Employment and Social Development Canada
Emploi et Développement social Canada
Canada

Gender-Based Analysis Plus Exploratory Evaluation Study on Selected Labour Market Programs

Presentation to the Quantitative Impact Assessment Workshop
March 2024

Presentation outline

- Background
- Data sources and indicators
- Methodology
- Results
- Conclusion
- Annexes
- References

Background

ESDC uses “matching” methods to assess the effectiveness of its labour market programs

- “Matching” is robust, but only provide average impacts.
- Not possible to estimate the distribution of program impacts across participants.
- Difficult to conduct subgroup analyses on different intersecting factors of identity.

Recent developments in machine learning have been applied to evaluate labour market programs in Europe (Belgium and Switzerland):

- Machine learning was used to estimate granular incremental impacts at the individual level, thereby also uncovering “what works for whom” (Wager and Athey, 2018; Lechner, 2019).
- Causal Machine Learning Evaluation of Training in Belgium (Lechner, 2019)

Scope of the study

- Test the effectiveness of a novel machine learning method to estimate incremental program impacts according to different GBA Plus intersecting identity factors.
- Examine two active labour market programs:
  - Labour Market Development Agreements (LMDA); and
  - Opportunities Fund for Persons with Disabilities (OFPD).

What is Gender-Based Analysis Plus (GBA Plus)?

- An analytical process used to assess the experience of different women, men and gender diverse people with regard to policies, programs and initiatives.
- The 'plus' in GBA Plus acknowledges that GBA goes beyond biological (sex) and sociocultural (gender) differences.

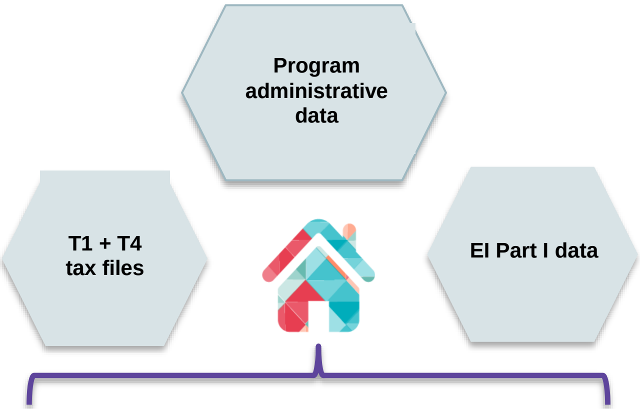
Source : Women and Gender Equality Canada[1]

[1] https://women-gender-equality.canada.ca/en/gender-based-analysis-plus/what-gender-based-analysis-plus.html

Data sources: Labour Market Program Data Platform

This study uses integrated datasets of rich administrative data.

- The initial stage cleans up duplicated records, build Action Plan Equivalents(APEs) from program interventions, and construct the final database from program data, CRA data and EI Part I data.



LMDA
2010-2012 cohort, followed up to 2018

OFPD
2011-2012 cohort, followed up to 2018

Two groups of interest to produce the incremental impacts:

- **Participant groups:** Individuals who participated in LMDA and OFPD
- **Control groups:** similar individuals who did not participate in LMDA or OFPD:
  - For LMDA: Active EI claimants who did not participate in LMDA.
  - For OFPD: individuals with disabilities who participated in Employment Assistance Service.

Control variables and indicators

Main indicators are the 5-year post-program average of:

- Incidence of Employment (pp.)
- Employment earnings ($)
- Dependence on income support (p.p)

* Observations with missing outcome indicators were excluded to ensure proper functioning of the chosen algorithm

Over 75+ variables used to build the covariate matrix:

Socio-Demographics

- Age
- Gender
- Indigenous status
- person with disability
- Immigrant
- Visible minority
- Marital status
- Education level
- Province of Residence
- Urban/Rural indicator

