Machine learning for imputing missing pharmacy costs in claims data

Presenter

Shiva Vojjala, MS

Senior Business Information Developer Consultant

Carelon Research

Team members

John Barron, PharmD

Aditya Kumar, MBA

Michael Grabner, PhD

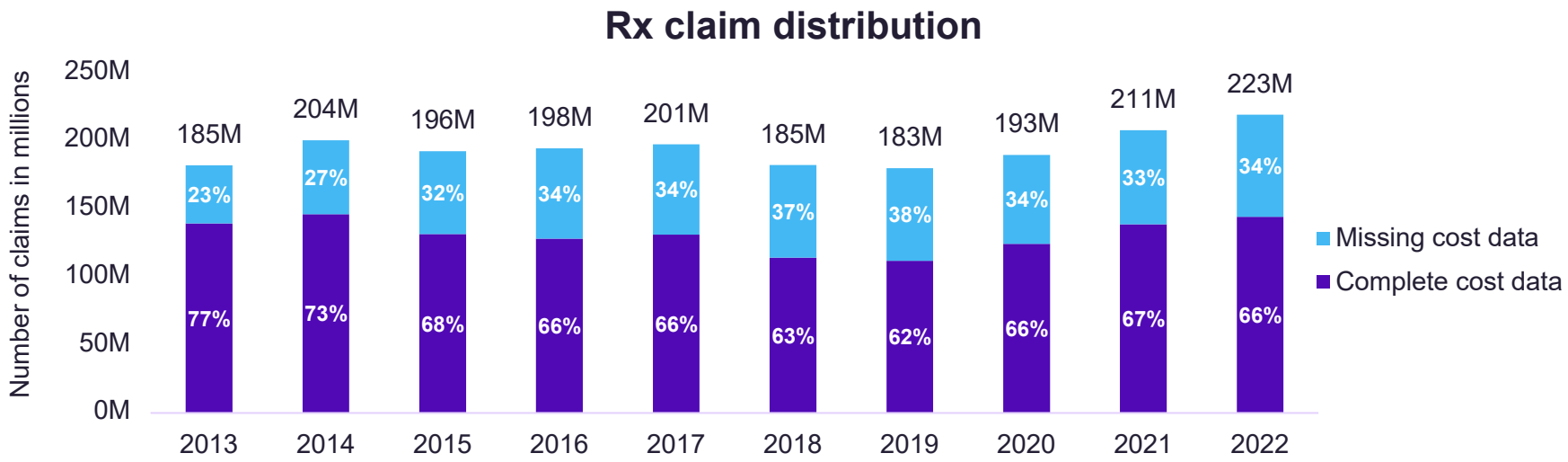
Biruk Eshete, MS

Hiangkiat Tan, MS

Vincent Willey, PharmD

Background

Missing data is an issue researchers often encounter at a database and/or study level. Particularly healthcare cost data can be missing or redacted.

