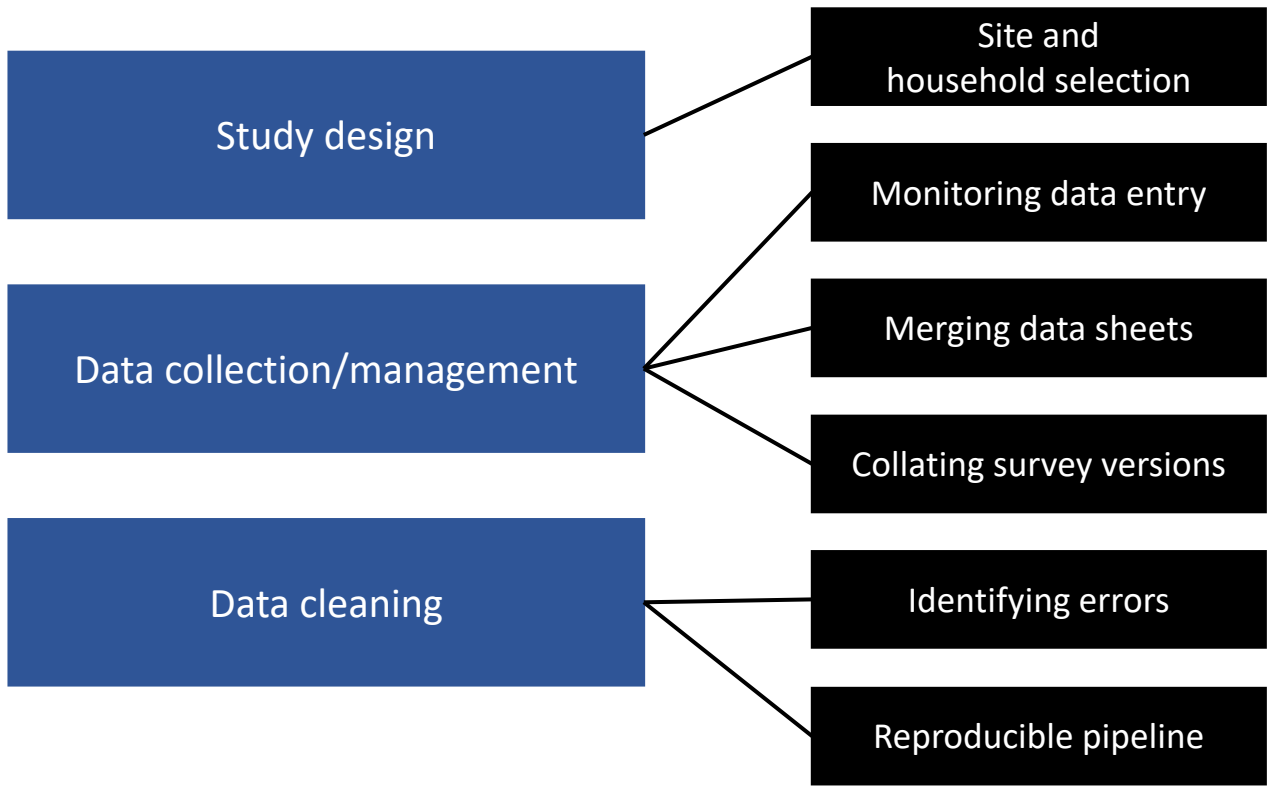
- Study design
  - Site and household selection
- Data collection/management
  - Monitoring data entry
  - Merging data sheets
  - Collating survey versions
- Data cleaning
  - Identifying errors

```
##############################################################################
## Reproducible pipeline
##############################################################################

#Recode rdt_result:
#(1) trim out invalid RDT result individuals
#(2) n = N = negative = 0; pan or panpf = 1

#use subsetting to keep only the individuals with a valid RDT result
raw_rdt_data <- raw_rdt_data[!raw_rdt_data$rdt_result == 'i',]

#Check: use a simple barplot to see if "i"s have been removed
plot_counts <- table(raw_rdt_data$rdt_result)
barplot(plot_counts)

#recode the rdt_results such that they are 1s (positive) or 0s (negatives)
raw_rdt_data$rdt_result <- recode(
  raw_rdt_data$rdt_result, N = 0, n = 0, pan = 1, panpf = 1)

#Check: use a simple barplot to see if the 1, 0 recoding was correct
plot_counts <- table(raw_rdt_data$rdt_result)
barplot(plot_counts)

#from str() we see that rdt results are stored as num even though just 1s and 0s.
#Convert to integer:
raw_rdt_data$rdt_result <- as.integer(raw_rdt_data$rdt_result)

#check with a histogram
hist(raw_rdt_data$rdt_result)
#------------------------------------------------------------------------------
```

