# Introduction to program evaluation in public health

Rapeepong Suphanchaimat, MD PhD

International Health Policy Program (IHPP) and Bureau of Epidemiology,
Department of Disease Control, Ministry of Public Health

8 July 2019

# Outline of presentation

- What is program evaluation?
- Steps of evaluation
- Some basic concepts about (impact) evaluation
- Exercise!

# What is program evaluation?

# Recap of the definitions

- Monitoring = The ongoing process of regularly collecting and analyzing relevant information to make sure you are doing what you set out to do
- Evaluation = When you assess whether what you have been doing is really making the difference that you intended it to
- Monitoring and evaluation are often called ‘M&E’.
- Review = When you look at the results of an evaluation and decide whether it needs to change.
- *Monitoring, evaluation and review are not isolated actions, they are parts of the same process.*

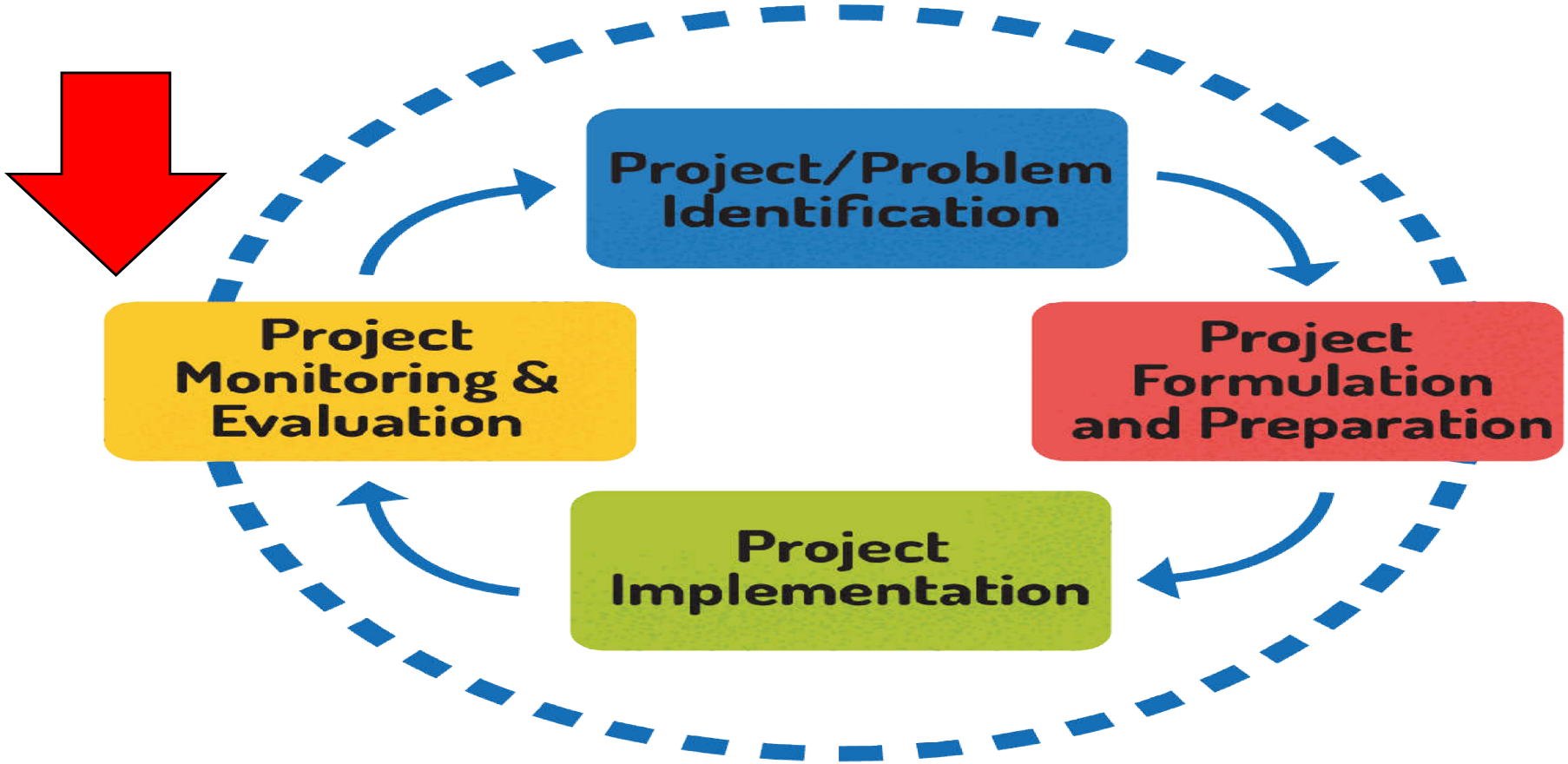


**Ref:** https://scvo.org.uk/running-your-organisation/business-planning/monitoring-evaluation-review

# Benefits of evaluation

- Obtain useful info
- Better inform policy makers
- Better plan for the next phase of the program
- Ensure good governance

# Plan-Do-Check-Act (PDCA)



# Characteristics of good M&E

- Integrate and link with overarching organizational plan
- Think of M&E before the program is implemented
- Clear concrete framework (reporting systems, indicators, timeline, etc)
- Regular
- Open to stakeholder engagement
- Able to communicate with wider public
- Must take ethical concerns into account
  - Fair to everybody
  - Do no harm
  - Fidelity
  - Ensure confidentiality

# Again on definitions! What literature says (1)

- 'Evaluation' has many overlapping definitions, depending on nature of study field, type of evaluation, and objectives of evaluation.
- CDC defines program evaluation as *'the systematic collection of information about the activities, characteristics, and outcomes of programs to make judgments about the program, improve program effectiveness, and/or inform decisions about future program development.'*



**Ref:** Patton MQ. Utilization-focused evaluation: The new century text. 3rd ed. Thousand Oaks, CA: Sage, 1997.

# Again on definitions! What literature says (2)

- CDC classifies evaluation as…
  - **Formative evaluation** ensures that a program or program activity is feasible, appropriate, and acceptable before it is fully implemented.
  - **Process/implementation evaluation** determines whether program activities have been implemented as intended.
  - **Outcome/effectiveness evaluation** measures program effects in the target population by assessing the progress in the outcomes or outcome objectives that the program is to achieve.
  - **Impact evaluation** assesses program effectiveness in achieving its ultimate goals.



