unicef | for every child

Implementing Causal Machine Learning in

Monday: Potential outcomes & causal effects (experiments, unconfoundedness) & programmes

August / **September** / October 2024

Michael Lechner
Professor of Econometrics | University of St. Gallen | Switzerland

unicef for every child

Lechner
Swiss Journal of Economics and Statistics (2023) 159:8
https://doi.org/10.1186/s41937-023-00113-y

Swiss Journal of
Economics and Statistics

CONFERENCE KEY NOTE Open Access

Causal Machine Learning and its use
for public policy

Michael Lechner[1*]

The workshop series | 2

The Astana Workshop September 30 to October 4, 2024 (10-13, 14:00-17:30)

- **Today**
  - **Morning: Identification with experiments & selection on observables**
  - Afternoon: Discussion of potential programmes to be evaluated
- Tuesday: Causal Machine Learning (theory) (ends at 16:00)
- Wednesday
  - Empirical examples: Active labour market programmes in Flanders
  - The mcf package – how to use it & how to interpret the results
- Thursday: Doing an empirical study in groups with the data introduced in online workshop 4
- Friday: Discussion of programmes to be evaluated continued (core team only)

unicef for every child

Personal introduction

Participants

- Professional background?
- Knowledge in the estimation of causal effects, machine learning, statistics, Python?

Myself

- Professor of Econometrics at the University of St. Gallen
- Co-head of The Swiss Institute for Empirical Economic Research at the University of St. Gallen
- Empirical Economic Research | SEW-HSG | University of St.Gallen (unisg.ch)

  www.michael-lechner.eu
- Research interest in Causal Machine Learning, AI, programme evaluation, …

Plan for today’s workshop

Potential outcome approach with multiple treatments

Definition of causal effects at different aggregation levels

Experiments

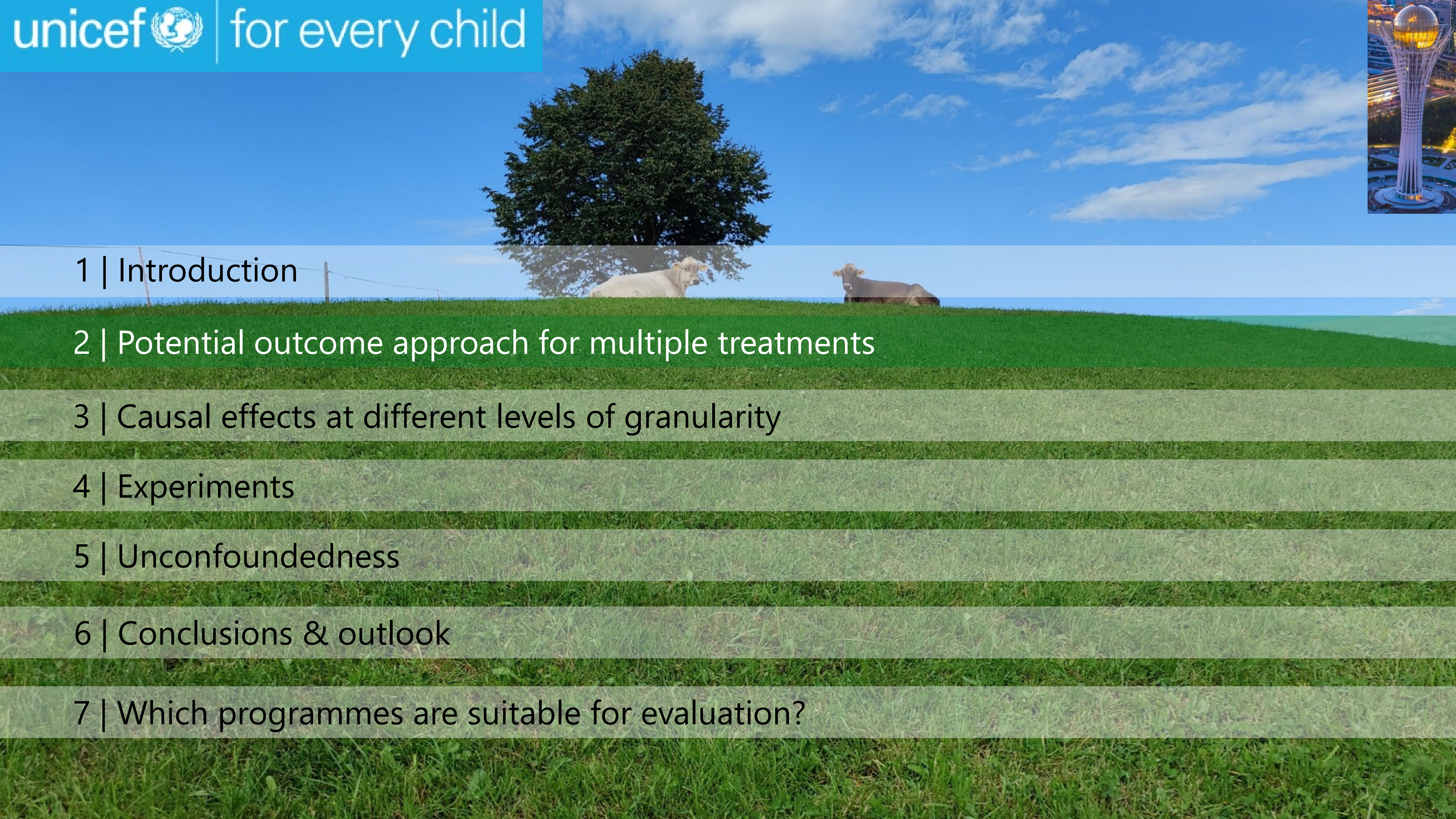
- Key design elements
- Identifying assumptions: Content & possible violations
- Formal proof of identification for the IATE, GATE & ATE