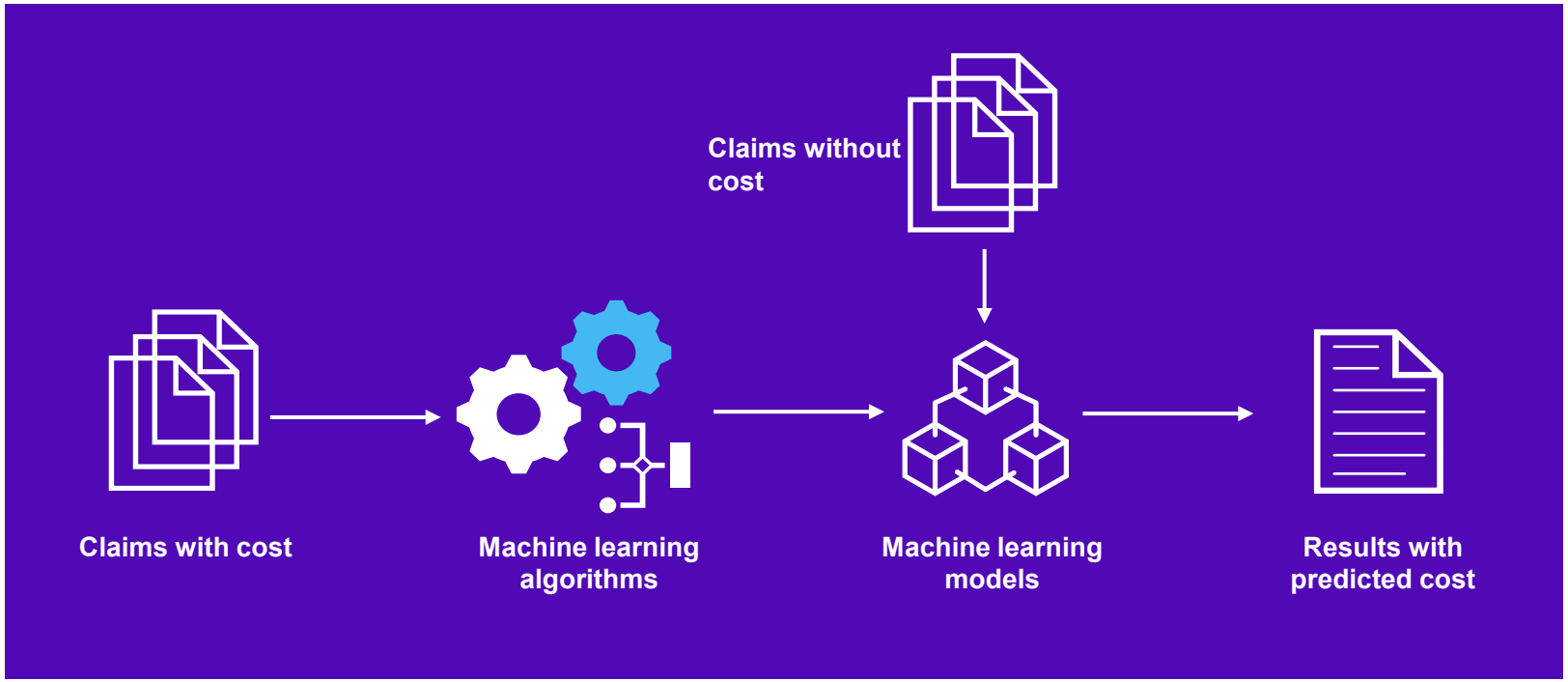


Data source: Carelon Research's Healthcare Integrated Research Database (HIRD) using researchable members with medical and pharmacy enrollment

Machine learning is a novel way to address this issue.

Objective

To impute missing pharmacy claims costs using the best-fitted model identified by developing and comparing several machine learning algorithms in a large US commercially insured population.



Data source

HIRD® – Healthcare Integrated Research Database

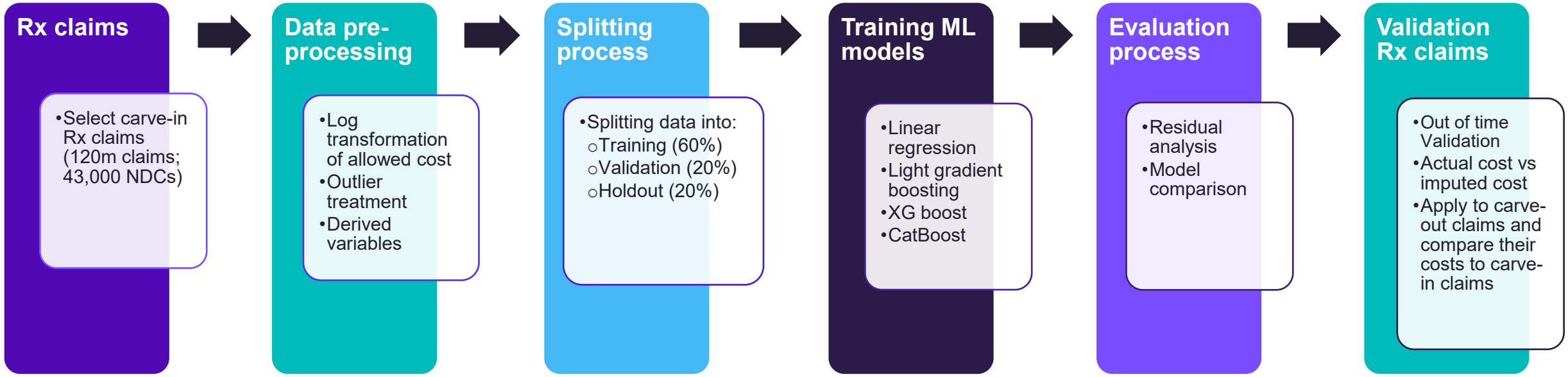
- Contains longitudinal claims and eligibility data from 2006 to the most recent quarter from a large national US payer.
- Commercial health plans in 14 states with members residing in all 50 states.
- Represents members with Commercial/Medicare Advantage/Supplement & Part D insurance.

Carve-out Rx claims

- Employer groups contract with health plans to provide health insurance packages to their employees.
- Some employers choose to “carve-out” pharmacy benefits to a third party, leading to missing cost data on the claims of affected members.

