Using Machine Learning to prepare CPCD data for RIW regression

36th PCSI Conference, Slovenia

Koffi Kpelitse, Rachel Zhang, Victoria Zhu,
Suren Rathnayake, and Sheril Perry

Canadian Institute for Health Information

May 30th, 2024

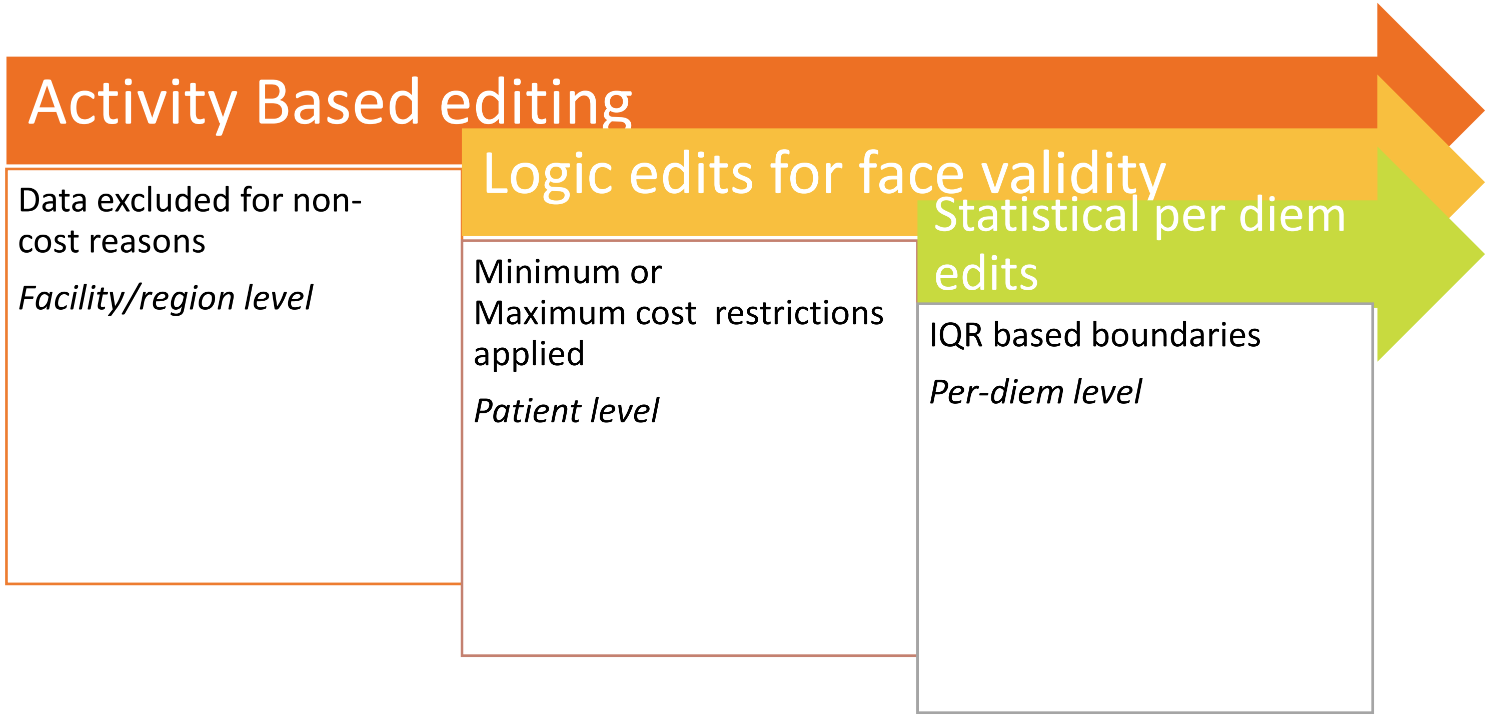
Casemix@cihi.ca cihi.ca @cihi_icis

CIHI

Outline

- Background: from cost to RIW production
- Issues and concerns
- New ML approach
- Outcome evaluation
- Next steps