Unconfoundedness

- Key design elements
- Identifying assumptions: Content & possible violations
- Formal proof of identification for the IATE, GATE & ATE

Important note

mcf has been updated to version 0.7.1. While working with 0.6.0 is still ok, it is highly recommended to install 0.7.1 in a Python 3.12 (instead of 3.11 for mcf 0.6.0) environment.



unicef | for every child

1 | Introduction

2 | Potential outcome approach for multiple treatments

3 | Causal effects at different levels of granularity

4 | Experiments

5 | Unconfoundedness

6 | Conclusions & outlook

7 | Which programmes are suitable for evaluation?

Notation of the potential outcome model

Treatment:	D	$(d\text{: } 0, 1, \ldots, M)$	observable
Potential outcome of treatment *d*:	$Y(d)$	*observable if D=d*, unobservable otherwise	
Observed outcome:	Y	$\left(=\sum_{d=0}^{M}\underline{1}(D=d)\,Y(d)\right)$	observable
Other variables:	X, Z	observable	
Data:	$(d_i, y_i, x_i, z_i), \quad i=1,\ldots,N$		

Notation

- Capital letters denote Random Variables (*X*), small letters denote specific values (*x*)
- Capital letters indexed by *i* (X_i) denote the *i*-draw of *X* before it is realised
- Small letters indexed by i (x_i) denote the realisation of the *i*-draw of *X*

1 | Introduction

2 | Potential outcome approach for multiple treatments

3 | Causal effects at different levels of granularity

4 | Experiments

5 | Unconfoundedness

6 | Conclusions & outlook

7 | Which programmes are suitable for evaluation?

General considerations

It is common in the literature to focus on pair-wise comparisons of treatments

- Alternatives that aggregate treatments have been proposed as well but are rarely used

Therefore, & for simplicity, below we stick to the binary treatment model

For the rest of the workshop, we will assume that causal effects are heterogeneous

- They may be different from one unit to the other (usually in an unrestricted way)

Individual treatment effect

$ITE_i = Y_i(1)-Y_i(0)$

This effect is fundamentally unidentifiable

- Observation *i* is only observed in 1 of the 2 treatments involved in this comparison

We can only identify effects for some larger groups for which units may be observed in both treatment states

Effects for different levels of granularity | 1

$D:$ Treatment (0 or 1)

$Y(1):$ Outcome when $D=1$

$Y(0):$ Outcome when $D=0$

$X:$ Confounder & heterogeneity variables

$Z:$ Specific heterogeneity variables (low dim.)

Observable: $X,\ Z,\ Y = DY(1)+(1-D)Y(0)$

Individualized **A**verage **T**reatment **E**ffect

- Average effect of different values of *D* for a unit (individual, firm, …) that has a specific values of many characteristics

Individualized (Conditional) Average Treatment Effects

$$IATE(x) = E(Y(1)-Y(0) \mid X = x)$$

D: Treatment (0 or 1)

Y^1: Outcome when $D = 1$

Y^0: Outcome when $D = 0$

X: Confounder & heterogeneity variables

Z: Specific heterogeneity variables (low dim.)

Observable: $X,\ Z,\ Y = DY^1 + (1 - D)Y^0$

Effects for different levels of granularity | 1

Group **A**verage **T**reatment **E**ffect (of the **T**reated)

- What is the average effect of different values of *D* for a larger group of interest

Group (Conditional) Average Treatment Effects

$$GATE(z) = E(Y(1) - Y(0) \mid Z = z) = E_{X|Z=z} IATE(x)$$

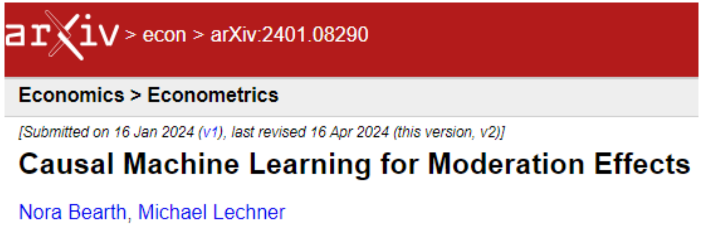
- Related effect

Balanced GATE (BGATE; Bearth, Lechner, 2024)

$$BGATE(z, \tilde{X}) = E_{\tilde{X}} E(Y(1) - Y(0) \mid Z = z, \tilde{X} = \tilde{x}) = E_{\tilde{X}} E_{X|Z=z,\tilde{X}=\tilde{x}} IATE(x)$$