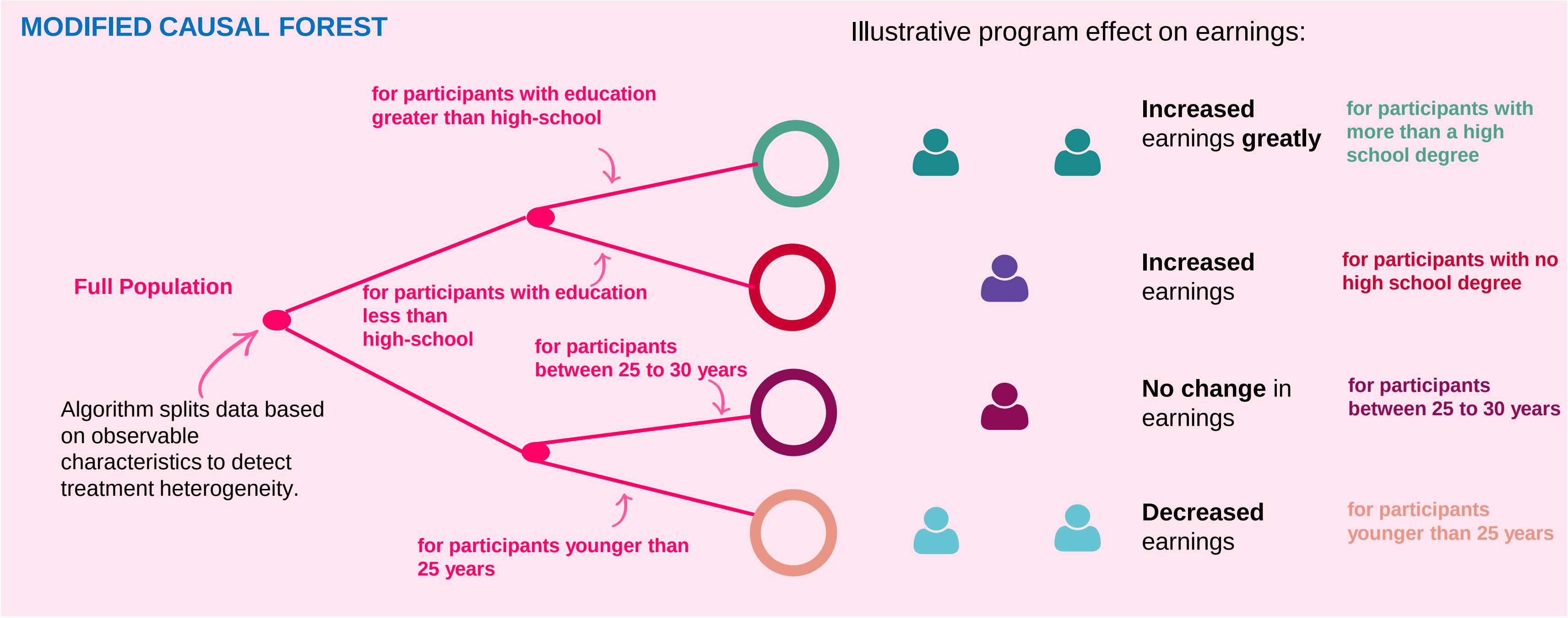
Labour Market Characteristics

- Reason for job separation
- Industry (NAICS)
- New entrant to labour force
- Past interventions taken
- Past Earnings up to 5 years
- Past EI Benefit usage up to 5 years
- Past social assistance income up to 5 years
- Past tuition amounts up to 5 years

Methodology: Modified Causal Forests

The study uses the Modified Causal Forest (MCF):

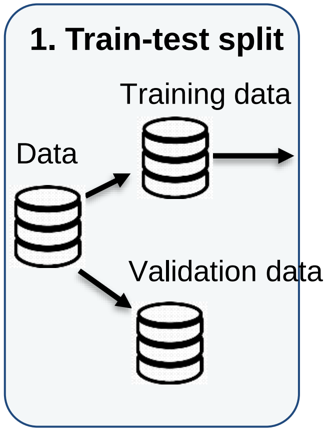
- A supervised causal machine learning algorithm that builds an ensemble of decorrelated trees, learns the characteristics from the data and estimates the program impacts in R-programming language (Lechner, 2019).



Causal Inference Literature

- The incremental impact refers to the measurable impact or change in an outcome that can be attributed to a specific intervention or treatment, often assessed by comparing the results between a treatment group that received the treatment and a control group that did not.
- Wager & Athey (2018) proposed the causal forests with valid statistical inferences based on the traditional machine learning method, random forest, proposed by Briedman (2001).
- Lechner (2019) modified the error terms of causal forests to provide more unbiased and granular estimators.