**Ref:** http://www.cdc.gov/std/program/ProgEvaluation.pdf.

# Again on definitions! What literature says (3)

- World Bank (WB) classifies evaluation as…
  - **Operational evaluation**
    - Examine how effectively programs were implemented and whether there are gaps between planned and realized outcomes
    - ensuring effective implementation of a program in accordance with the program's initial objectives
  - **Impact evaluation**
    - Studies whether changes are indeed due to the programme intervention and not to other factors
    - Ensuring effective implementation of a program in accordance with the program's initial objectives



**Ref:** World Bank 2010. Handbook on Impact Evaluation

# Again on definitions! What literature says (4)

- Lavis J et al suggested that evaluation could be divided into 4 components:
  - Process evaluation
    - Address adequacy
    - What process of changes lead to observed effects
  - Context evaluation
    - Address transferability
    - Explain whether the observed effects are due to the intervention
  - Effect evaluation
    - Describes and quantifies the intervention's health outcomes as well as its impact on effective coverage, quality of care and equity
  - Economic evaluation
    - Address value for money
    - Look at incremental costs of implementing the intervention compared with the status quo or other alternatives

**Ref:**
https://apps.who.int/iris/bitstream/handle/10665/44204/9789241563895_eng.pdf;jsessionid=BE8E654C05AFA5D0A6058BE4C3DCB750?sequence=1

# Again on definitions! What literature says (5)

- Habicht JP suggested that evaluation should answer two axes:
  - First axis: What do you want to measure? (provision, utilization, coverage, impact)
  - Second axis: How sure you want to be?
    - Adequacy assessment
      - Did the expected change occur?
    - Plausibility assessment
      - Did the programme seem to have an effect above and beyond other external influences?
    - Probability assessment
      - Did the programme have an effect (with significant level at P < x%)

**Ref:** https://www.ncbi.nlm.nih.gov/pubmed/10195658

# Again on definitions! What literature says (6)

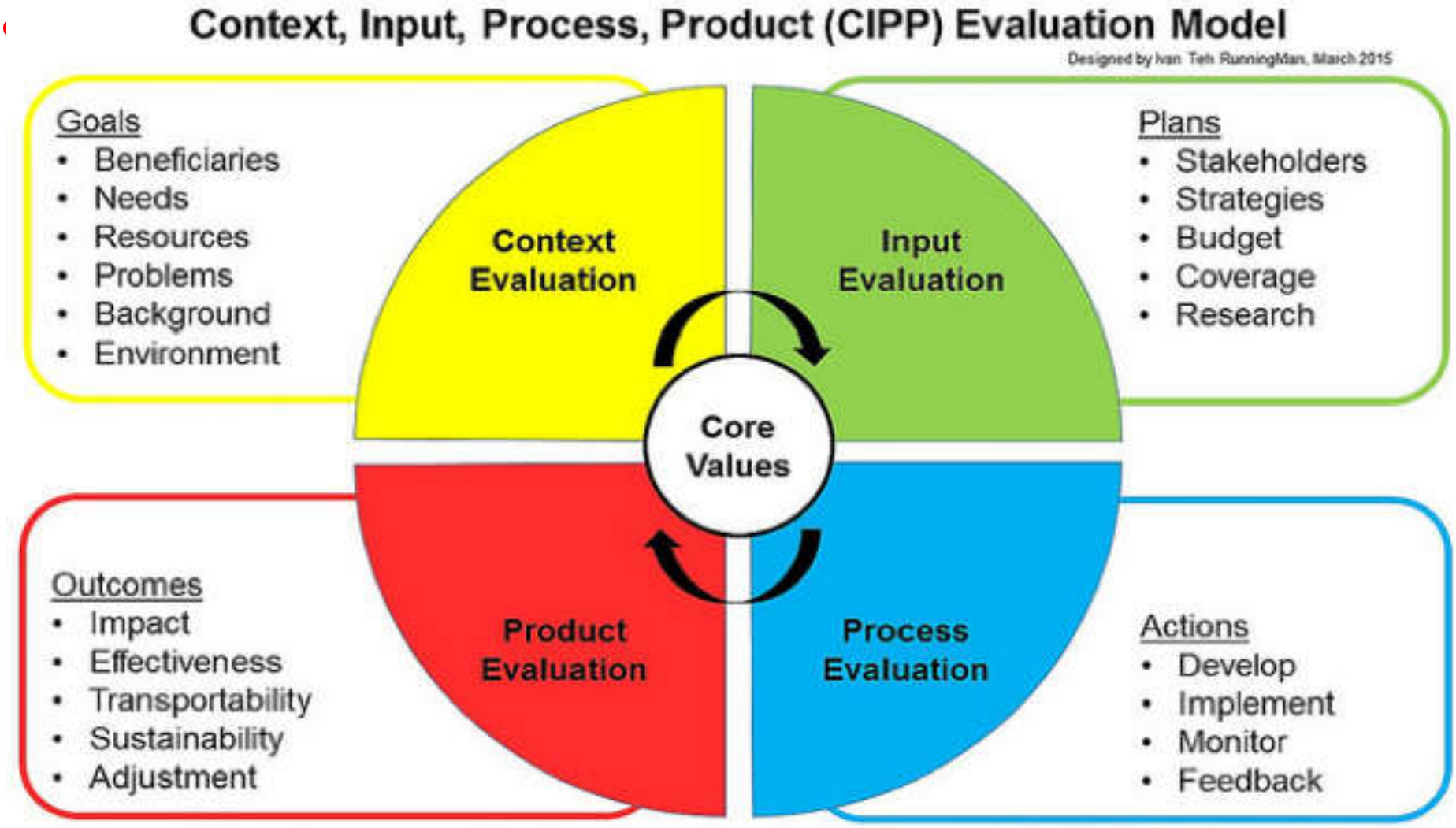
- Examples of program evaluation on ORS supplement

| Type of evaluation | Provision | Utilization | Coverage | Impact |
|---|---|---|---|---|
| **Adequacy** | Changes in availability of ORS in health centres | Changes in numbers of ORS packets distributed in health centres | Measurement of percentage of all diarrhoeal episodes treated with ORT in the population | Measurement of trends in diarrhoeal mortality in intervention area |
| **Plausibility** | As above, but comparing intervention with control services | As above, but comparing intervention with control services | Comparison of ORT coverage between intervention and control areas (or dose-response) | Comparison of diarrhoeal mortality trends between intervention and control areas (or dose-response) |
| **Probability** | As above, but intervention and control services would have been randomized | As above, but intervention and control services would have been randomized | As above, with previous randomization | As above, with previous randomization |

How can we do if we cannot do randomization? We will explore this point later.