GATE of the Treated

$$GATET(z) = E(Y(1) - Y(0) \mid Z = z, D = 1) = E_{X|Z=z,D=1} IATE(x)$$



arXiv > econ > arXiv:2401.08290

Economics > Econometrics

[Submitted on 16 Jan 2024 (v1), last revised 16 Apr 2024 (this version, v2)]

Causal Machine Learning for Moderation Effects

Nora Bearth, Michael Lechner

unicef for every child

Effects for different levels of granularity | 1

D: Treatment (0 or 1)

$Y(1)$: Outcome when $D = 1$

$Y(0)$: Outcome when $D = 0$

X: Confounder & heterogeneity variables

Z: Specific heterogeneity variables (low dim.)

Observable: $X,\ Z,\ Y = DY(1) + (1 - D)Y(0)$

Average **T**reatment **E**ffect (of the **T**reated)

- What is the average effect of different values of *D* for a population of interest?

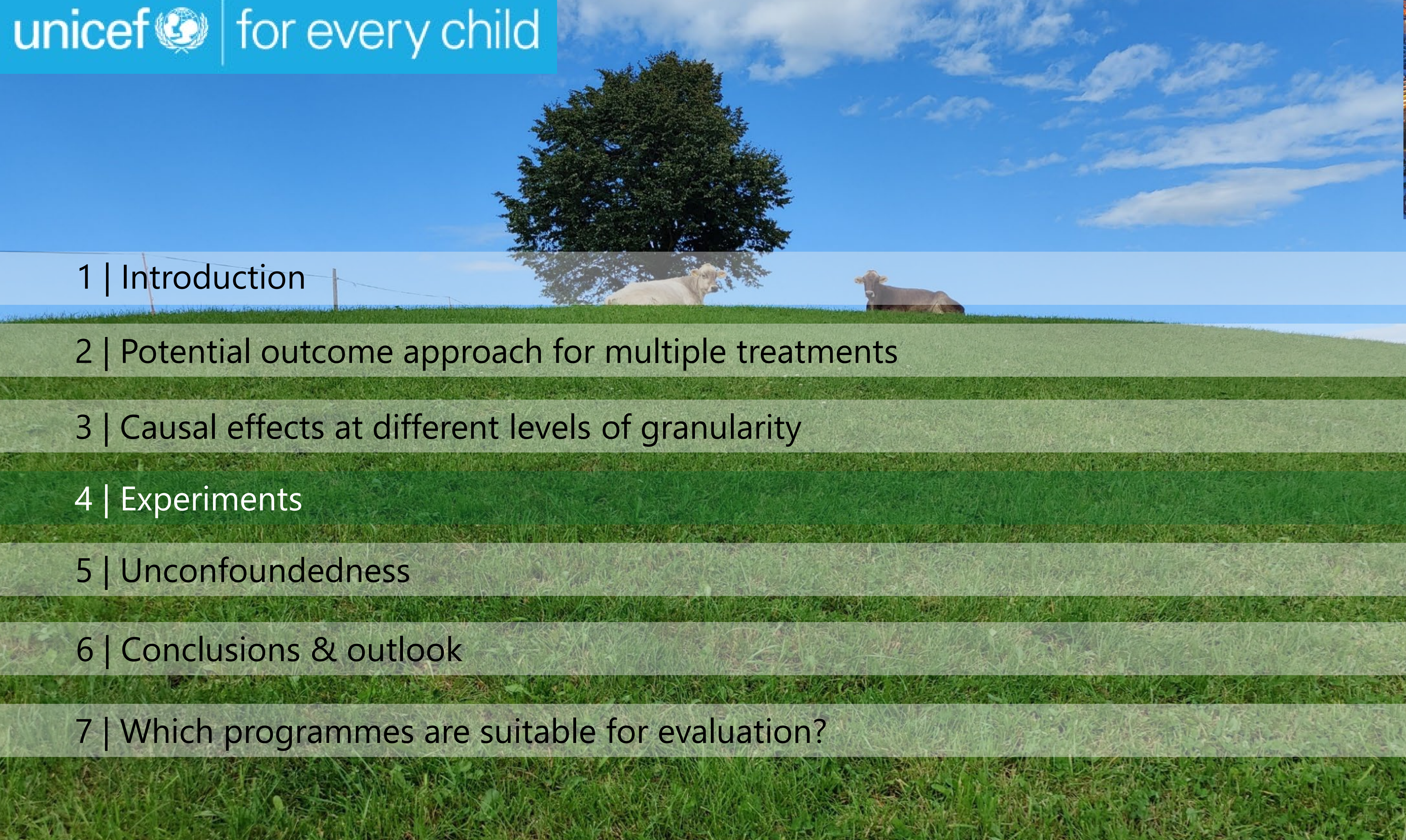
Average Treatment Effects

$$ATE = E(Y(1) - Y(0)) = E_X IATE(x)$$

Average Treatment Effects of the Treated

$$ATE = E(Y(1) - Y(0) \mid D = 1) = E_{X|D=1} IATE(x)$$

unicef | for every child



1 | Introduction

2 | Potential outcome approach for multiple treatments

3 | Causal effects at different levels of granularity

4 | Experiments

5 | Unconfoundedness

6 | Conclusions & outlook

7 | Which programmes are suitable for evaluation?

Introduction

Well-run experiments solve the identification problem convincingly

- The only empirical design in which the researcher has full control of selection / assignment mechanism
- But: SUTVA must hold nevertheless & is not guaranteed by randomization

In the past, experiments have been very rare in economics, now they are far more wide-spread

- Digitalisation will make many experimental designs even cheaper & easier to conduct

Note: Most of this lecture follows closely the Athey-Imbens (2017) chapter in the *Handbook of Field Experiments*.

- This chapter also contains many other useful references & more technical details & derivations.

unicef for every child

Experiments | Identification

Assumption

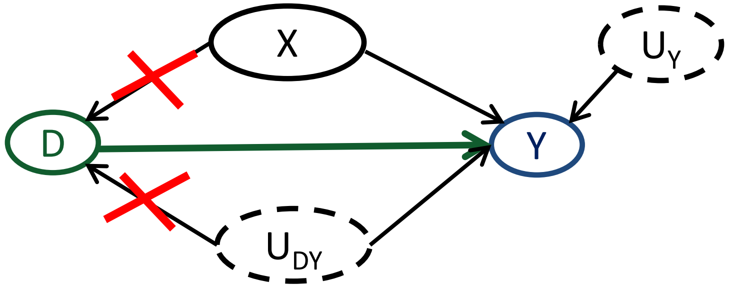
- Treatment is (as good as) randomly assigned

$$Y(0), Y(1) \coprod D \quad \Rightarrow \quad E(Y(d)) = E(Y \mid D = d)$$

- SUTVA
  - Treatment well defined for everybody, treatment assignments of others do not matter for own effect

$$Y = DY(1) + (1 - D)Y(0)$$

Experiments | The corresponding causal graphs



Experiments | Implication of identifying assumption

Any potentially confounding factors have the same distribution for participants & non-participants

- Thus, unadjusted comparisons of the outcomes of participants with the outcomes of non-participants reveal causal effects