Step 1: Train-test split

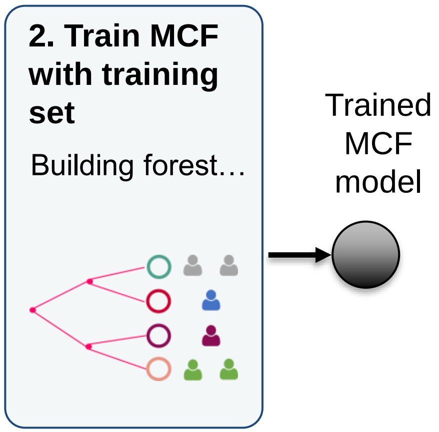


The linked program administrative data is randomly split into:

- 50% **Training data**: to train the MCF algorithm
- 50% **Validation data**: to estimate the effects by applying the trained MCF

This is to avoid the MCF to over-perform on the data it has "seen" before. By showing it a new set of data, the testing data, we can make sure we "generalize" the MCF.

Step 2: Train MCF with training set

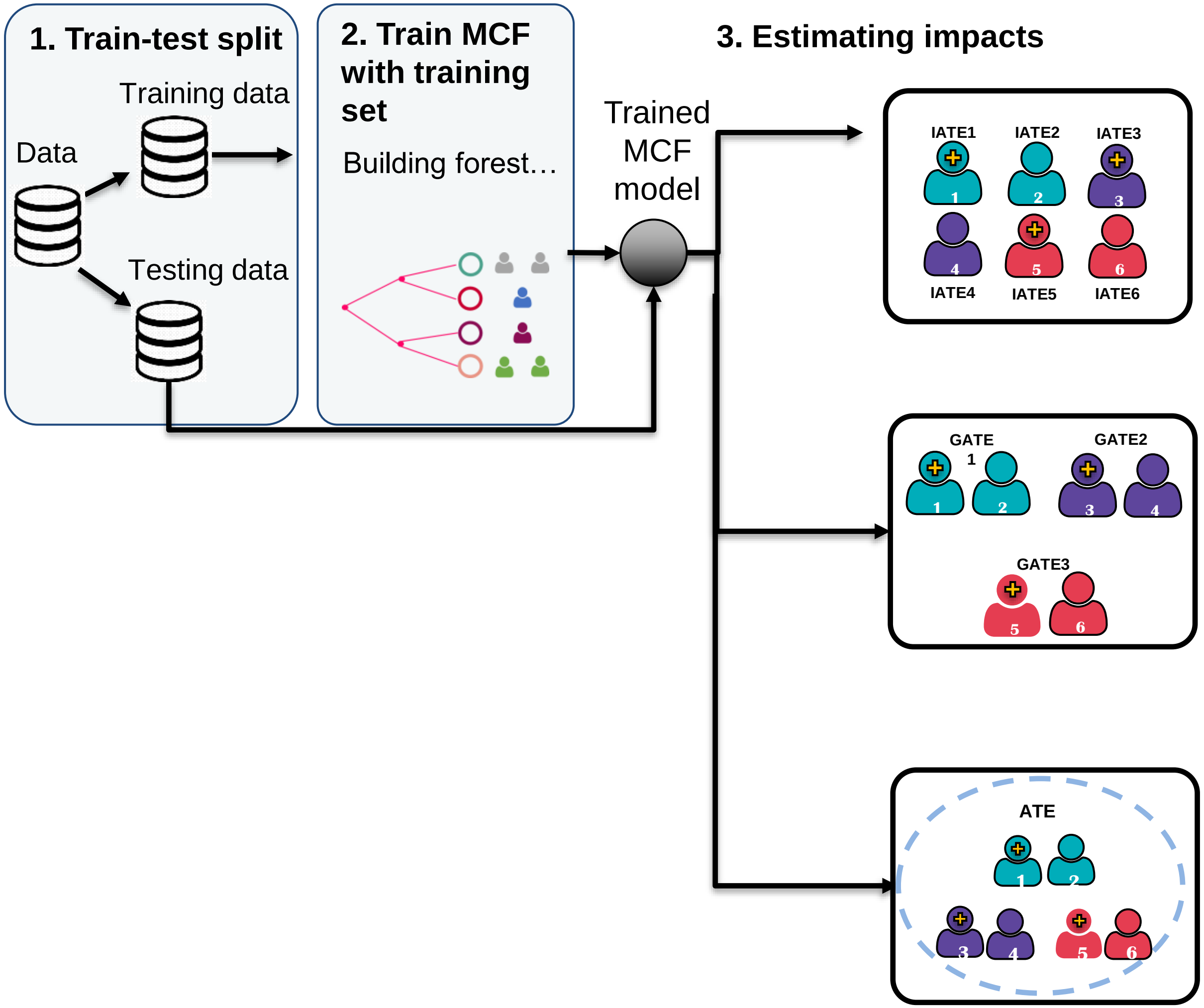


While training the MCF, the algorithm internally splits the data into two parts to build the forests:

- **Training data**: to learn how to make splits
- **Honest data**: to estimate the treatment effects among individuals belonging to the same split

* For objective function to determine the optimal splitting rules, refer to the [Annex]

Step 3: Estimating impacts



Individualized Average Treatment Effect (IATE) *NEW*

Measures the impact of a program on an individual with a given set of characteristics or profile. Represents causal program impact at the *finest level of granularity.*

Grouped Average Treatment Effect (GATE) *NEW*

Estimated by aggregating and weighting the IATEs over specific subgroups. Unlike traditional subgroup analyses, GAPIs can be *compared across groups.*

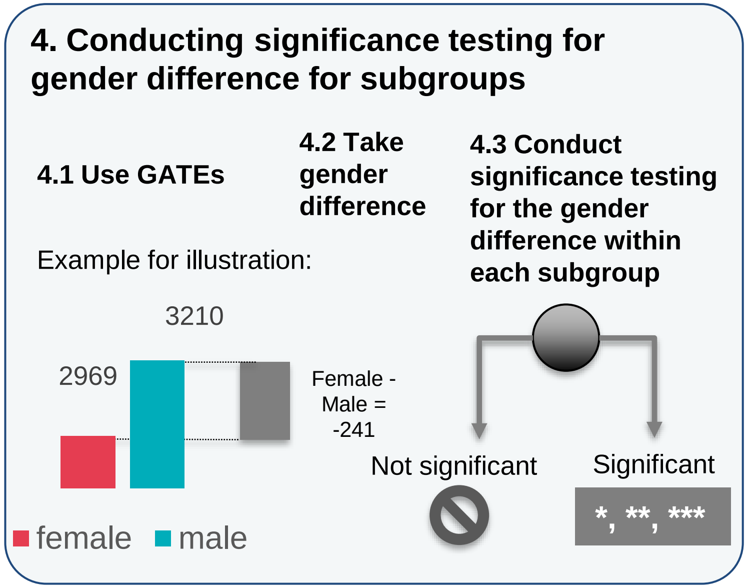
Average Treatment Effect (ATE)

Represents the population average program impact.

Note: ✚ Indicates that the individual is a participant

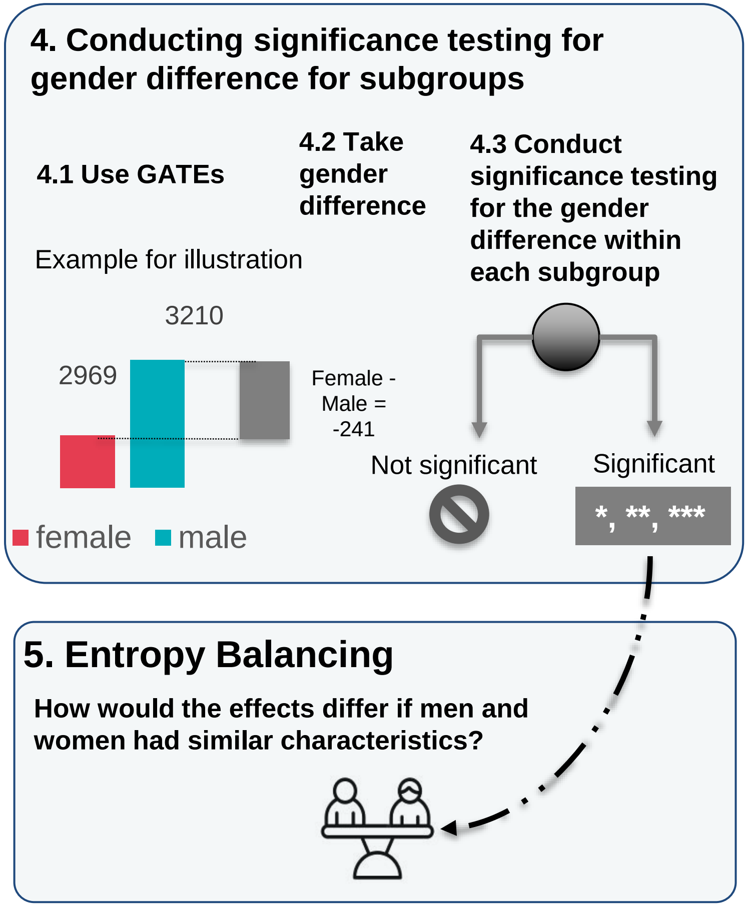
Step 4: Conducting significance testing

- The MCF algorithm determines if two GATEs are statistically significantly different from each other.
- If so, we send the GATE into next step for further investigation.