Background: before RIW production starts…



Activity Based editing

Data excluded for non-cost reasons

Facility/region level

Logic edits for face validity

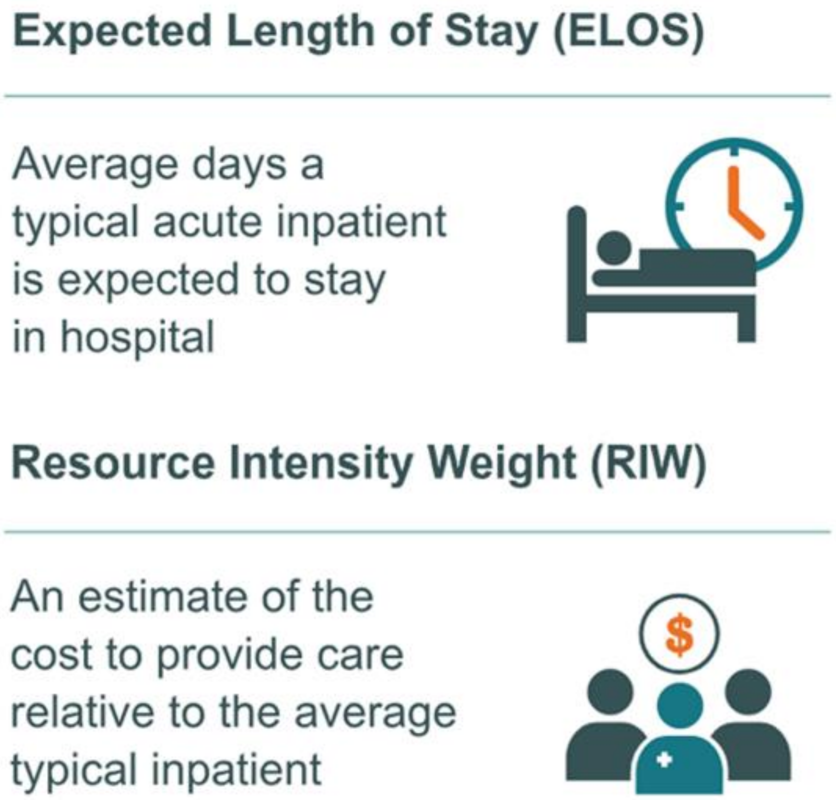
Minimum or Maximum cost restrictions applied

Patient level

Statistical per diem edits

IQR based boundaries

Per-diem level



Expected Length of Stay (ELOS)

Average days a typical acute inpatient is expected to stay in hospital

Resource Intensity Weight (RIW)

An estimate of the cost to provide care relative to the average typical inpatient