Formal proof:

$$IATE(x) = E(Y(1) - Y(0) \mid X = x) =$$
$$= E(Y \mid X = x, D = 1) - E(Y \mid X = x, D = 0)$$

- Since *IATE(x)* is identified, *GATE(z)*, ATE etc. are identified as well

Internal & external validity

Internal validity: Valid causal inference in the study population

- Big advantage of well-conducted experiments

External validity: Generalizability to population of interest

- Is the sample representative for something of interest?
- Most trivial: *Past* data useful for policy advice (which by definition relates to the *future*)?
- Experimental population may be selective
  - If units can choose to participate (usually ‘informed consent’ needed)
  - If some assignment algorithm chooses ‘easy to get’ units
    - E.g. students in behavioural labs
- Sometime external validity can be achieved by reweighting

These issue arises only when effects are heterogeneous

Experiments | Estimation | ATE

Mean outcome of participants – mean outcome of non-participants

- ATE (=ATET)

$$\hat{\gamma} = \frac{1}{N^1}\sum_{i=1}^{N} d_i y_i - \frac{1}{N^0}\sum_{i=1}^{N}(1-d_i)y_i$$

$$Var(\hat{\gamma}) = Var\left[\frac{1}{N^1}\sum_{i=1}^{N} d_i y_i\right] + Var\left[\frac{1}{N^0}\sum_{i=1}^{N}(1-d_i)y_i\right]$$

$$= \frac{1}{N^1} Var(Y \mid D=1) + \frac{1}{N^0} Var(Y \mid D=0)$$

$$\hat{V}ar(\hat{\gamma}) = \frac{1}{N^1}\hat{V}ar(Y \mid D=1) + \frac{1}{N^0}\hat{V}ar(Y \mid D=0)$$

Experiments | Estimation | GATE

Discrete *Z*

- Stratify w.r.t. values of *Z*
- Same estimator as for ATE within strata

Continuous Z & IATE

- Stratification does not work
- Use Causal Machine Learning Methods to be discussed tomorrow
- Note: All estimators that work for *unconfoundedness*, also work for *experiments*

Experiments | Advantages & disadvantages in practise

Advantages

- If experiments were implemented cleanly, causal effects have very high credibility
- Usual causal parameters of interest are identified

Disadvantages

- Not every variation in *D* can be generated by a reasonable experiment
- Experiments may be expensive
- … may take long (waiting time from starting treatment until measuring outcomes)
- … may be messed up by administrators or subjects not obeying the rules
- Results may be not externally valid
  - Internal vs external validity

1 | Introduction

2 | Potential outcome approach for multiple treatments

3 | Causal effects at different levels of granularity

4 | Experiments

5 | Unconfoundedness

6 | Conclusions & outlook

7 | Which programmes are suitable for evaluation?