**Ref:** https://www.ncbi.nlm.nih.gov/pubmed/10195658

# Again on definitions!
# What literature says (7)

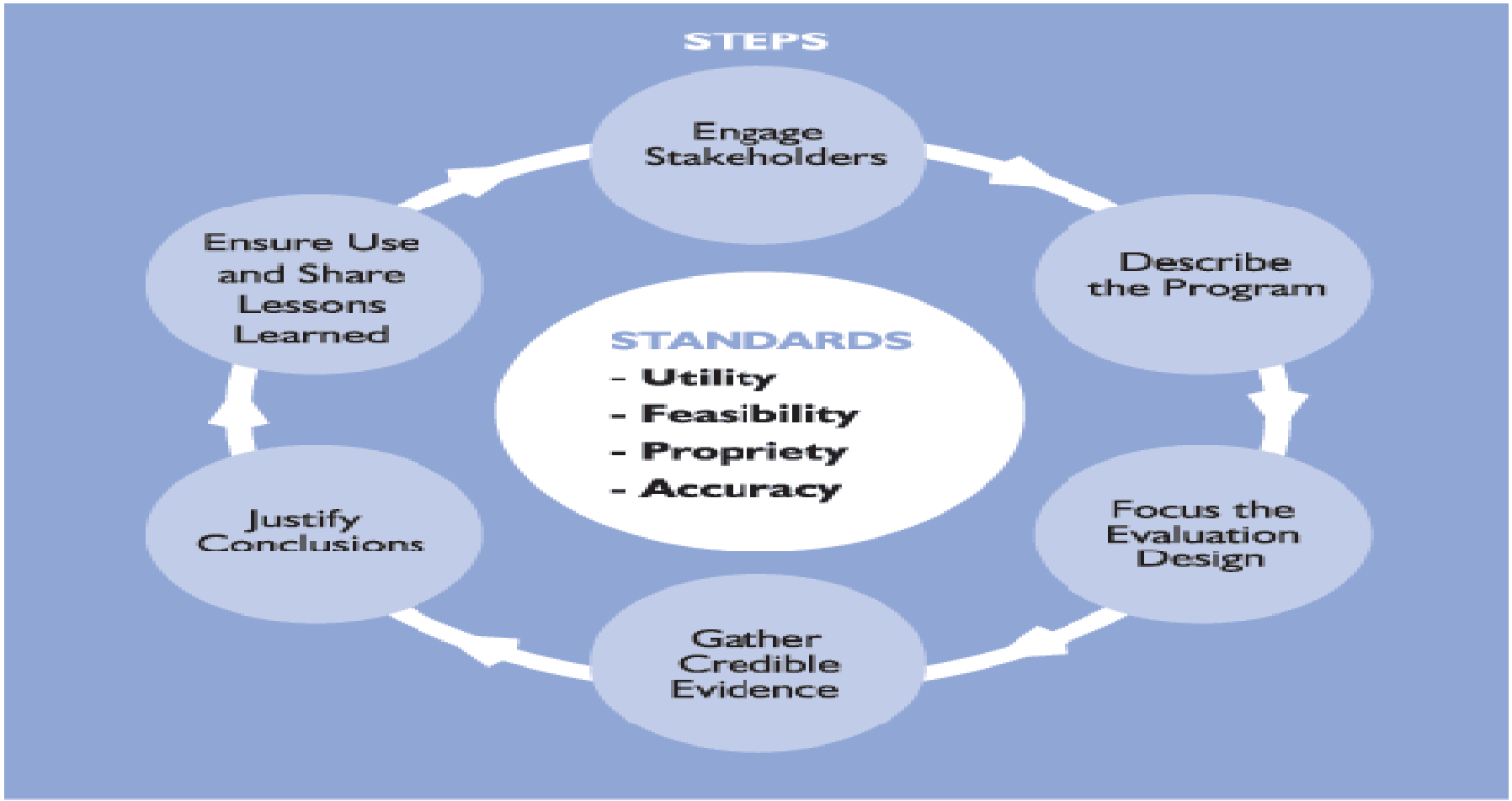



**Ref:** Stufflebeam, D. (2003). The CIPP model of evaluation. In T. Kellaghan, D. Stufflebeam & L. Wingate (Eds.), Springer international handbooks of education: International handbook of educational evaluation.

# Steps of evaluation

# All steps proposed by various textbooks are very alike despite few differences (1)

- CDC proposes the following steps.




**Ref:** https://www.cdc.gov/eval/guide/CDCEvalManual.pdf

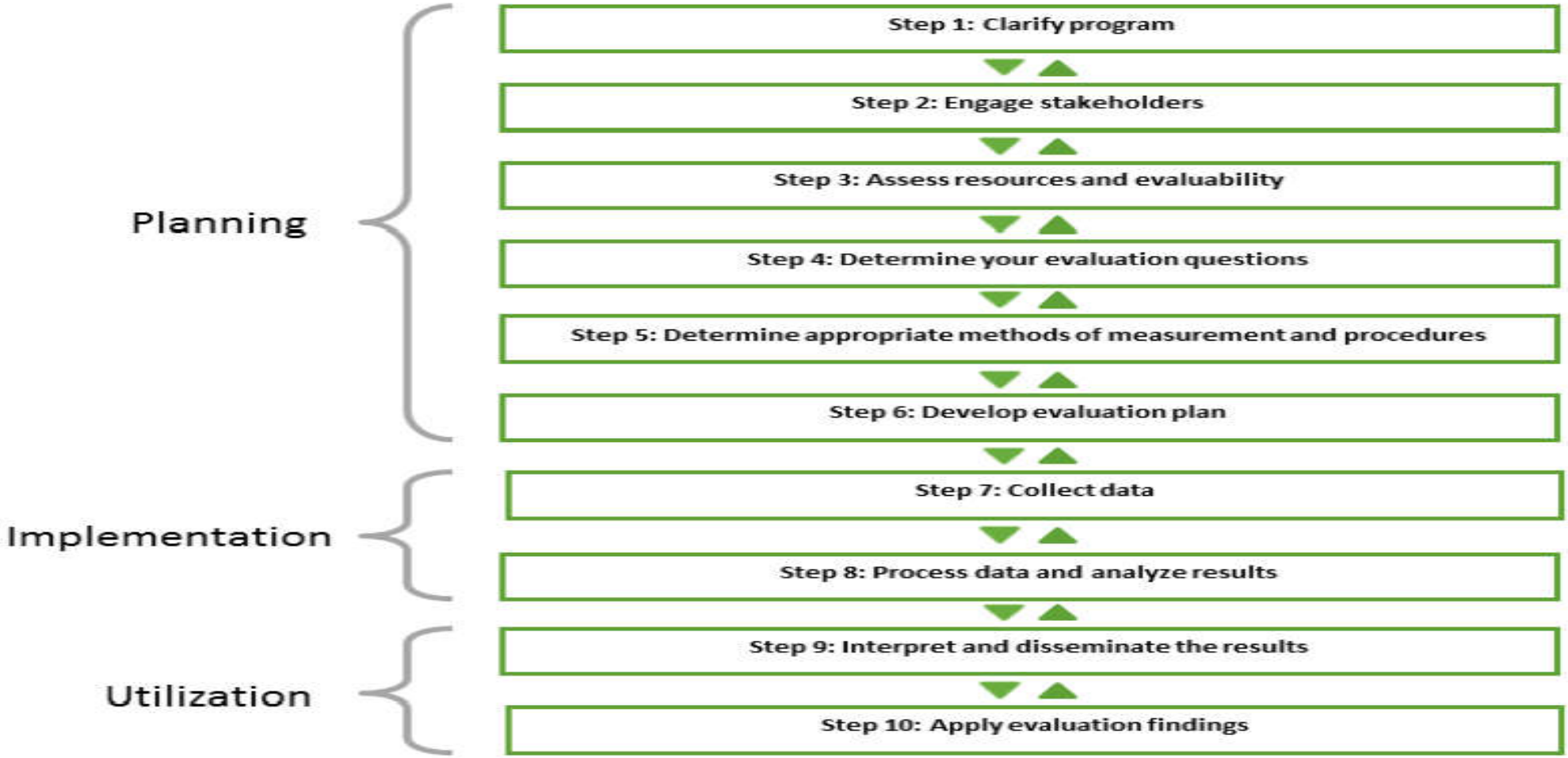
# All steps proposed by various textbooks are very alike despite few differences (2)

- CDC proposes the six-step approach.
  - Engage Stakeholders
  - Describe The Program
  - Focus The Evaluation
  - Gather Credible Evidence
  - Justify Conclusions
  - Ensure Use of Evaluation Findings and Share Lessons Learned

**Ref:** https://www.cdc.gov/eval/guide/CDCEvalManual.pdf

# All steps proposed by various textbooks are very alike despite few differences (3)

- Public Health Ontario proposes ten-step approach.