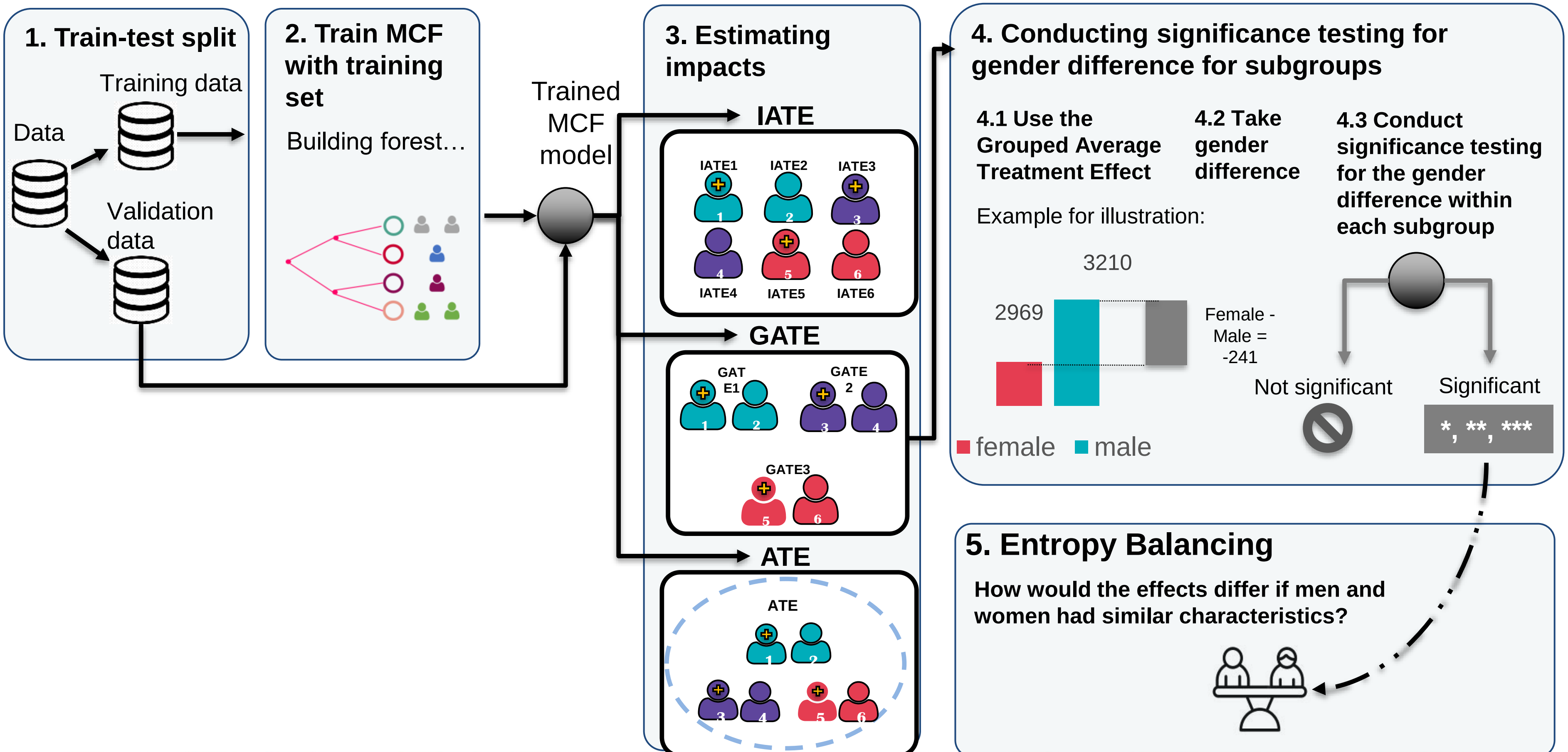
Step 5: Entropy Balancing

- We estimate and balance the entropies on the IATEs so that the characteristics of male individuals can be similar to the female individuals in the data.
- The balancing is done on the control of gender.
- This allows us to demonstrate how the effects would differ if men and women had similar characteristics