Identifying assumptions

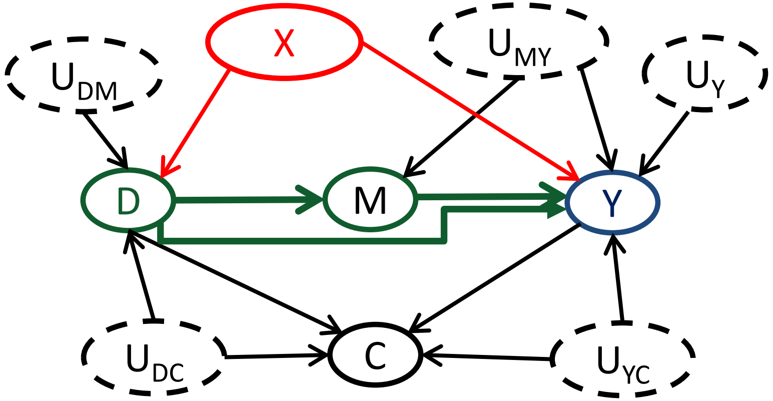
1) Potential outcomes are conditionally independent of treatment for any given values of the confounding variables (***CIA***, *no confounding on observables, unconfoundedness,…, stratified experiment)*

2) For any given value of the confounding variables, a unit could potentially be observed with *D=1* or *D=0* (***common support***)

3) The confounding variables are not influenced by the treatment in a way that is related to the outcome variables (***exogeneity*** of confounders)

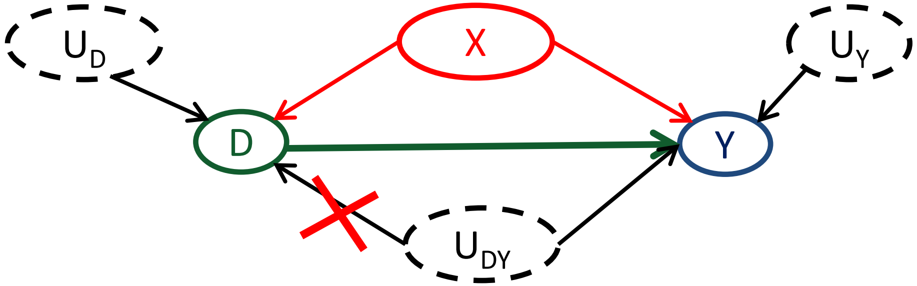
4) The observed outcomes in one treatment state correspond to the potential outcomes of that treatment state for the participants in that state (stable unit treatment value assumption, ***SUTVA***)

The corresponding causal graphs



Causal effect of interest: D → Y
D: Treatment
Y: Outcome
M: Mediator
C: Collider
X: Confounder

Simplified version



Identifying assumptions | CIA

Potential outcomes are conditionally independent of treatment for given values of the confounding variables

$$Y(0), Y(1) \coprod D \mid X = x; \quad \forall x \in \chi$$

This assumption defines the control variables (*X*) required: If all variables that ***jointly*** influence *potential* outcomes (*Y(d)*) & selection (*D*) are observed, CIA must hold. If such variables are missing, it is unlikely to hold.

This needs to be true only for the population of interest (in terms of χ).

Data hungry identification strategy.

Researcher must know & decide which control variables are needed.

unicef for every child

Identifying assumptions | Common support

For any given value of the confounding variables, a unit could potentially be observed with *D=1* or *D=0* (***common support***).

$$0 < P(D=1 \mid X=x) < 1, \quad \forall x \in \chi$$

Since identification is based on comparing units with the same *X* & different *D*, common support ensures that such units exist

- NB: For the estimation of mean effects, comparing units with the same values of *P(D=1|X=x)* instead of *X* is sufficient

unicef | for every child

Identifying assumptions | Exogeneity | 1

The confounding variables are not influenced by the treatment in a way that is related to the outcome variables (***exogeneity*** of confounders)

- Define effect of *D* on *X* similarly to *D* → *Y*:

$$X(1), X(0); \qquad X = DX(1) + (1-D)X(0) \qquad\qquad D \rightarrow X: \quad X(1) \neq X(0)$$

A *sufficient* condition for exogeneity is: $X(1) = X(0)$

Example: If *Y* is used as conditioning variable in CIA, all consistent estimators converge to 0 (whatever the real effect is)

Problem is due to conditioning on part of the effect, thus reducing the 'remaining' effect