Current methodology: Issues and concerns

Activity data based

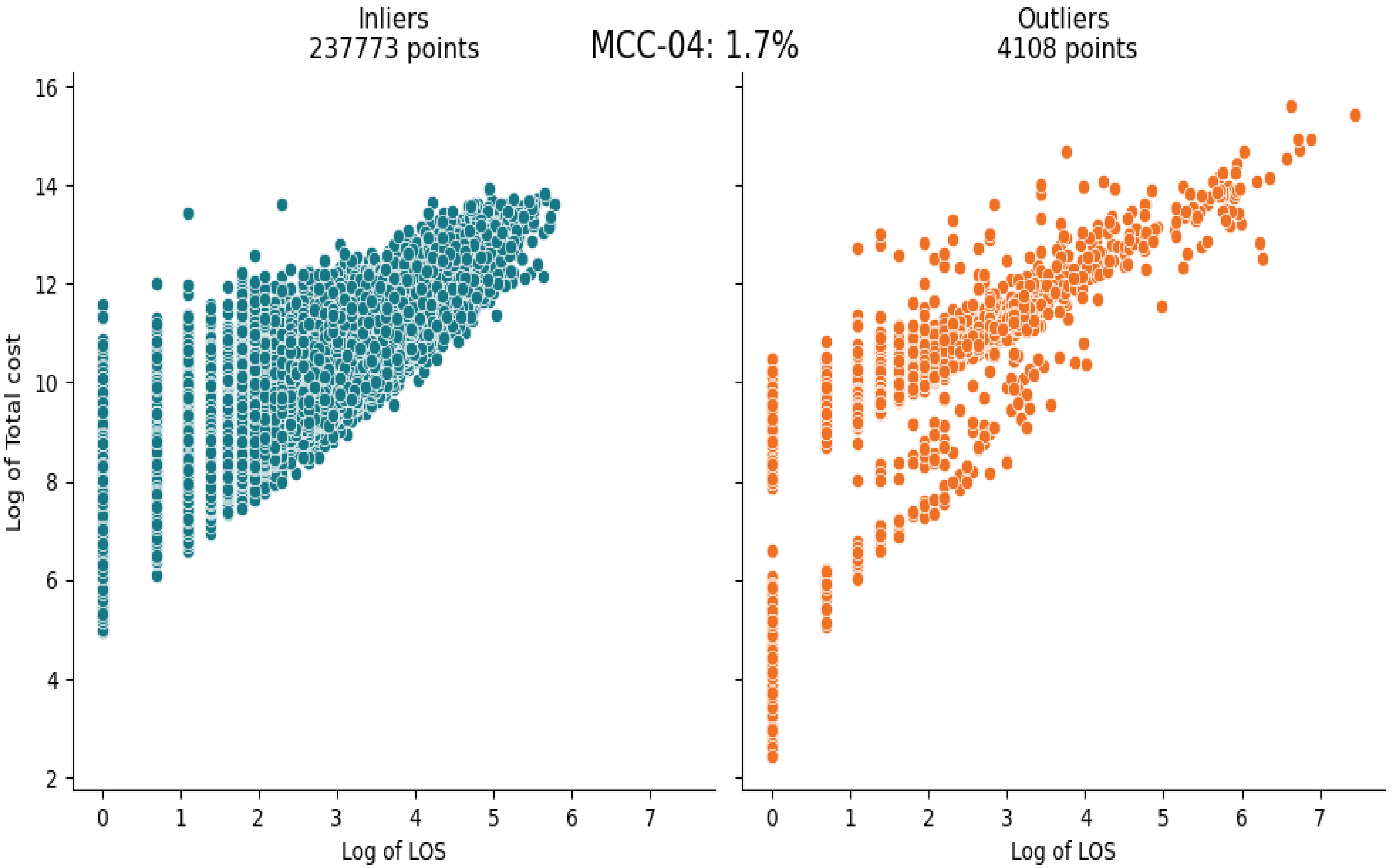
Logic edits for face validity

- Arbitrary high/low boundaries were set up based on historical findings
- Traditional methods are time-consuming