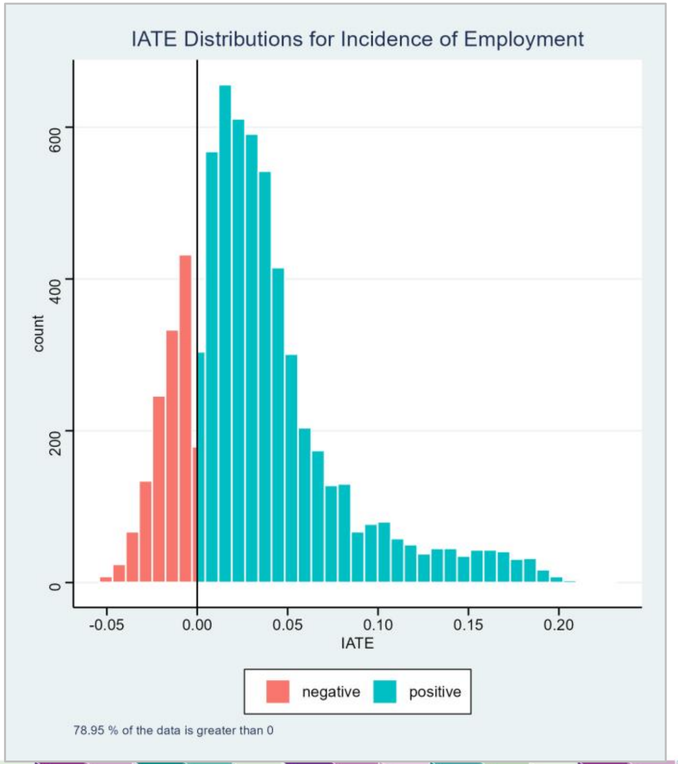
Methodology: Recap

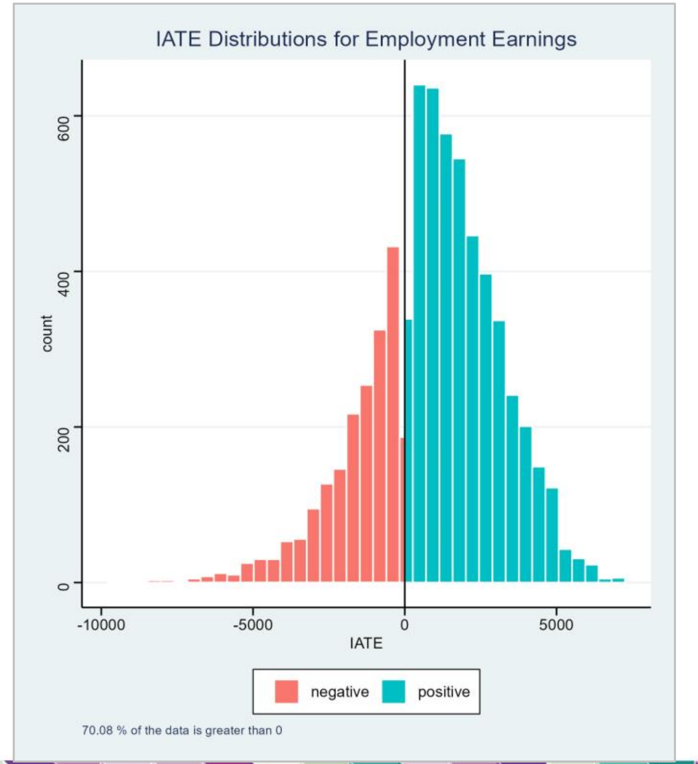
Using the results from the MCF, the methodology includes significance testing and entropy balancing to assess gender differences.