Identifying assumptions | SUTVA

The observed outcomes in one treatment state correspond to the potential outcomes of that treatment state for the participants in that state (stable unit treatment value assumption, ***SUTVA,*** *observation rule, consistency condition*):

$$Y = DY(1) + (1 - D)Y(0)$$

This condition requires that there are …

- no unrepresented treatments in the population of interest (everybody is either 0 or 1)
- no relevant interactions between treatments
  - The fact that '*i*' participates does not change the potential outcome of '*j*' $Y(d_i)\ vs.\ Y(d_1,...,d_N)$
  - Example: Large ALMP programmes may change demand and supply relations and thus wages ➔ **non**participation labour market outcome with and without a large programme may differ; "*no general equilibrium effects*")

Unconfoundedness | Implication of identifying assumption

Potential confounders may be distributed differently for participants & non-participants

Thus, unadjusted comparisons of outcomes of participants & outcomes of non-participants do not reveal causal effects

Experiment-like comparisons for units **with the same values of *X*** are causally valid

Proof of identification of *IATE(x)*

Same as for experiments

- Data as good as coming from a stratified experiment

$$IATE(x) = E(Y(1) - Y(0) \mid X = x) =$$
$$= E(Y \mid X = x, D = 1) - E(Y \mid X = x, D = 0)$$

Thus, ATEs & GATEs are identified as well

However, this proof is only partially instructive for estimation

- Next, we look at more instructive proof

unicef for every child

Identification proofs for ATE | Conditional exp. of outcome

$$ATE = E(Y(1)) - E(Y(0))$$

$$E(Y(1)) = E_X E(Y(1) \mid X = x) \underset{CIA}{=} E_X E(Y(1) \mid X = x, D = 1) \underset{SUTVA}{=} E_X E(Y \mid X = x, D = 1) = E_X \mu(1, x)$$

$$E(Y(0)) = E_X E(Y(0) \mid X = x) \underset{CIA}{=} E_X E(Y(0) \mid X = x, D = 1) \underset{SUTVA}{=} E_X E(Y \mid X = x, D = 0) = E_X \mu(0, x)$$

$$ATE = E_X \left[\mu(1, x) - \mu(0, x) \right]$$

These identification results suggests to base estimation on consistent estimators of the conditional-on-*X* expectations in subsamples by *D* (**outcome regressions** & matching estimators)

unicef for every child

Identification proofs | Weighted outcomes

$$ATE = E(Y(1)) - E(Y(0))$$

$$E(Y(1)) = E_X \mu(1,x) = \ldots = E\left(\frac{DY}{p(x)}\right)$$

$$E(Y(0)) = E_X \mu(0,x) = \ldots = E\left(\frac{(1-D)Y}{1-p(x)}\right)$$

$$ATE = E\left(\frac{DY}{p(x)}\right) - E\left(\frac{(1-D)Y}{1-p(x)}\right)$$

These identification results suggest to use consistent estimators of the propensity score, *p(x)*, to obtain weighted averages of the outcomes (***Inverse Probability Weighting***)

unicef for every child

Identification proofs | Double robustness | 1

The previous results can be combined

$$ATE = E(Y(1)) - E(Y(0))$$

$$E(Y(1)) = E\mu(1,x) + E\left[\frac{\left(Y-\mu(1,x)\right)D}{p(x)}\right] \qquad E(Y \mid X = x, D = 1) = \mu(1,x)$$

$$E(Y(0)) = E\mu(0,x)E\left[\frac{\left(Y-\mu(0,x)\right)(1-D)}{\left(1-p(x)\right)}\right] \qquad E(Y \mid X = x, D = 0) = \mu(0,x)$$

$$ATE = E\left[\mu(1,x)-\mu(0,x)\right] + E\left[\frac{\left(Y-\mu(1,x)\right)D}{p(x)} - \frac{\left(Y-\mu(0,x)\right)(1-D)}{\left(1-p(x)\right)}\right]$$

Based on average influence / efficient score functions (Hahn, 1998, p. 328)

These functions suggest to base estimators on consistent estimators of conditional outcome expectations **&** propensity scores ➔ **Important for CML**

- Estimators remain consistent even if *p(x)* **or** $\mu(\mathrm{d},x)$ is completely misspecified

Unconfoundedness | Estimation | 1

All estimators must do the following (implicitly or explicitly)

- Estimate causal effects for all different observed values of *X* → aggregate them to obtain ATE
  - Matching estimators do this explicitly
- Estimate weights that would make the distribution of the confounders among treated & non-treated identical → use these weights for weighted mean comparison of the outcomes of treated & non-treated
  - Special case 1: Methods that remove the effects of other variables (*X*) (e.g. linear regressions) $Y = D\alpha + X\beta$
  - Special case 2: (1) Weight outcomes of treated by estimated *P(D=1|X)* [=: propensity score]