Analysis Time period – 1/1/2013 to 12/31/2021

Methodology overview



ML – Machine Learning

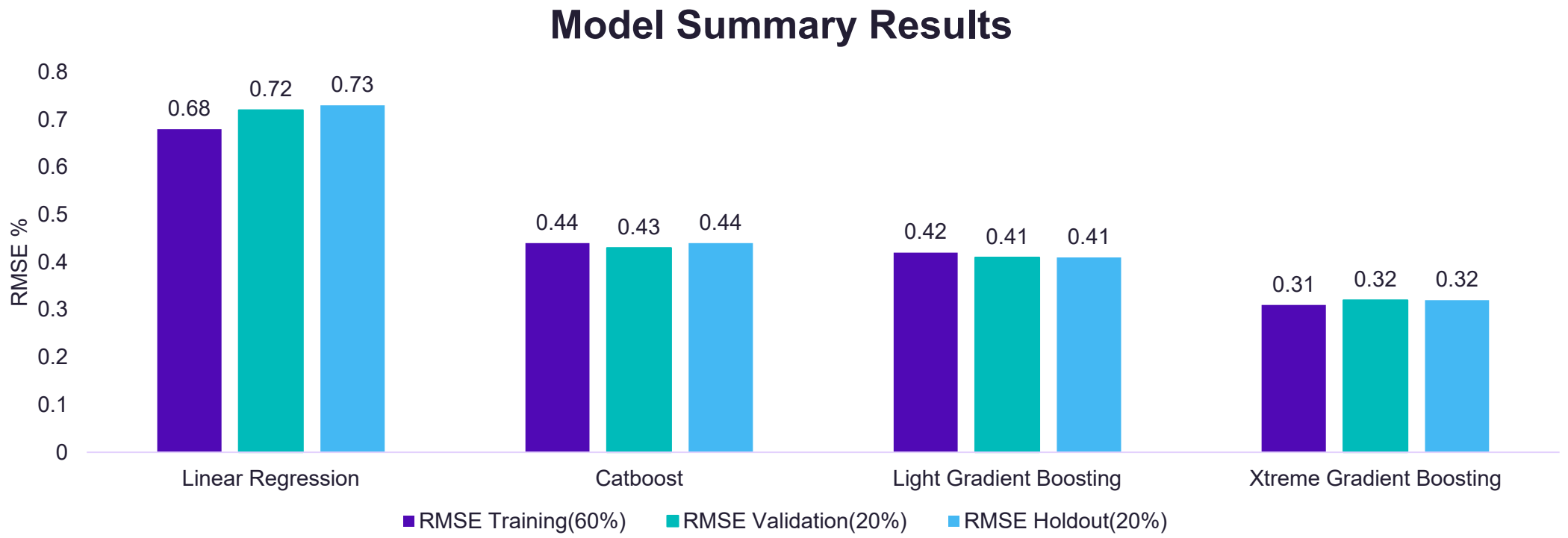
Predictors

Independent variables
NDC (National Drug Code)
Quantity dispensed
Wholesale unit price
Insurance account type (local/national)
Dispensing pharmacy (grouped)
Patient age on fill date
Insurance group type (large/small/individual)
Mail-order or retail
Month and year of the fill
Insurance product type (HMO/PPO/CDHP/Other)
Patient state of residence
Dispensing pharmacy state
Urban or rural – based on patient zip code
Refill or new fill
Number of all Rx claims until the fill date (Yearly)
Patient sex
NDC properties (e.g., route of administration, generic drug indicator)

Top Predictors

HMO: Health Maintenance Organization
PPO: Preferred Provider Organization
CDHP: Consumer Driven Health Plan