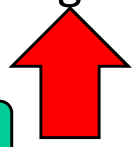





**Ref:** https://www.publichealthontario.ca/-/media/documents/evaluating-hp-programs-workbook.pdf?la=en

# All steps proposed by various textbooks are very alike despite few differences (4)

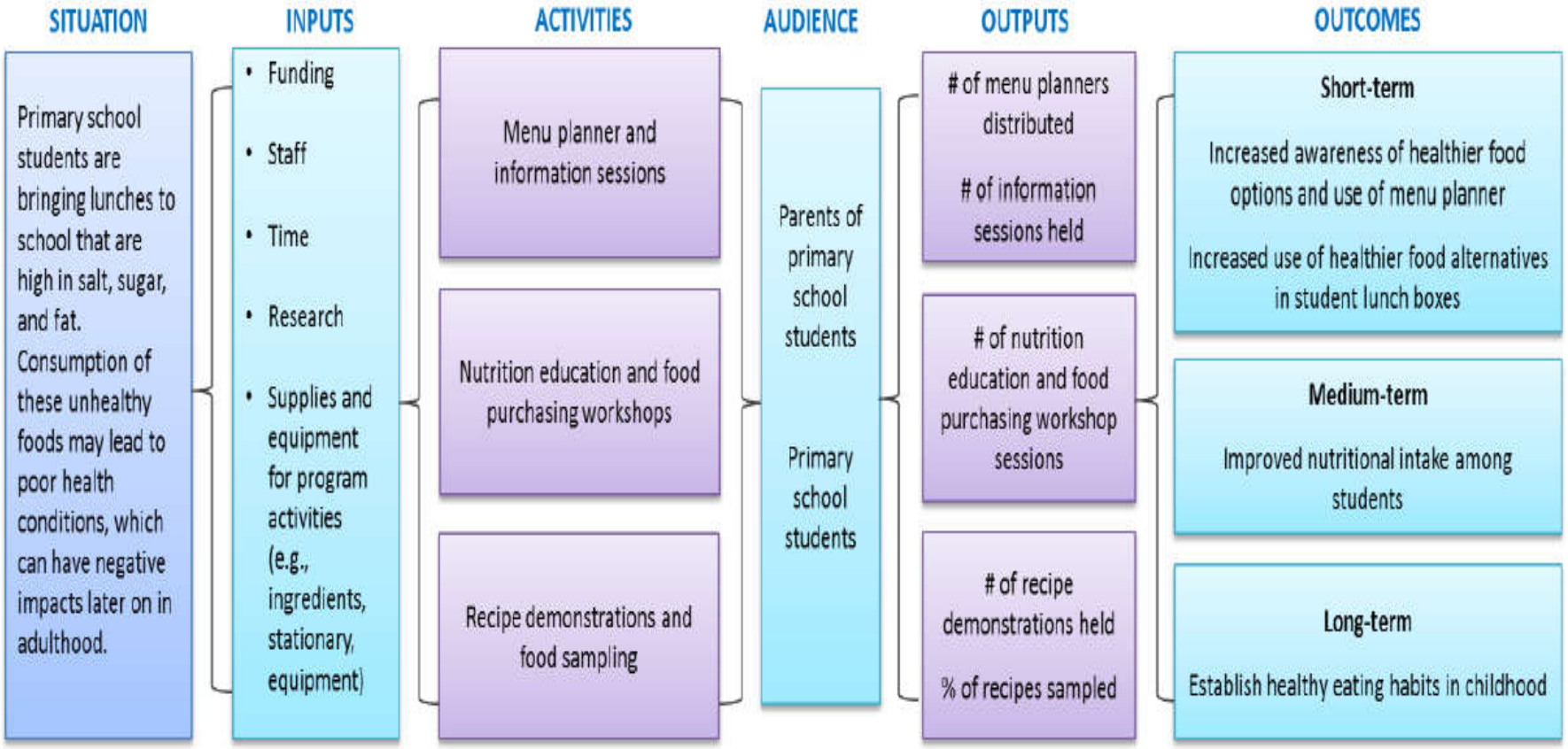
- Using logic model
  - A logic model is a visual illustration of a program's resources, activities and expected outcomes. It is a tool used to simplify complex relationships between various components and can be used during program planning, implementation and evaluation.
  - Common components of logic model
    - Input = The resources invested into a program or initiative.
    - Activities = Activities or interventions that will be carried out as part of the program.
    - Output = Products that are produced from program activities or interventions.
    - Outcome = The changes expected to result from the program on participants
    - Impact = Higher level strategic (usually long-term) change developed by an intervention.



Do not be confused with 'impact evaluation'

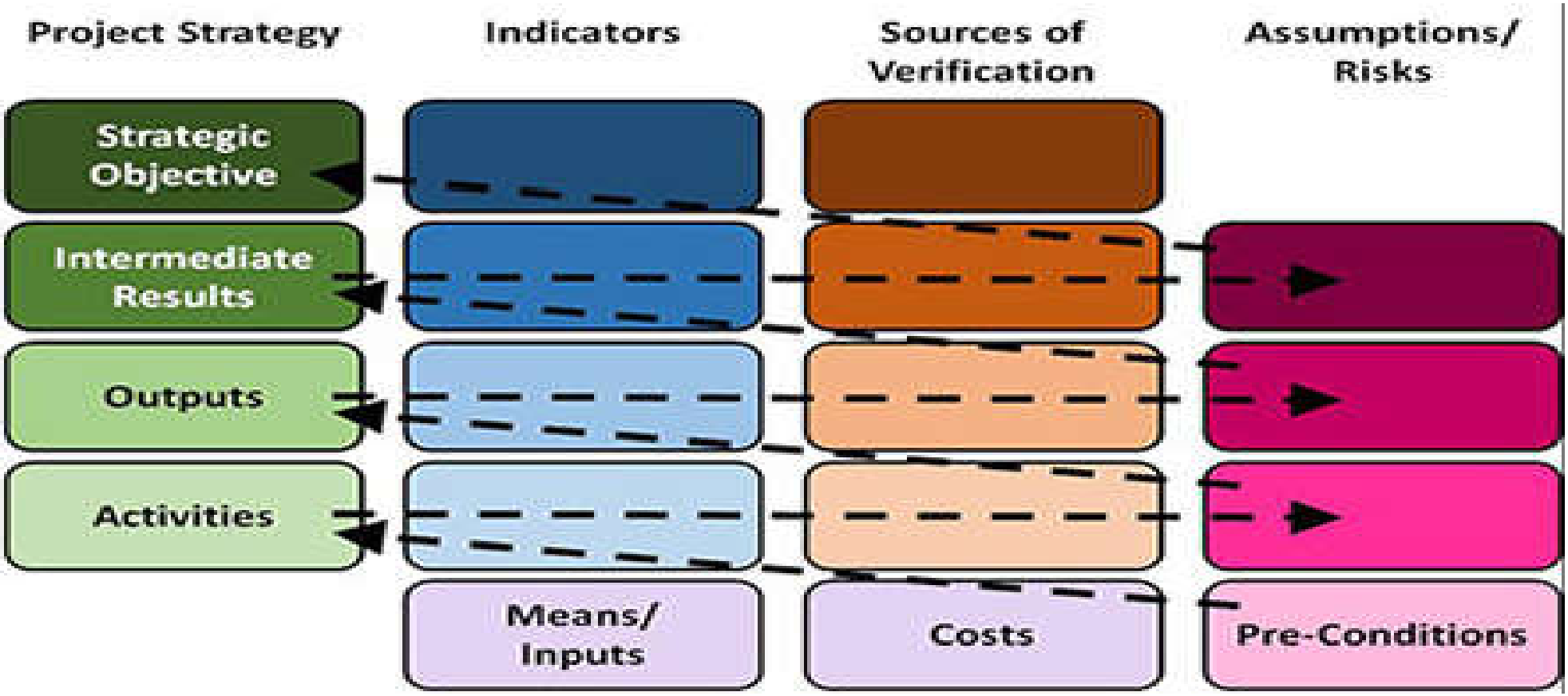
# All steps proposed by various textbooks are very alike despite few differences (5)