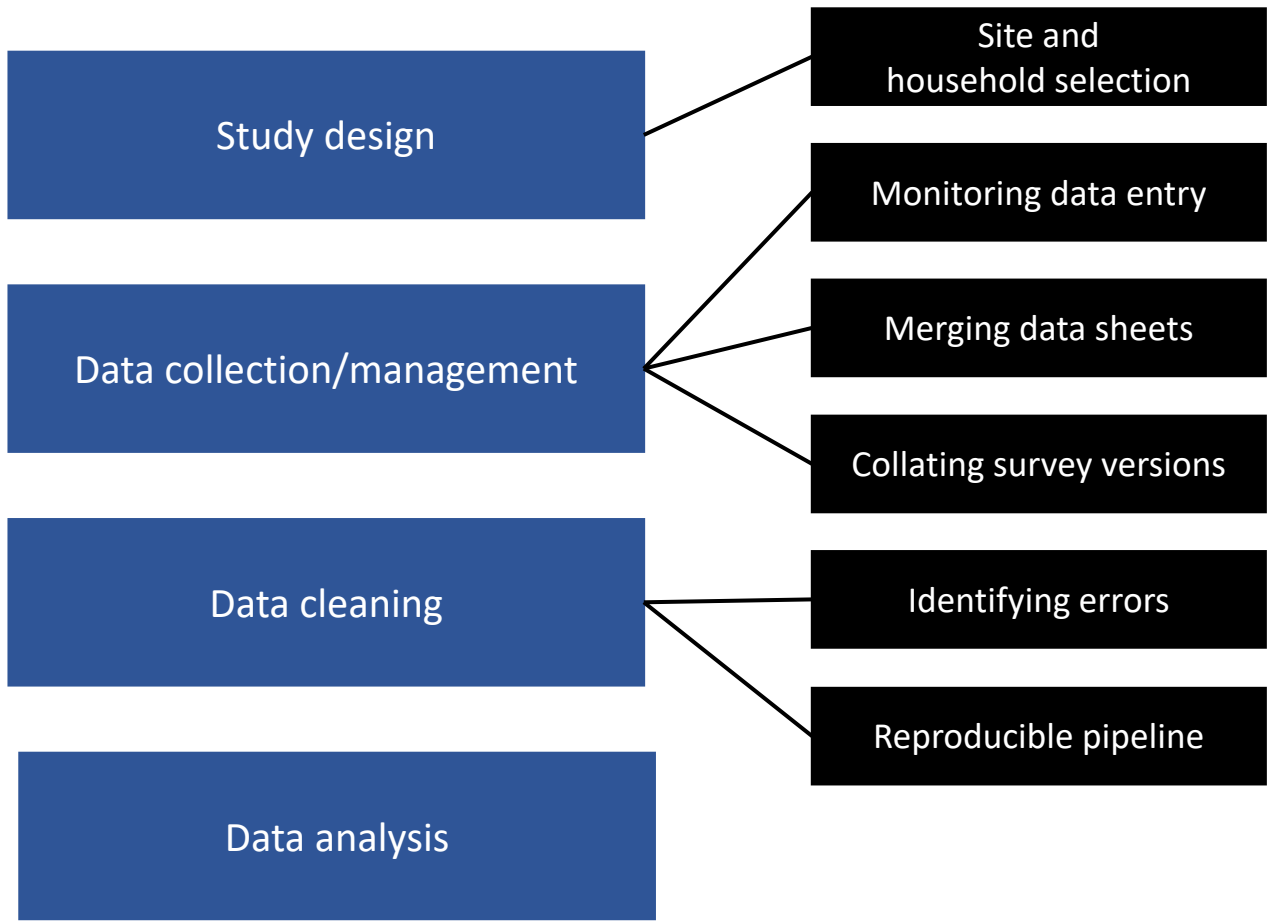


Using histograms to check for incorrect data values

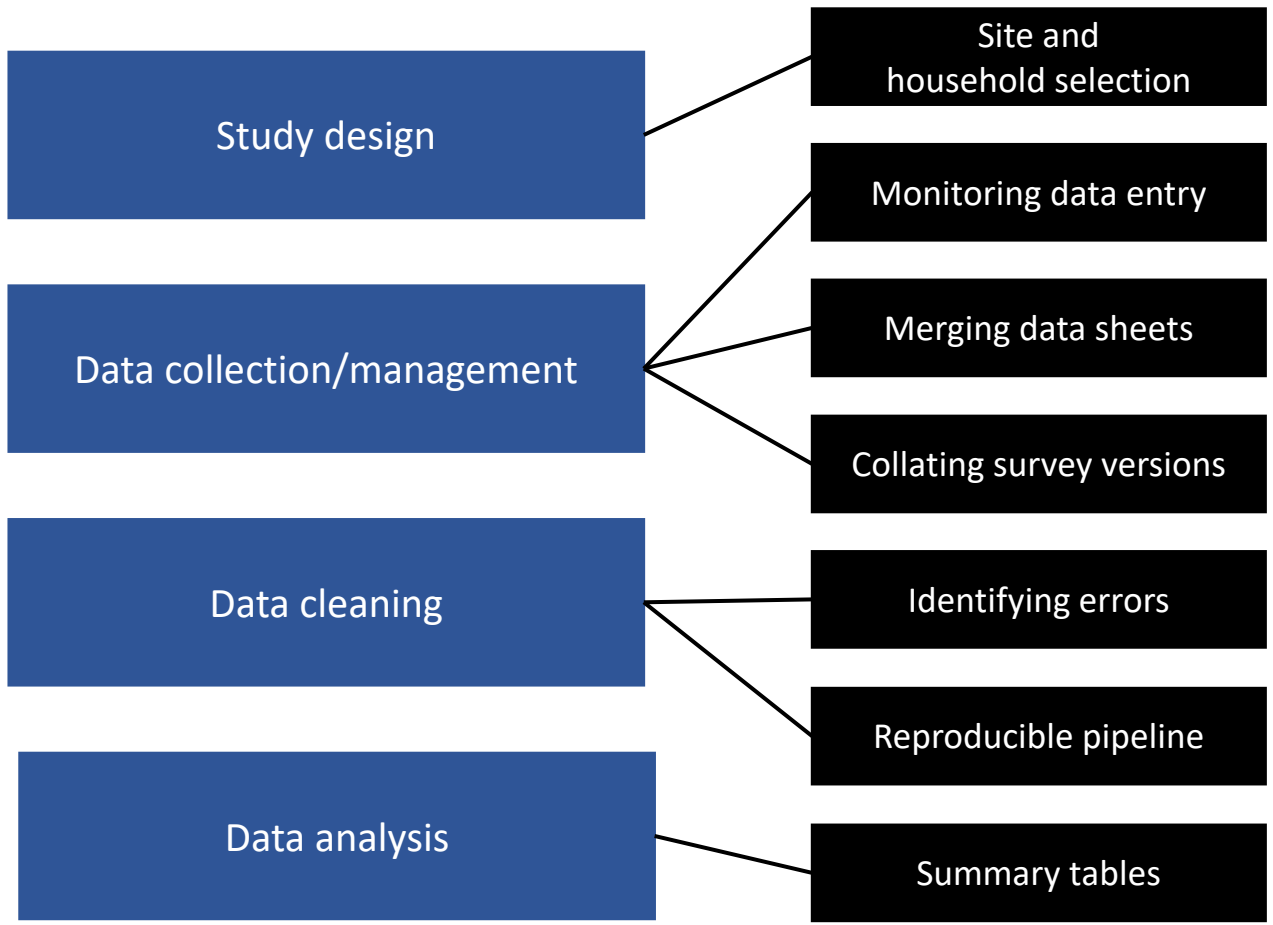
Using R in the MAHERY-CRS Project



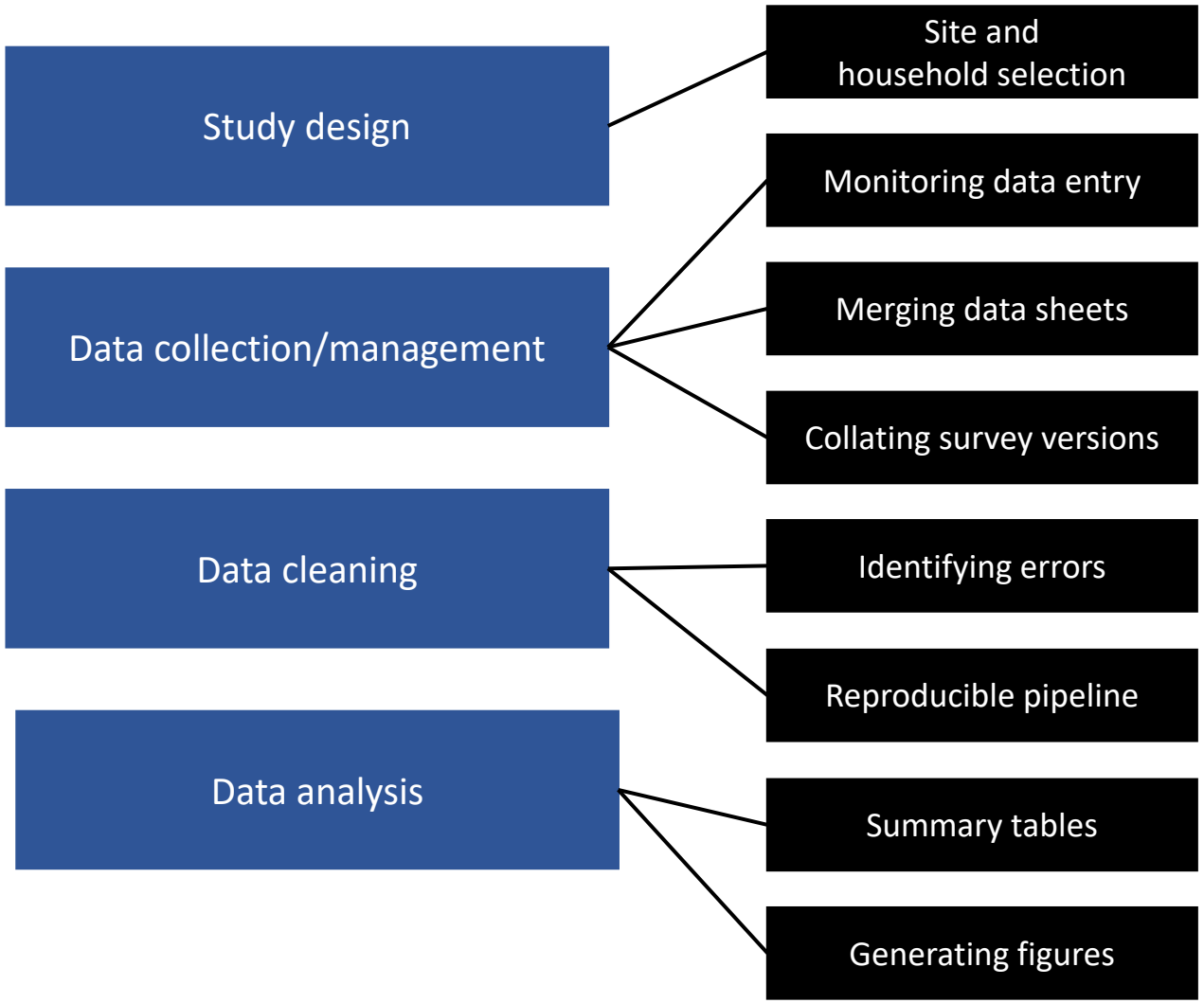
Using R in the MAHERY-CRS Project



Using R in the MAHERY-CRS Project



Using R in the MAHERY-CRS Project



Using R in the MAHERY-CRS Project

- Study design
  - Site and household selection
- Data collection/management
  - Monitoring data entry
  - Merging data sheets
  - Collating survey versions
- Data cleaning
  - Identifying errors
  - Reproducible pipeline
- Data analysis
  - Summary tables
  - Generating figures

Code to generate age-structure figures

```
##############################################################################
## Generate figures: Age structure of malaria infections
##############################################################################

p.region2 <- ggplot(data = melted.plot_age_structure_region2,
                    aes(x=Age_bins,
                        y=ifelse(test = Sex == "Male",
                                 yes = -value, no = value),
                        fill=variable,
                        color=variable,
                        alpha = variable)) +
  geom_bar(stat="identity", position ="identity") +
  coord_flip() +
  scale_colour_manual(values=c("dimgray", "tomato")) +
  scale_fill_manual(values=c("dimgray", "tomato")) +
  scale_alpha_manual(values=c(1, 0.2)) +
  labs(y = "Frequency", x = "Age") +
  scale_y_continuous(labels = abs, limits = 0.25 * c(-1,1))
```