Examples of results – LMDA Targeted Wage Subsidies Distribution of IATEs

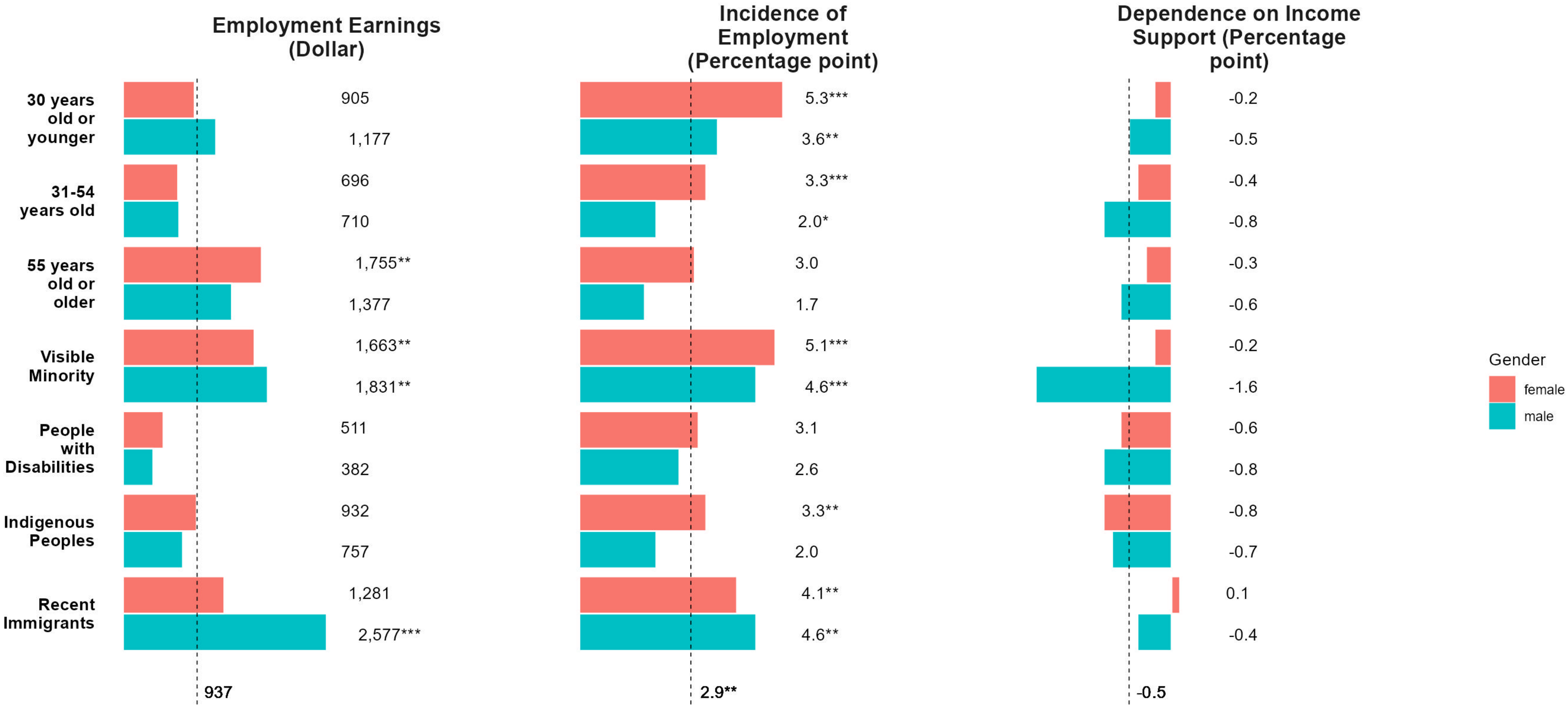
- The incremental impacts revealed that there is limited heterogeneity in program impacts.
- The majority of active EI claimant participants in TWS benefited from it.
- The results indicate that:
  - 79% of participants experienced an increase in the incidence of employment
  - 70% of participants increased their employment earnings





Examples of results – LMDA Targeted Wage Subsidies

Incremental impacts by gender and by other subgroups, 5-year post-participation period, annual averages



- Overall, all gender subgroups increased their incidence of employment. Two groups saw a larger increase in their incidence of employment and employment earnings:
  - Both female and male participants who were visible minorities
  - Male recent immigrants

<u>Notation for significance levels</u>:

*** 1% level

** 5% level

* 10% level

The overall average treatment effect on the participants annotated as the dashed line.

Example of Entropy Balancing– LMDA Targeted Wage Subsidies

Overall, we found no gender differences in the program impacts

- Initial results for TWS suggested gender differences between men and women who were recent immigrants
  - Men increased their employment earnings by $1,296 more than women (statistically significant at 1%), which suggests a difference in program impacts.
- But after controlling for their socio-demographic characteristics, the differences became non-statistically significant, suggesting no difference in program impact.

	Employment earnings (dollars)
Without controlling for socio-demographic characteristics	-1,296**
When men have similar socio-demographic characteristics as women	-328

Notation for significance levels: *** 1% level, ** 5% level, * 10% level.

Conclusion of the study

The machine learning method was successful in generating robust results for key program interventions:

- Overall, results align with previous evaluations and provide a new level of granularity to examine program impacts through a GBA Plus lens.
- Results can help understand the distribution of impacts on various groups and inform policy development and support program design from the perspective of “what works best for whom”.