- Example of logic model



# All steps proposed by various textbooks are very alike despite few differences (6)

- Example of log frame






**Ref:** https://programs.online.american.edu/online-graduate-certificates/project-monitoring/resource/what-is-a-logframe

# Basic concepts of impact evaluation

# Qualitative or quantitative

- What is program evaluation?
- **Qualitative assessment** can help identify mechanisms through which programs might be having an impact. But a qualitative assessment on its own cannot assess outcomes against relevant alternatives or *counterfactual* outcomes.
- **Quantitative assessment** can help address *counterfactual* problems.
  - Ex ante ➔ predict future impact
  - Ex post ➔ estimate actual impact

# Problems of randomization

- Difficult for evaluator to be involved in every step of the programme
- Can be a self selection into the program
- Difficult to overcome political influences affecting the choice of where to deploy the programme
- Lack of external validity (difficult to extrapolate the outcome)
- What are other options?
  - Historical control group (before *versus* after)
  - Internal control group (institutions or geographical areas or individuals that should have received the full intervention but did not)
  - External control group (institutions or geographical areas without the program)

# Problems of counterfactual (1)

- The problem of evaluation is that while the program's impact (independent of other factors) can truly be assessed only by comparing actual and counterfactual outcomes, the counterfactual is not observed. So the challenge of an impact assessment is to create a convincing and reasonable comparison group for beneficiaries in light of this missing data.
- Ideally, one would like to compare how the same household or individual would have fared with and without an intervention or treatment. But one cannot do so because at a given point in time a household or an individual cannot have two simultaneous existences

# Problems of counterfactual (2)

- Suppose $Y_i(1)$ = outcome of individual i under treatment
- Suppose $Y_i(0)$ = outcome of individual i under non-treatment
- $D = E(Y_i(1)|T_i=1) - E(Y_i(0)|T_i=0)$
- $D = E(Y_i(1)|T_i=1) - E(Y_i(0)|T_i=0) + E(Y_i(0)|T_i=1) - E(Y_i(0)|T_i=1)$
- $D = [E(Y_i(1)|T_i=1) - E(Y_i(0)|T_i=1)] + [E(Y_i(0)|T_i=1) - E(Y_i(0)|T_i=0)]$
- D = ATE + B
- ATE = average treatment effect = average gain in outcomes of participants relative to non-participants as if non-participants were also treated.
- B = selection bias



**Ref:** World Bank 2010. Handbook on Impact Evaluation

# Problems of counterfactual (3)

- However selection bias may disappear if we can assume that selection bias would disappear if one could assume that whether or not households or individuals receive treatment (conditional on a set of covariates, *X*) were independent of the outcomes that they have.
- This assumption is called the *assumption of unconfoundedness*, also referred to as the *conditional independence assumption.*

$$(Y_i(1), Y_i(0)) \perp T_i \mid X_i$$

# (examples) Techniques to handle selection bias in impact evaluation (1)

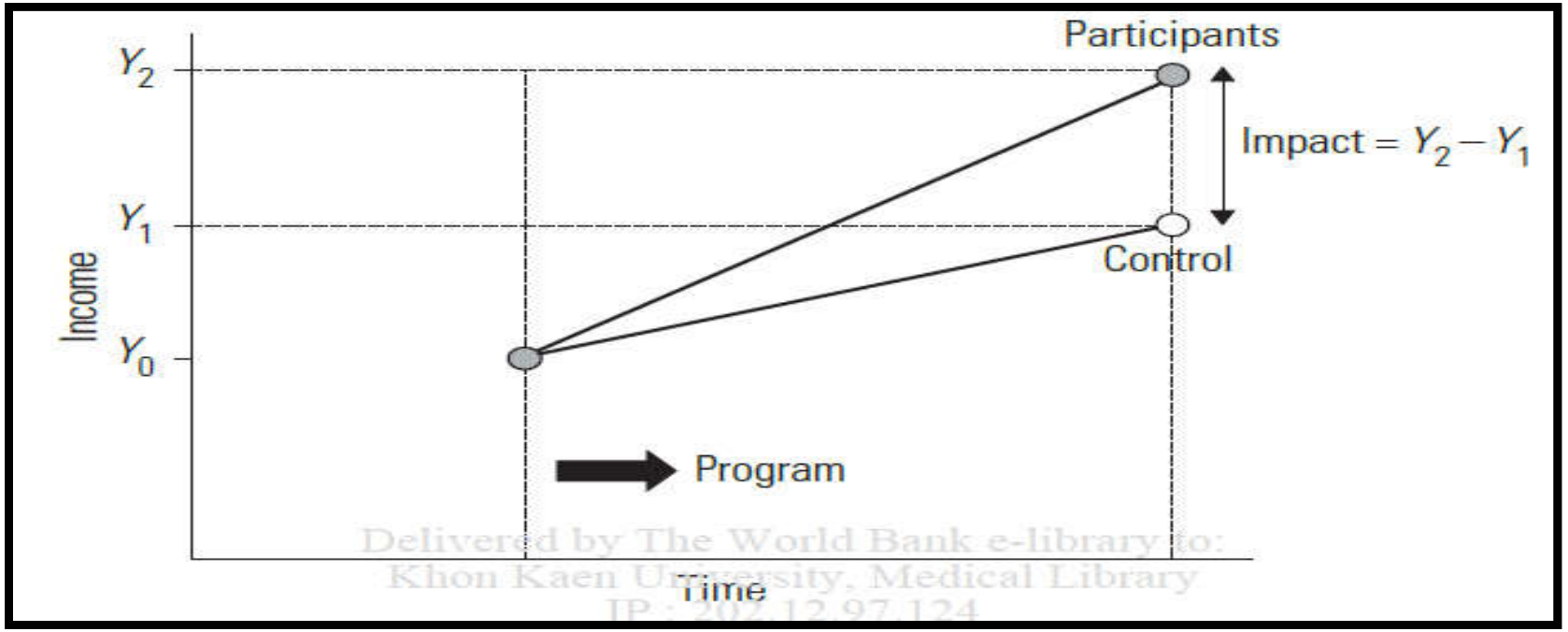
- Randomization
- Propensity score matching (PSM)
- Double difference (DD)
- Instrumental variable (IV)
- Regression discontinuity (RD)
- Distributional impact
- Structural and other modeling approaches



**Ref:** World Bank 2010. Handbook on Impact Evaluation

# (examples) Techniques to handle selection bias in impact evaluation (2)

- Randomization
  - ATE = TOT where TOT is treatment effect of the treated
  - TOT = $E(Y_i(1)-Y_i(0)|T_i=1)$, the difference in outcomes from receiving the program as compared with being in a control area for a person or subject i randomly drawn from the treated sample