Statistical per diem edits

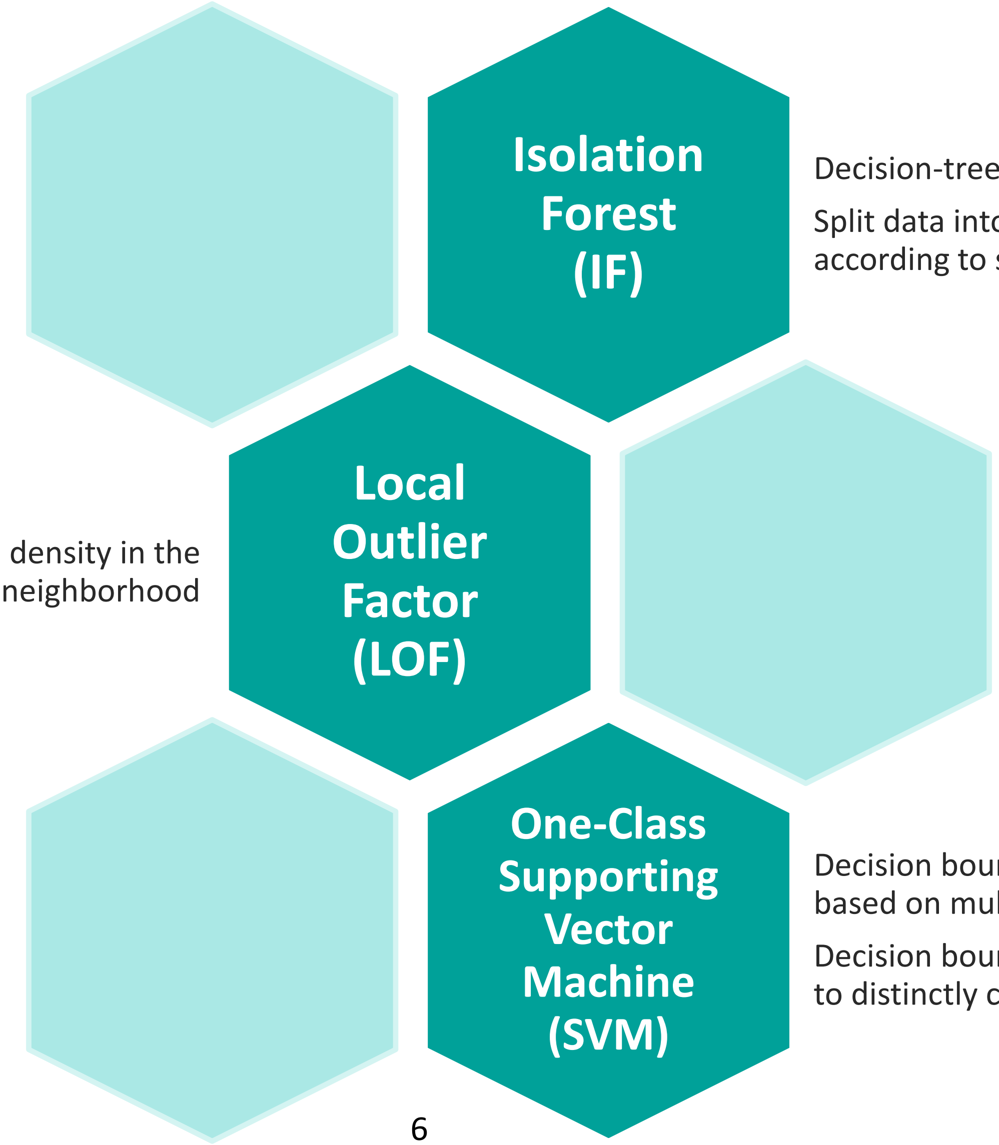
- Correlation between LOS and total cost
- Statistical distributions vary across groups
- High volume of outliers
- High impact to low volume CMGs

CMG+ Inliers and outliers in current approach



- Inliers and outliers have similar distribution patterns
- All cases at the higher end of cost and LOS distributions are excluded

Explore unsupervised ML approaches



Isolation Forest (IF)

Decision-tree-based algorithm

Split data into sub-samples according to some features

Local Outlier Factor (LOF)

Focus on the density in the neighborhood

One-Class Supporting Vector Machine (SVM)