As part of future evaluation cycles:

- Machine learning results could provide a new line of evidence to explore differentiated impacts on subgroups when feasible.
- Complementary qualitative research and analysis would be required to contextualize these results. This could be done as part of future program-specific evaluation cycles.

Limitations

- This study is limited to the information available in administrative data:
  - Biological sex was used as a proxy for gender and data was not available for some GBA Plus factors of identity.

- Pre-existing differences might exist between participants and non-participants that were not measured during the matching process:
  - For example: ability, health, and motivation to seek employment.

- Results are not directly comparable between programs:
  - This analysis used comparison groups built by program intervention.

- The study does not capture participation in multiple interventions:
  - By using Action Plan Equivalents, the analysis attributed the longest intervention as the principal intervention in the unit of analysis.

Ways forward

- Explore using the MCF method as part of upcoming evaluations of labour market programs.
  - When only smaller datasets are available, the traditional matching method will remain the preferred method for conducting net impact analysis.

- Continue to collaborate with Prof. Lechner on ways to measure the effect of gender and other intersecting factors of identity.

- Sharing our experience with exploratory ML studies with others.

Annex A: Potential Outcomes Framework

For a set of i.i.d individuals $i = 1, \ldots, n$, we observe a tuple of (X_i, Y_i, D_i), comprised of

- A covariate X_i
- An outcome Y_i
- A treatment assignment D_i

Let D denote the treatment that may take a known number of M different integer values from 0 to $M - 1$. The (potential) outcome of interest that realises under treatment d is denoted by Y^d.

Goal is to find $IATE(m, l; x, \Delta) = E\left(Y^m - Y^l \middle| Z = z,\ D \in \Delta\right)$

$IATE(m, l; x, \Delta)$ measure the mean impact of treatment m compared to treatment l for units with features x that belong to treatment groups Δ, where Δ denotes all treatments of interest.

Annex B: Identification Assumptions

$$\{Y^0,\ldots,Y^m,\ldots,Y^{M-1}\} \coprod D \mid X = x, \qquad \forall x \in \chi; \qquad (CIA)$$
$$0 < P(D = d \mid X = x) = p_d(x), \qquad \forall x \in \chi, \forall d \in \{0,\ldots,M-1\}; \quad (CS)$$
$$Y = \sum_{d=0}^{M-1} \underline{1}(D = d) Y^d; \qquad (SUTVA)$$
$$X^d = X, \qquad \forall d \in \{0,\ldots,M-1\}. \qquad (EXOG)$$

- **Conditional Independence Assumption (CIA):** No features other than X jointly influence treatment and potential outcomes within the range of interest (χ).

- **Common Support Assumption (CS):** Every value within χ allows for the observation of all treatments.

- **Stable-Unit-Treatment-Value Assumption (SUTVA):** The observed treatment value is independent of the treatment allocation of other individuals, ruling out spillover and treatment size effects.

- **Exogeneity Assumptions (EXOG):** The observed values of X are not dependent on the treatment status, thereby ruling out any causal effect of D on X.

Annex C: Finding an estimator for IATE

When all identification assumptions hold, IATE can be also expressed as:

$$IATE(m, l; x) = \mu_{m(x)} - \mu_{l(x)};\ \forall \mathrm{x} \in \chi, \forall m \neq l \in \{0, \dots, M-1\}$$

By denoting the conditional expectations of Y given X in the subpopulation $D = d$ by $\mu_d(x)$

- An easy-to-implement estimator involves separately estimating two conditional expectations using standard ML tools and then taking the difference.
  - The disadvantage of this approach is that standard ML methods prioritize maximizing out-of-sample predictive power for each estimator separately.
  - Using methods like Random Forest can lead to variability in the estimated treatment effects, especially when features are highly predictive of Y, but the treatment effects are relatively constant.
  - Unequal treatment shares* can also cause issues, with the forest for one treatment being finer than the other due to differences in sample sizes.

- An alternative approach involves using the same splitting rules for both subsamples, estimating $\mu_m(x)$ and $\mu_l(x)$ separately and then finding a plausible splitting rule for a 'joint' forest.

*: the proportion of individuals receiving treatment

References

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523), 1228-1242.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

Lechner, M. (2019). Modified causal forests for estimating heterogeneous causal effects.

Cookiecutter Data Science
A logical, reasonably standardized, but flexible project structure for doing and sharing data science work.
https://drivendata.github.io/cookiecutter-data-science/

Athey S. Solving Heterogeneous Estimating Equations Using Forest Based Algorithms
https://www.youtube.com/watch?v=CPz0HdUM3dE