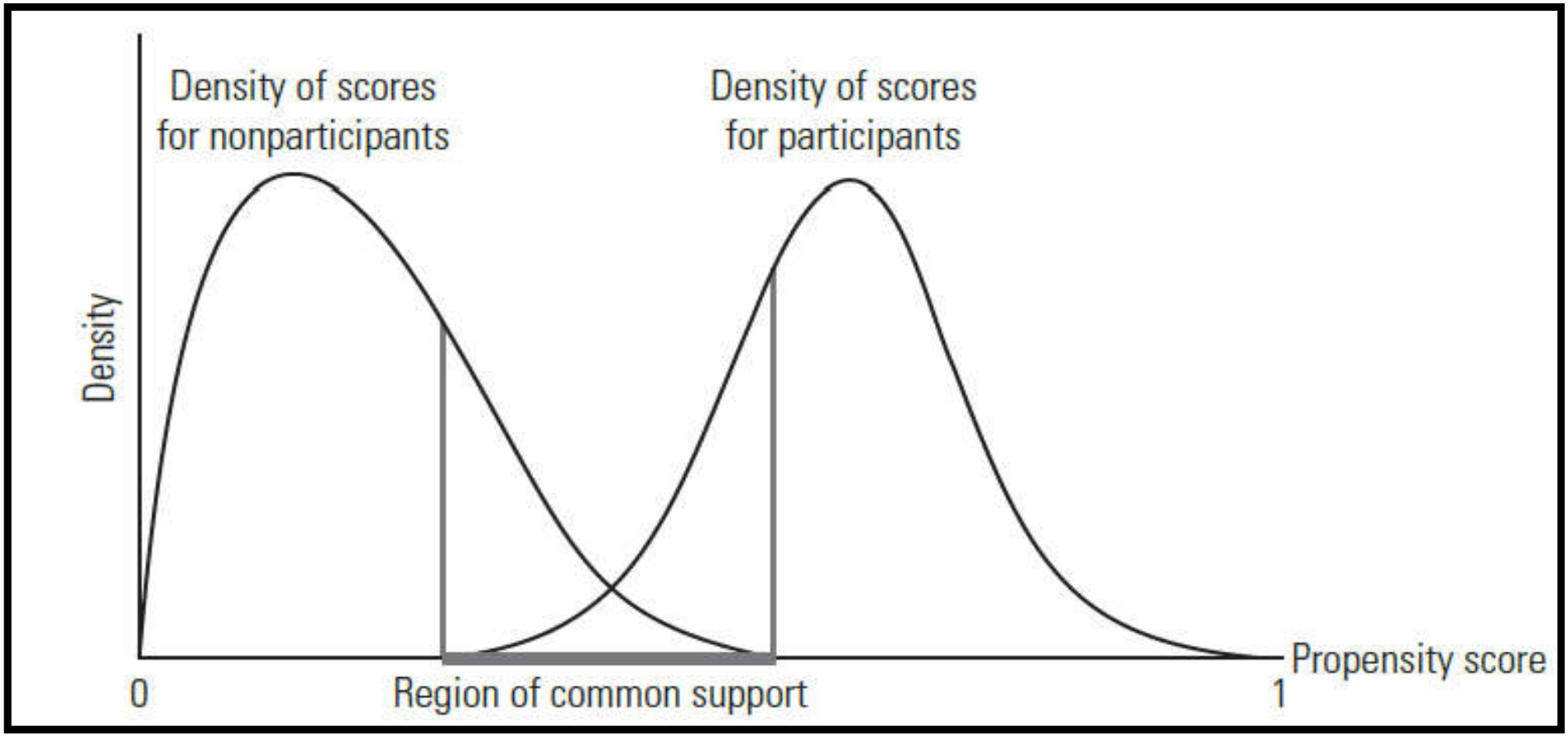


# (examples) Techniques to handle selection bias in impact evaluation (3)

- PSM
  - Participants are then matched on the basis of this probability, or *propensity score*, to nonparticipants. The average treatment effect of the program is then calculated as the mean difference in outcomes across these two groups.
  - The validity of PSM depends on two conditions: (a) conditional independence (namely, that unobserved factors do not affect participation) and (b) sizable common support or overlap in propensity scores across the participant and nonparticipant samples.
  - Matching techniques: nearest neighbors (NN), caliper and radius matching, stratification and interval matching, and kernel matching and local linear matching (LLM)

# (examples) Techniques to handle selection bias in impact evaluation (4)

- PSM
  - PSM does not require baseline data and it is semi-parametric method.

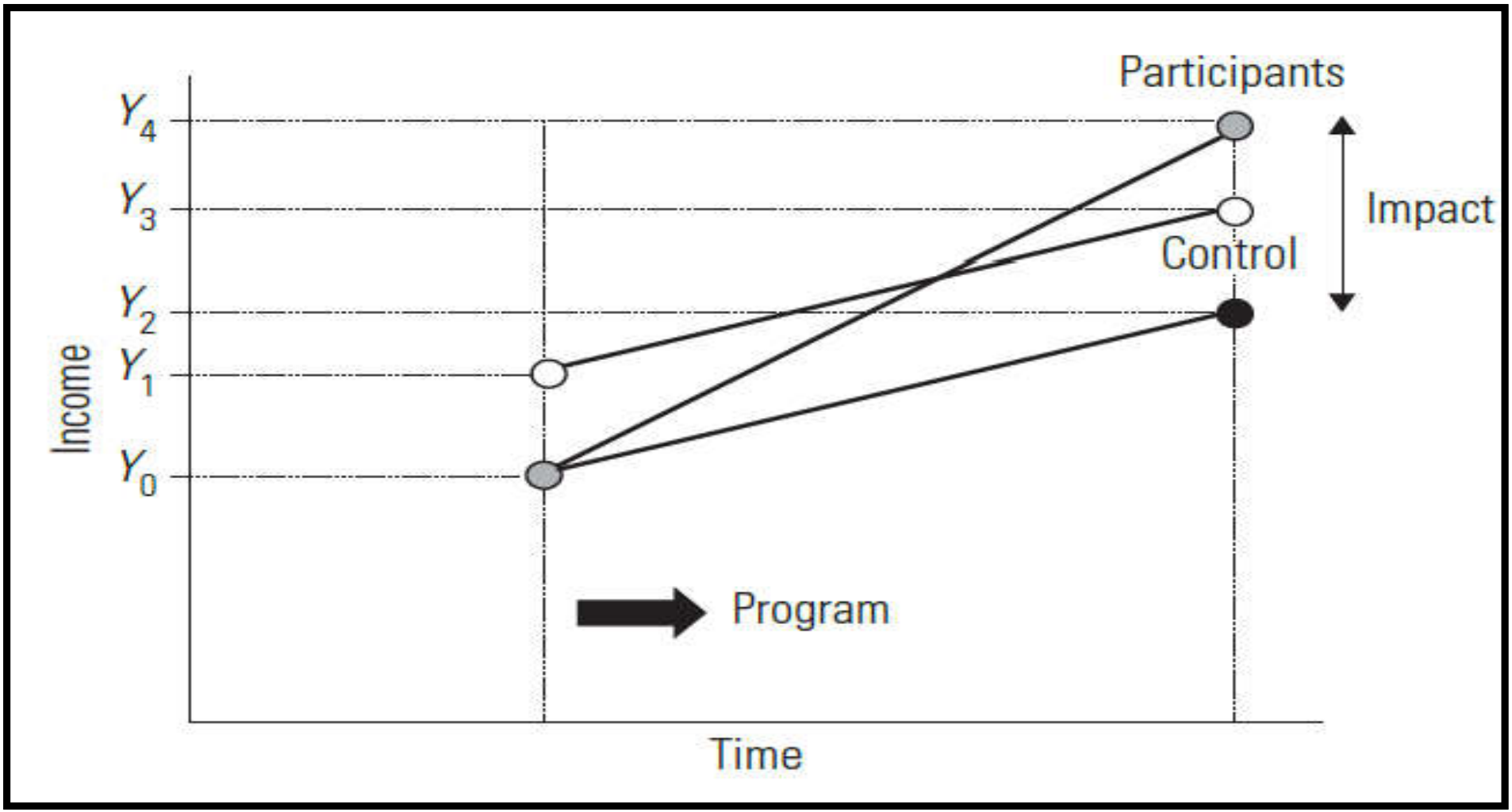


# (examples) Techniques to handle selection bias in impact evaluation (5)

- DD
  - Double-difference (DD) methods, compared with propensity score matching (PSM), assume that unobserved heterogeneity in participation is present—but that such factors are time invariant.
  - With data on project and control observations before and after the program intervention, therefore, this fixed component can be differenced out.