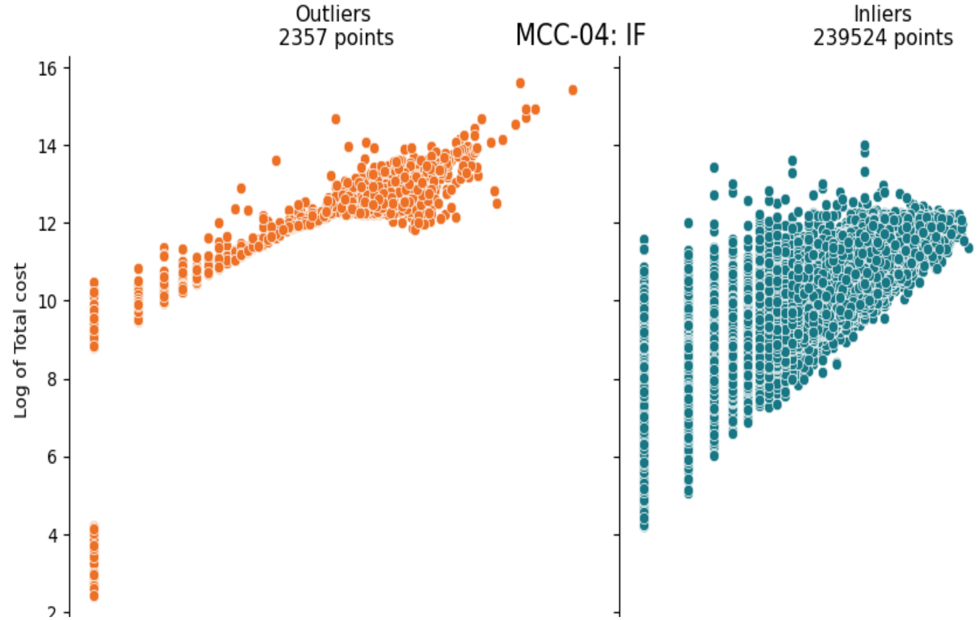
Decision boundaries are set up based on multiple dimensions

Decision boundaries are used to distinctly classify data points

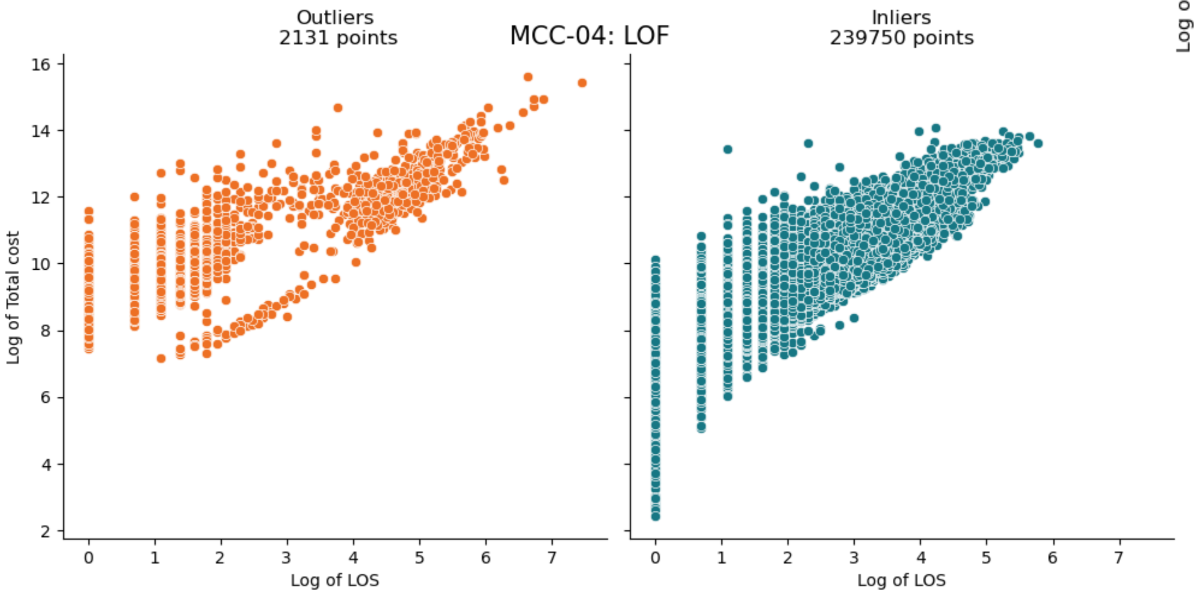
CMG+ costs cleaning: Model specifications

- **Analysis performed at the MCC level instead of CMG**
  - Outliers rate set to 1% and 2%
- **3 set of features tested**
  - Model 1: LOG total cost
  - Model 2: LOG total cost + LOG acute LOS
  - Model 3: LOG Total cost, LOG acute LOS + cost per diem
- **Final features selected**
  - Model 3 and outlier rate of 1%