    (2) Weight outcomes of non-treated by *1-P(D=1|X)*

    (3) Mean of (1) minus mean of (2)

unicef for every child

Unconfoundedness | Estimation | 2

The value of Causal Machine Learning

- Avoid additional assumptions in the estimation steps (e.g., for regressions, propensity scores)
  - As would be required by classical regression-type & weighting type estimators
- More powerful in estimating heterogeneities

unicef for every child

Unconfoundedness | Advantages & disadvantages in practise

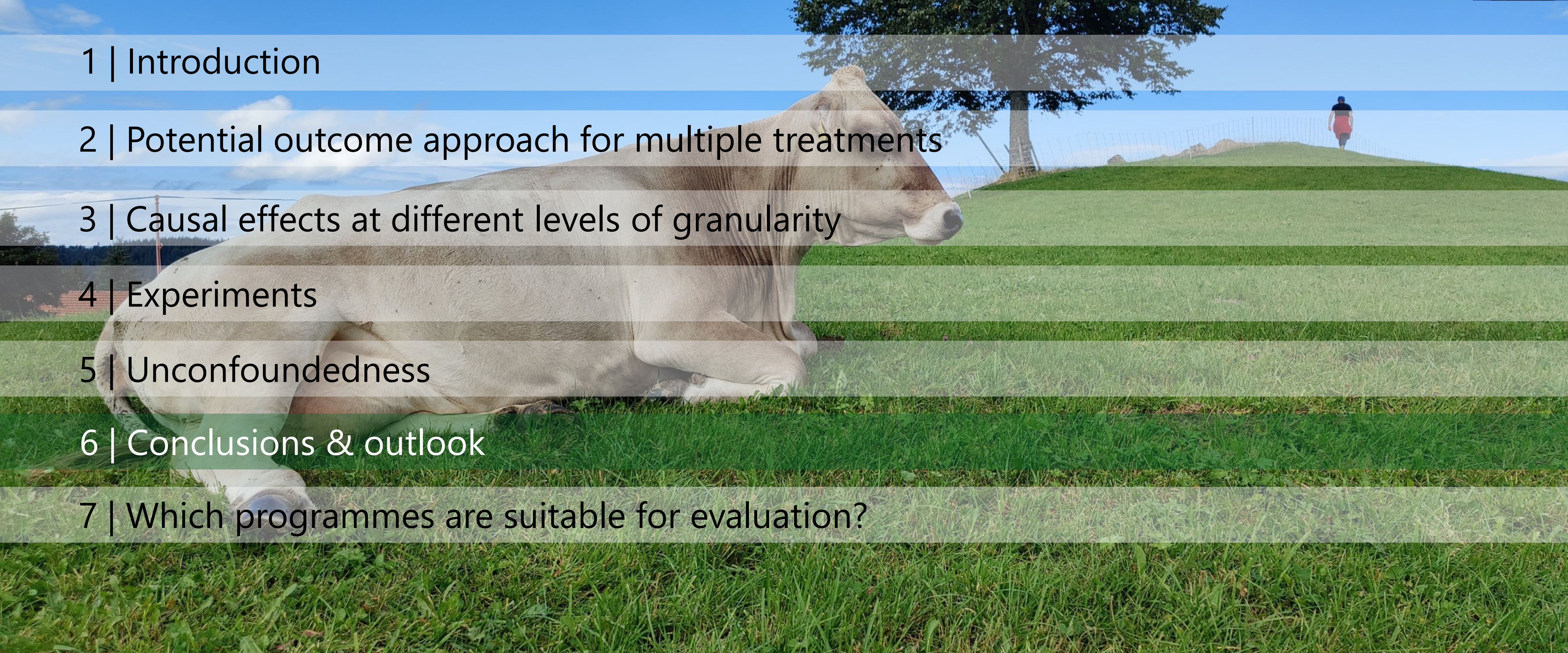
Advantages

- Credibility could be high
- Usual causal parameters of interest are identified

Disadvantages

- Substantial knowledge about assignment process is needed to identify relevant confounders
- Data hungry strategy (many features - *X*)
- More fancy estimators needed to perform confounder adjustments
  - Loss of precision compared to experiments → more observations needed (*N* larger)

unicef | for every child



1 | Introduction

2 | Potential outcome approach for multiple treatments

3 | Causal effects at different levels of granularity

4 | Experiments

5 | Unconfoundedness

6 | Conclusions & outlook

7 | Which programmes are suitable for evaluation?