Figure 3

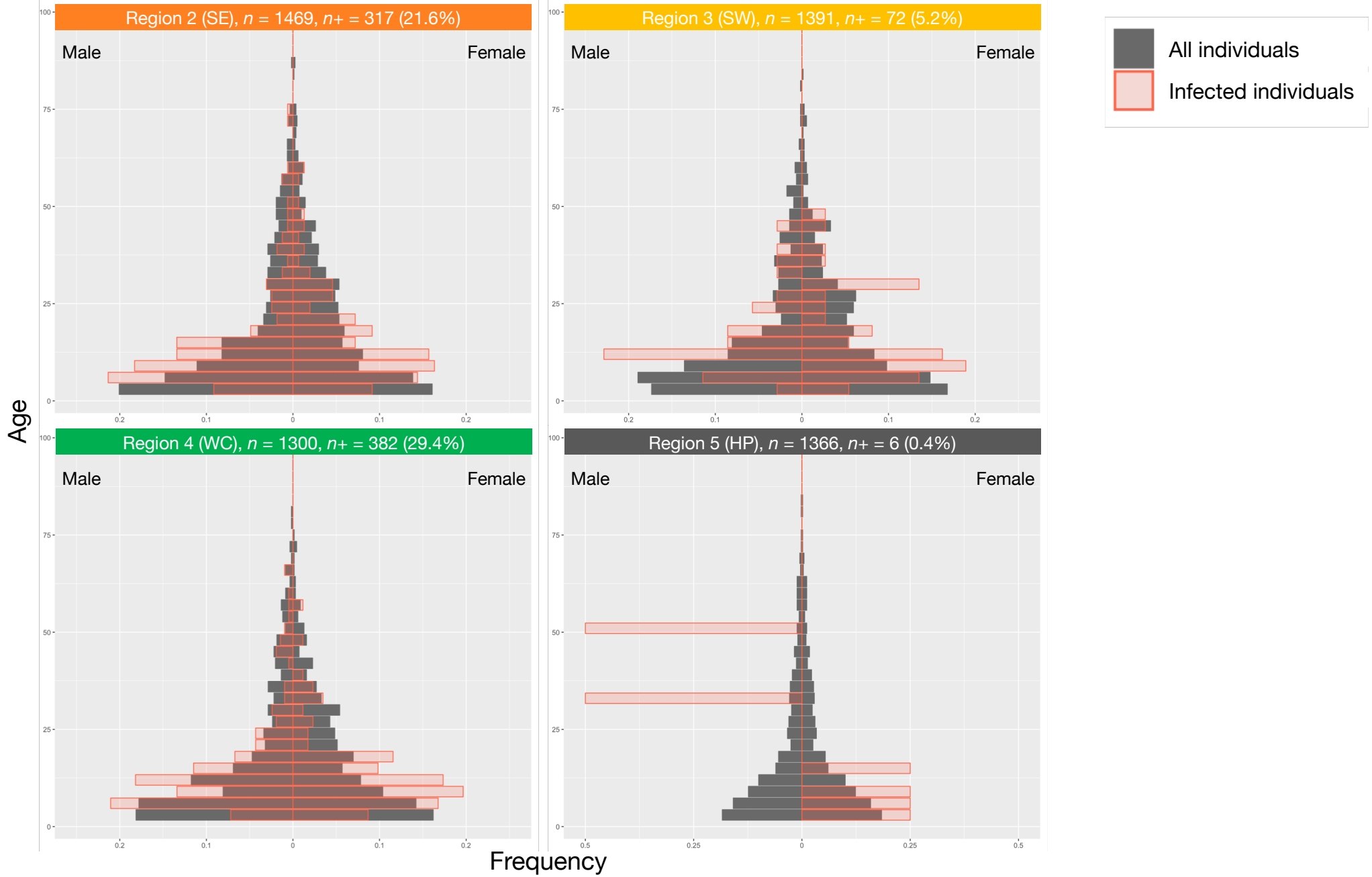