# (examples) Techniques to handle selection bias in impact evaluation (6)

- DD

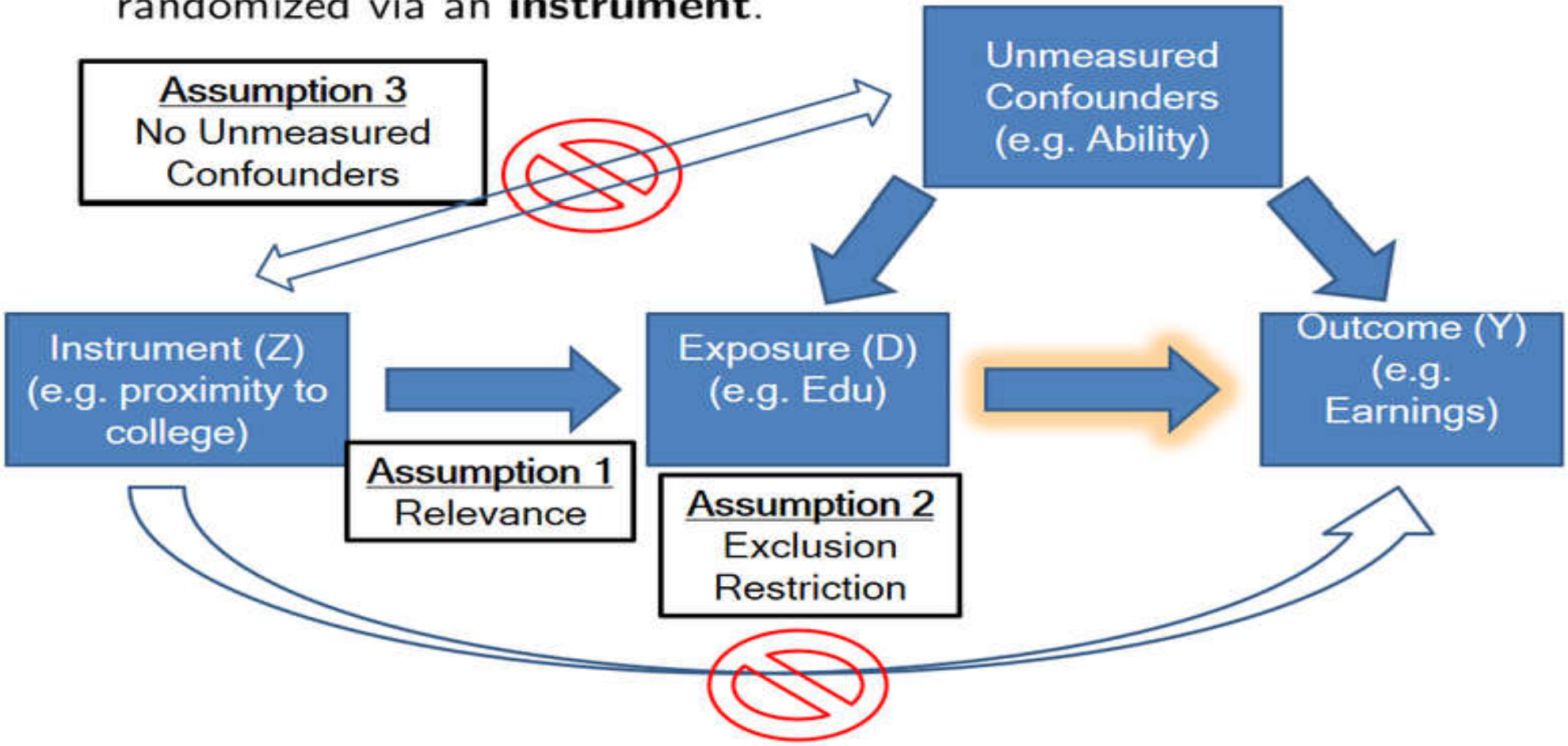


# (examples) Techniques to handle selection bias in impact evaluation (7)

- IV
  - Instrumental variable (IV) methods allow for endogeneity in individual participation, program placement, or both. With panel data, IV methods can allow for time-varying selection bias. Measurement error that results in attenuation bias can also be resolved through this procedure.
  - The IV approach involves finding a variable (or instrument) that is highly correlated with program placement or participation but that is not correlated with unobserved characteristics affecting outcomes.
  - Endogeneity = Programs are placed deliberately in areas that have specific characteristics that may or may not be observed and that are also correlated with outcomes *Y*.

# (examples) Techniques to handle selection bias in impact evaluation (8)

- IV
  - Key idea: "induce" randomization on exposure that can't be randomized via an **instrument**.