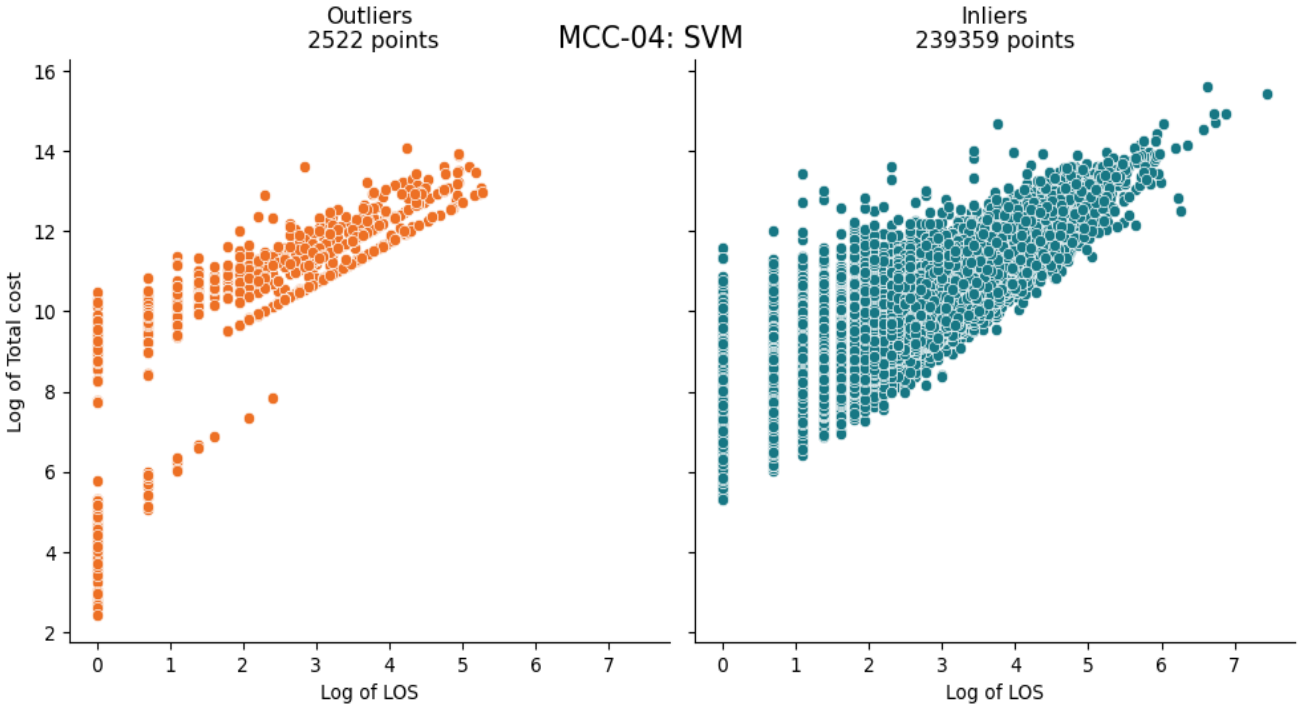
Preliminary findings – comparing 3 methods