Experiments

Experiments are the most credible research designs

- Implement whenever possible & reasonable

Estimation

- ATE estimation is just a mean comparison
- Estimation & use of effect heterogeneity requires CML

Unconfoundedness

Unconfoundedness (selection-on-of-observables) could be a credible research design

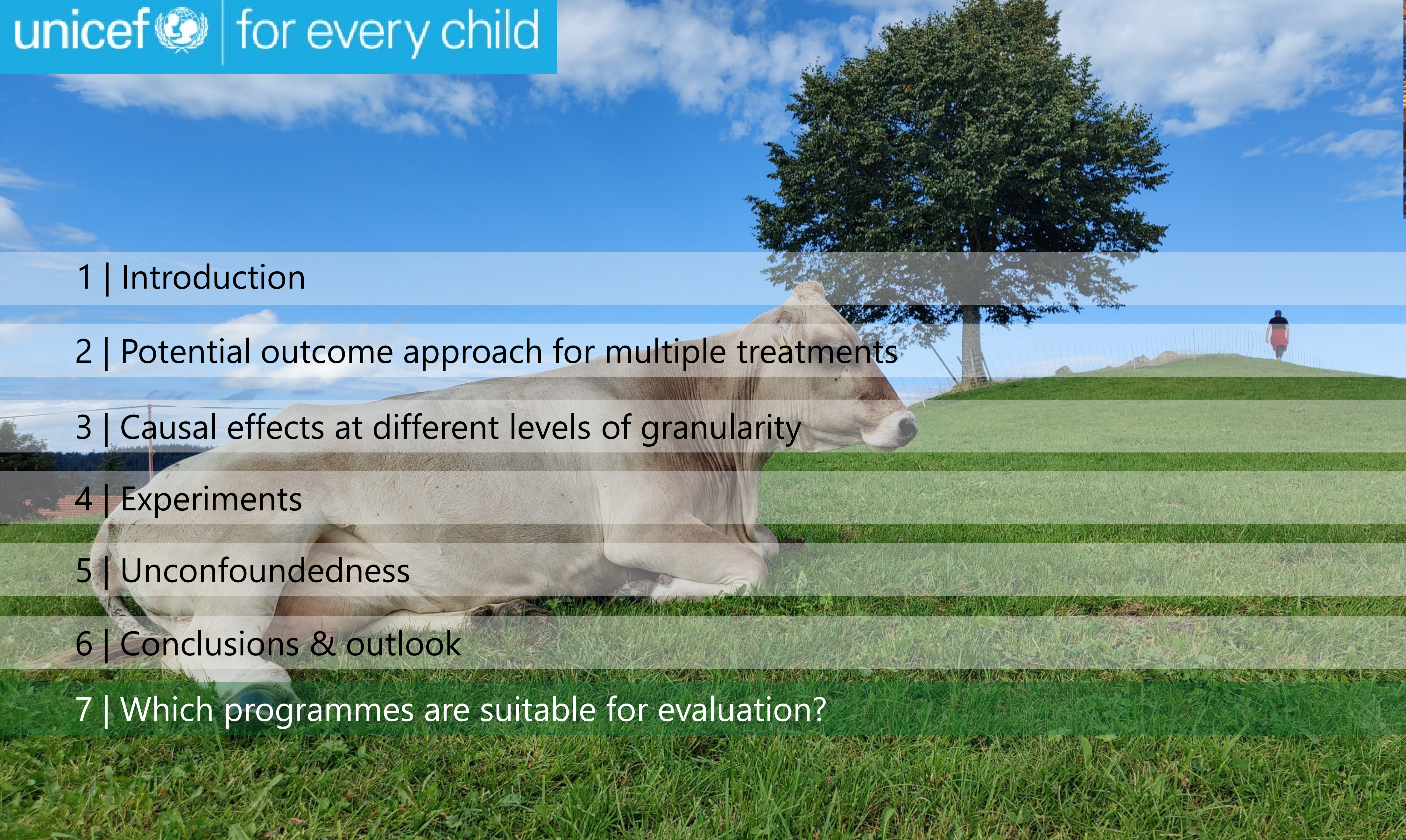
Credibility requires …

- institutional knowledge (in particular of assignment process)
- Informative data to be able to account for confounding

Estimation

- CML has advantages for all effects (ATE, GATE, IATE, …)
- Estimation & use of effect heterogeneity requires CML

unicef | for every child



1 | Introduction

2 | Potential outcome approach for multiple treatments

3 | Causal effects at different levels of granularity

4 | Experiments

5 | Unconfoundedness

6 | Conclusions & outlook

7 | Which programmes are suitable for evaluation?

unicef for every child

Questions we need to answer to select a programme | 1

1st discussion of programmes that might be actually evaluated by core team

Institutional questions

- Which programmes are important for UNICEF & government?
- Which programmes are large enough?
- Which programmes are not compulsory / universally taken-up?
  - i.e. can we expect to have common support?
- Are treated & non-treated (or alternatively treated) groups large enough?
- Can we find out how selection into the programme works?

Questions we need to answer to select a programme | 2

Data related questions

- Large enough?
  - Enough treated & enough controls?
- Informative?
  - Are good measures for treatment, outcome, confounders, heterogeneity available?
- Representative for a relevant population?
  - Which population is relevant?

unicef for every child

Tomorrow
The methodology of Causal Machine Learning & Optimal Policy

Michael Lechner

Swiss Institute for Empirical Economic Research (SEW)
University of St. Gallen | Switzerland