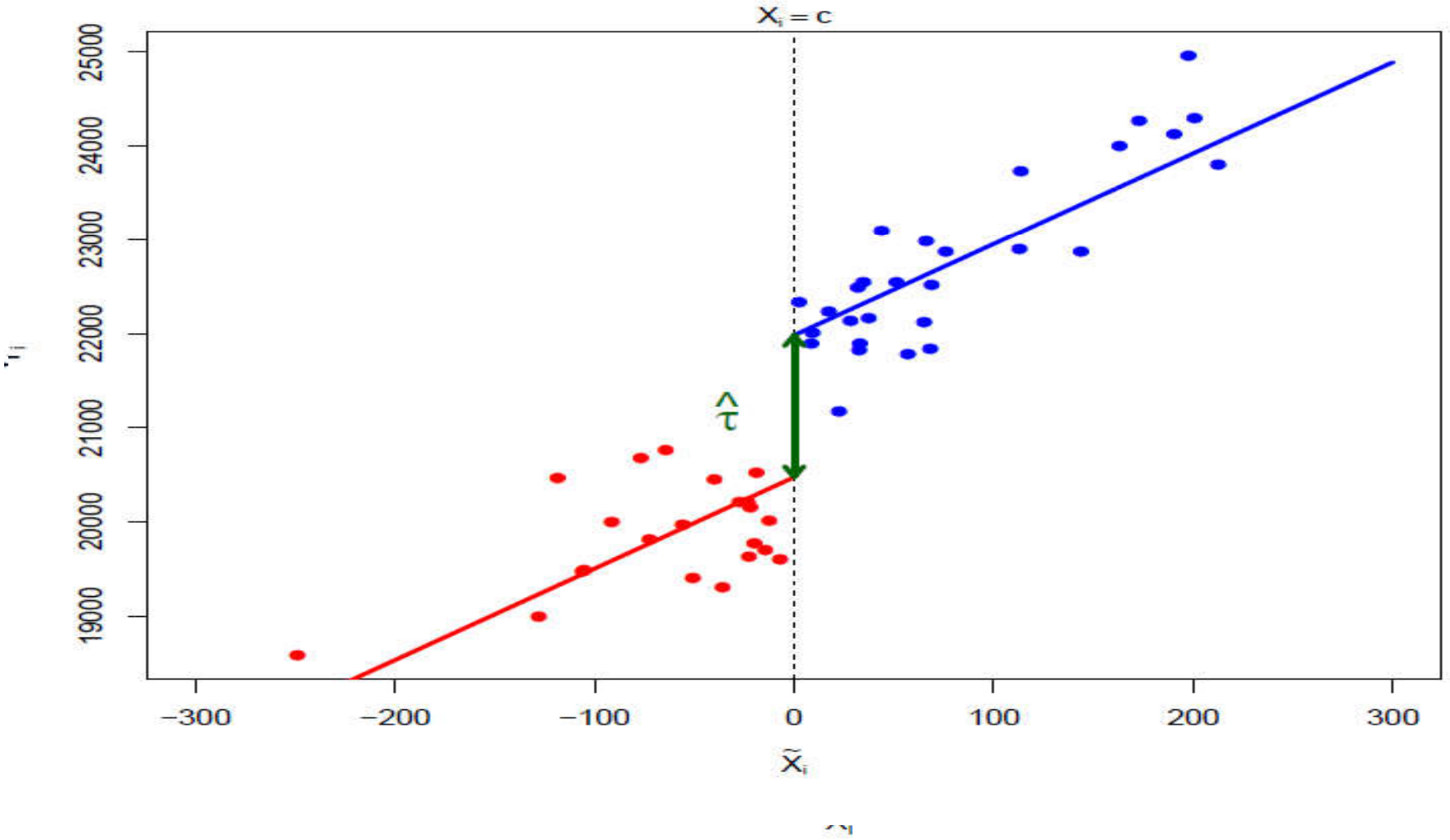


# (examples) Techniques to handle selection bias in impact evaluation (9)

- RD
  - RD is a method that accounts for potential selection or participation on observed and unobserved characteristics.
  - Discontinuities in program implementation, based on eligibility criteria, can be very useful in non-experimental program evaluation. People above and below the threshold, assuming they are similar in observed characteristics, can be distinguished in terms of outcomes.
  - Unobserved heterogeneity may be a factor if people within the eligible targeting range exhibit variation in actual take-up of the program, leading to selection bias. In that case, eligible and non-eligible samples close to the eligibility cutoff would be taken to compare the average program effect.
  - RD may have some disadvantages. For instance, (1) it produces local average treatment effects that are not always generalizable; and (2) that the effect is estimated at the discontinuity, so, generally, fewer observations exist.

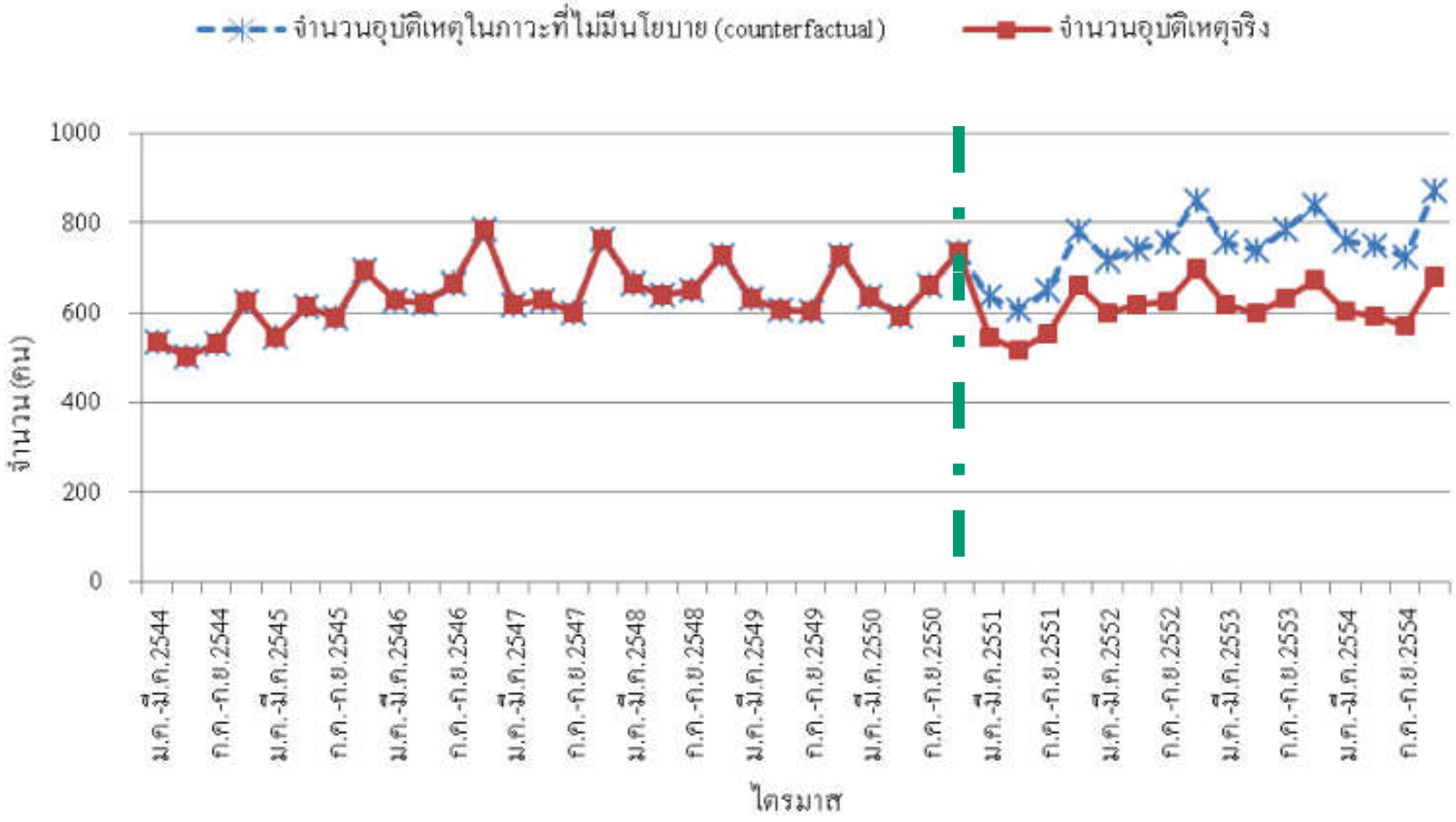
# (examples) Techniques to handle selection bias in impact evaluation (10)

- RD



# (examples) Techniques to handle selection bias in impact evaluation (11)

- Other models, such as interrupted time series (ITS)




**Ref:** Suphanchaimat W, Suphanchaimat R: Traffic accident prevention measures and the reduction of patients from traffic injuries: a case study of Khon Kaen Regional Hospital. Journal of Health Science. 2013;22:765-75.

# Conclusion on what we have learned so far

- Definition of monitoring and evaluation
- Several types of evaluation
- Some basic theories about evaluation
- Step in evaluations
- Impact evaluation and ideas of counterfactual
- Some techniques to handle with selection bias in program evaluation

# Thank you

# Questions and comments