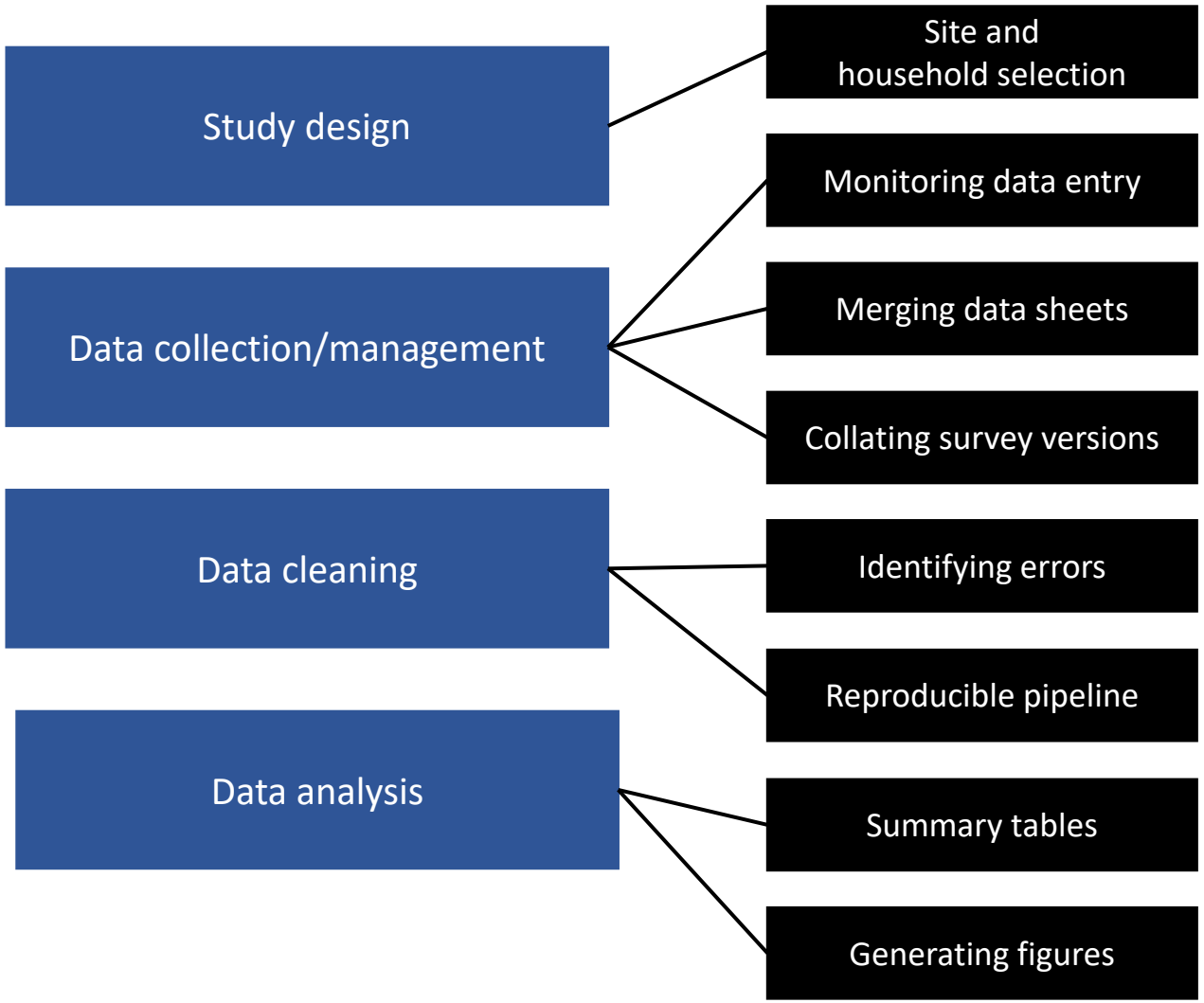
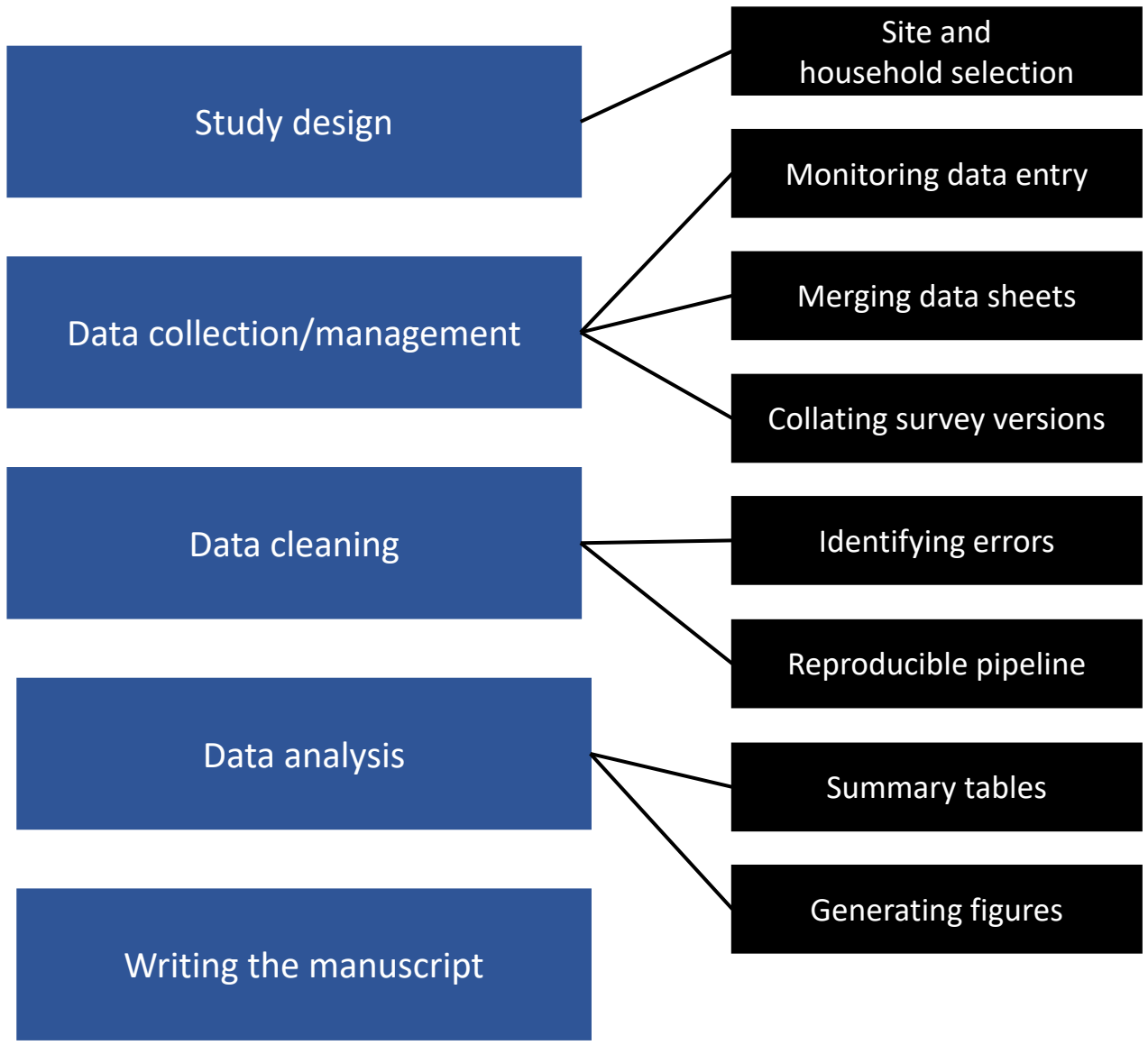