MCC-04: IF



MCC-04: LOF



MCC-04: SVM

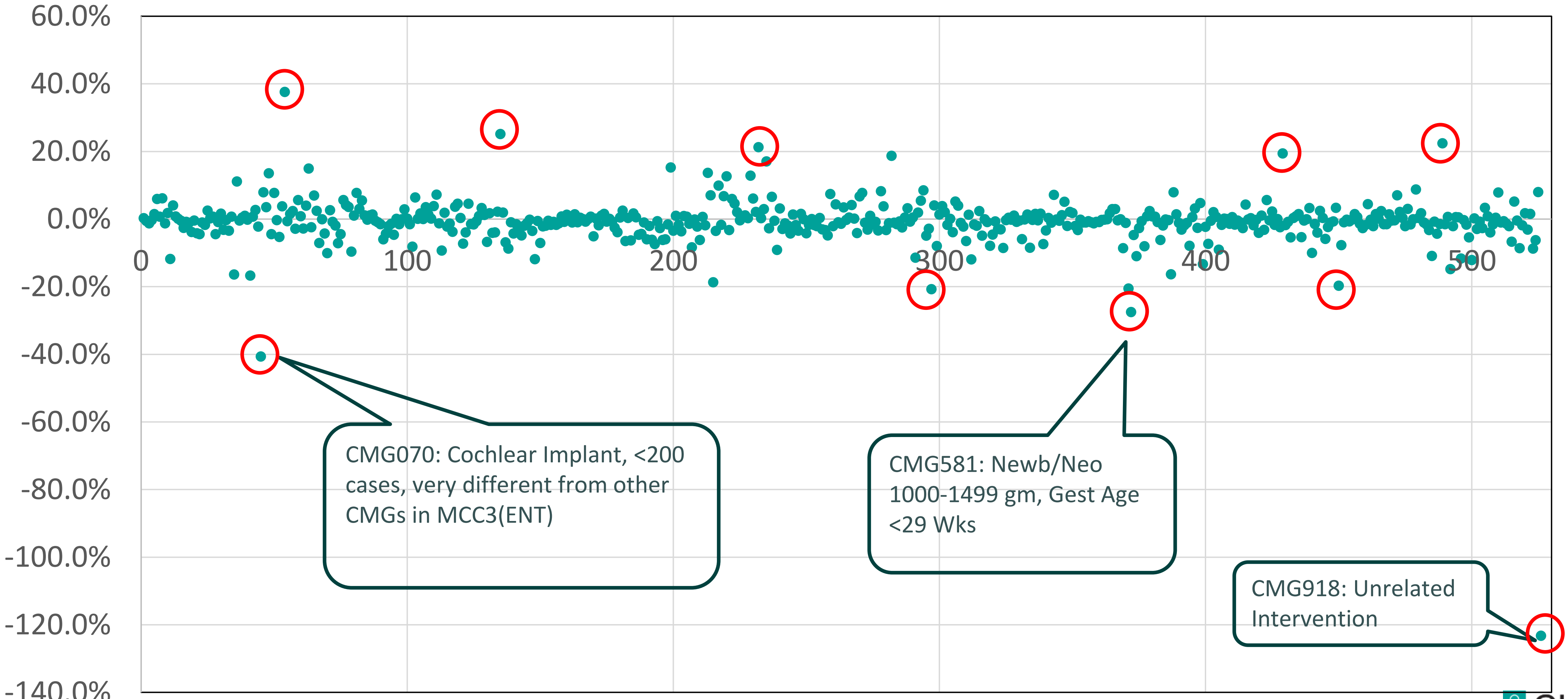
MCC 04: Respiratory

Impact on CPCD data used for RIW

- **0.36 % fewer cases identified as outliers**
- **By MCC, approximately 20-40% of outlier cases overlap**
- ***64% of outliers had 1 day stay in 2022 product, almost 73% have 1 day stay in new ML approach***

Data Set	**Volume**	**Actual Mean**	**Predicted Mean**	**Bias**	**MAE**	**R-Square**
2022 Production	2,107,864	9,218.57	9,233.03	-14.46	3,216	81.5%
MCC Only	2,114,406	9,216.81	9,222.53	-5.72	3,257	80.3%

Change in GOF by CMG



CMG070: Cochlear Implant, <200 cases, very different from other CMGs in MCC3(ENT)

CMG581: Newb/Neo 1000-1499 gm, Gest Age <29 Wks

CMG918: Unrelated Intervention

What we see on the journey…

CONTINUOUS LEARNING

EFFICIENCY & FLEXIBILITIES

OPPORTUNITIES

Canadian Institute for Health Information

Better data. Better decisions. Healthier Canadians.

cihi.ca

casemix@cihi.ca

@cihi_icis