Figure 3: Age distribution of all individuals (gray) and infected individuals (red) shown by region. The proportion of individuals within a 3 year age bin is shown mirrored by sex with males are on the left and females on the right. Sample size, *n*, and the number of individuals positive, *n*+, are shown for each region (southeast, SE; southwest, SW; west coast, WC; high plateau, HP).

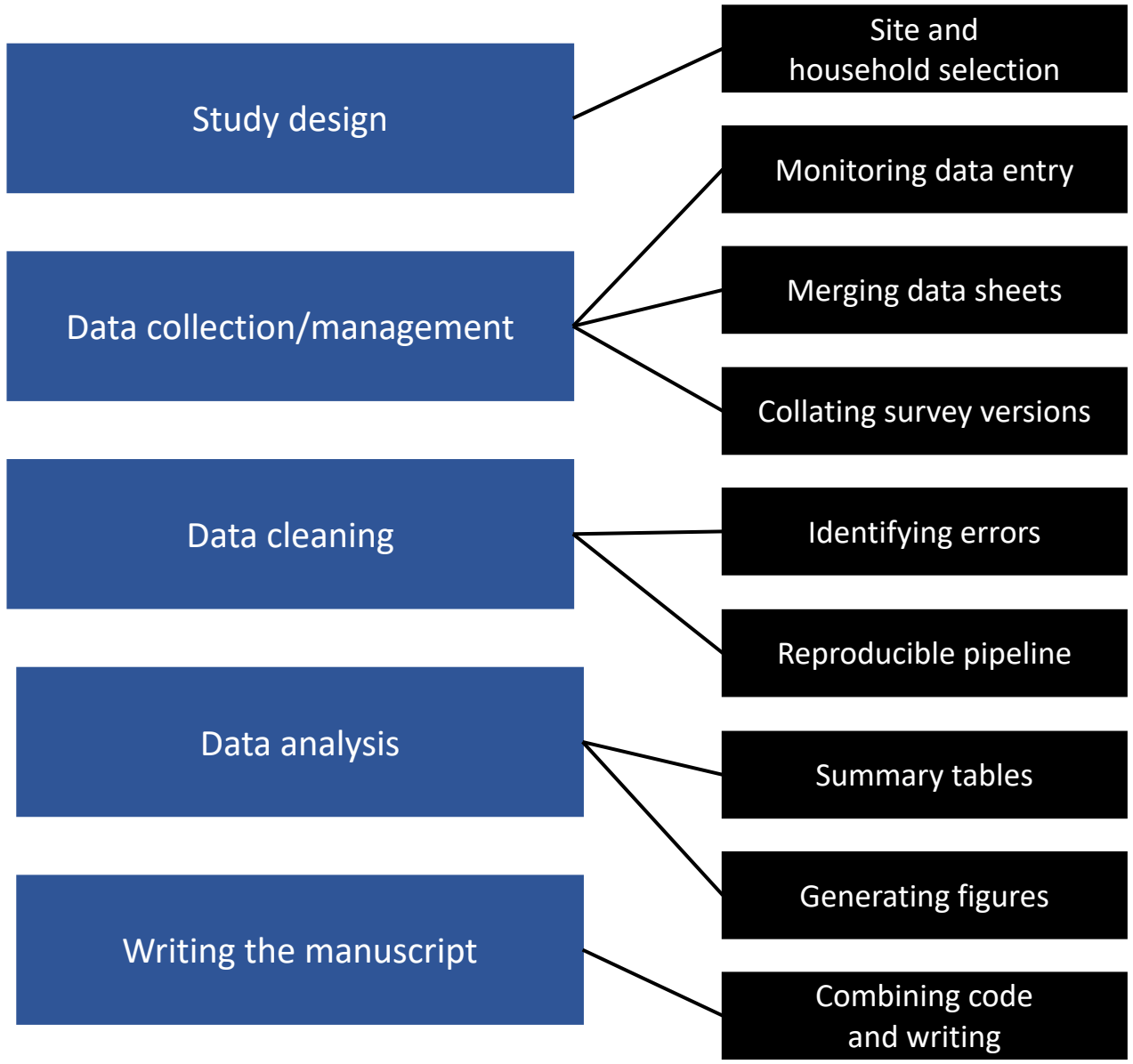
Using R in the MAHERY-CRS Project



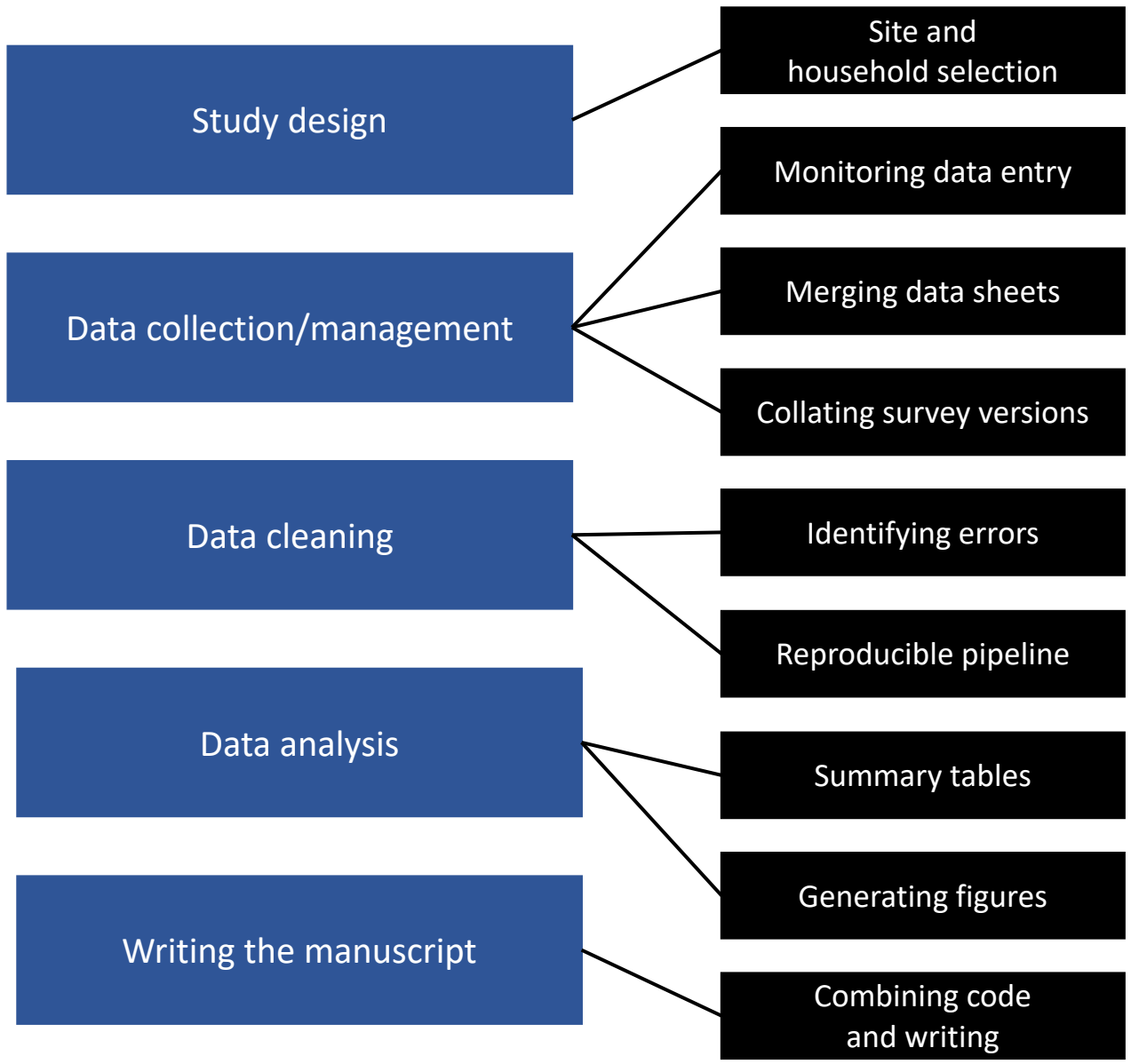
Using R in the MAHERY-CRS Project



Using R in the MAHERY-CRS Project



Using R in the MAHERY-CRS Project



R Markdown
https://rmarkdown.rstudio.com/

How can R help at every stage from day zero to the day of publication?

(*Amin'ny 'study', manomboka andro voalohany hatramin'ny andron'ny 'publication':*

Mora mora kokoa ve raha miasa amin'ny R?)