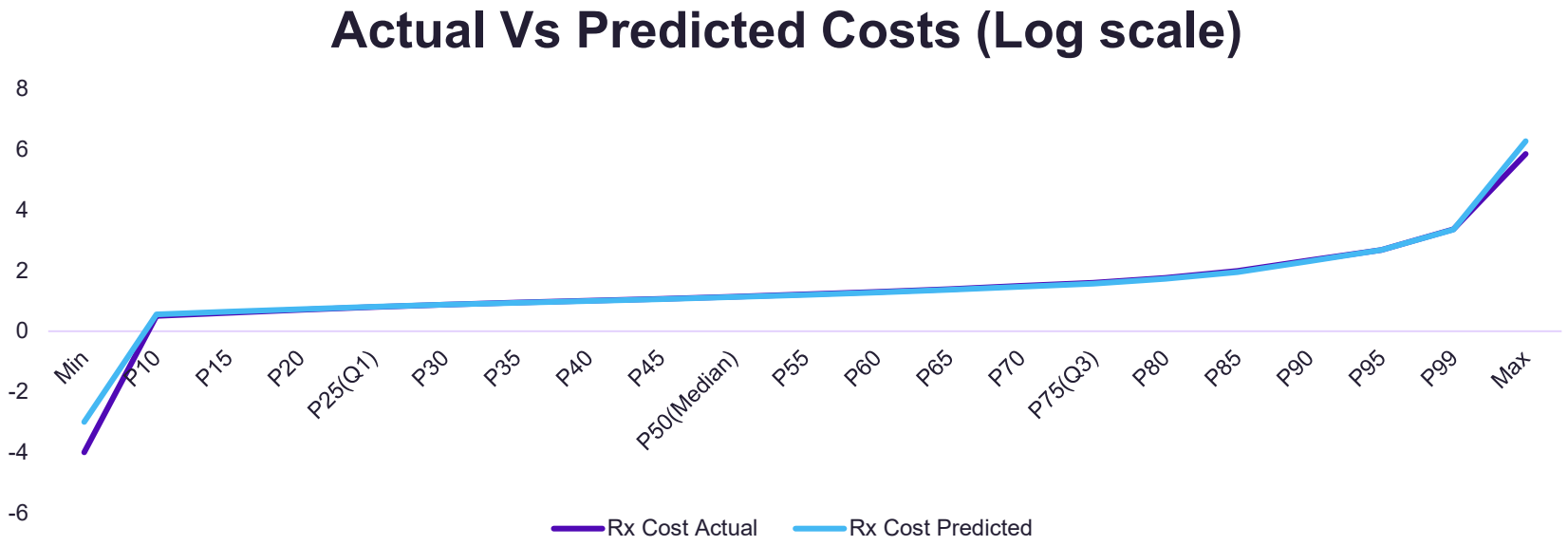
Model comparison



Evaluation measurement

RMSE - Root Mean Square Error: Measures the average difference between values predicted by a model and the actual values; can range from 0 to ∞; lower numbers are better

XGBoost – Validation



The mean cost differences between the imputed and actual costs were:

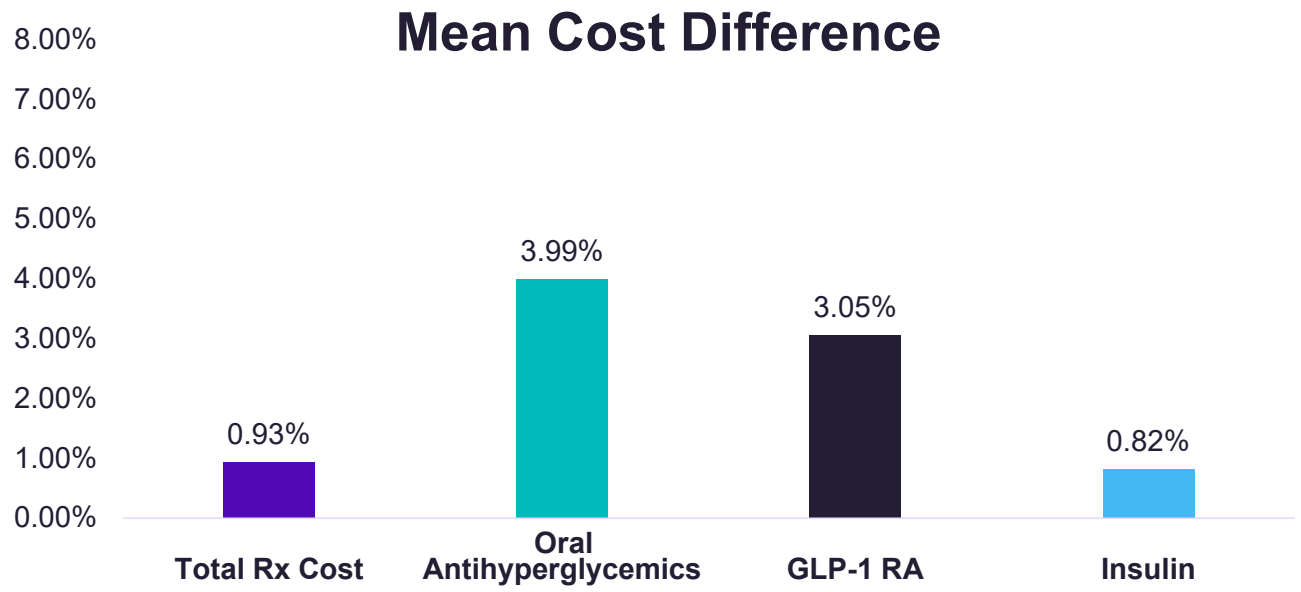
- less than $10 across over 90% of low cost (<$100) NDCs
- less than 25% across over 90% of medium (≥$100) and high cost (≥$1,000) NDCs

Case study validation #1

Cohort: Diabetes medication users

- Members with ≥12 months of continuous medical and Rx enrollment following index date, defined as first fill of a diabetes medication between 01/01/2013 - 12/31/2021

Cohort Size		
Using carve-in Rx claims only	Combining with carve-out	Increase
79,312	145,006	83%



GLP-1 RA: Glucagon-like Peptide-1 Receptor Agonists

Case study validation #2

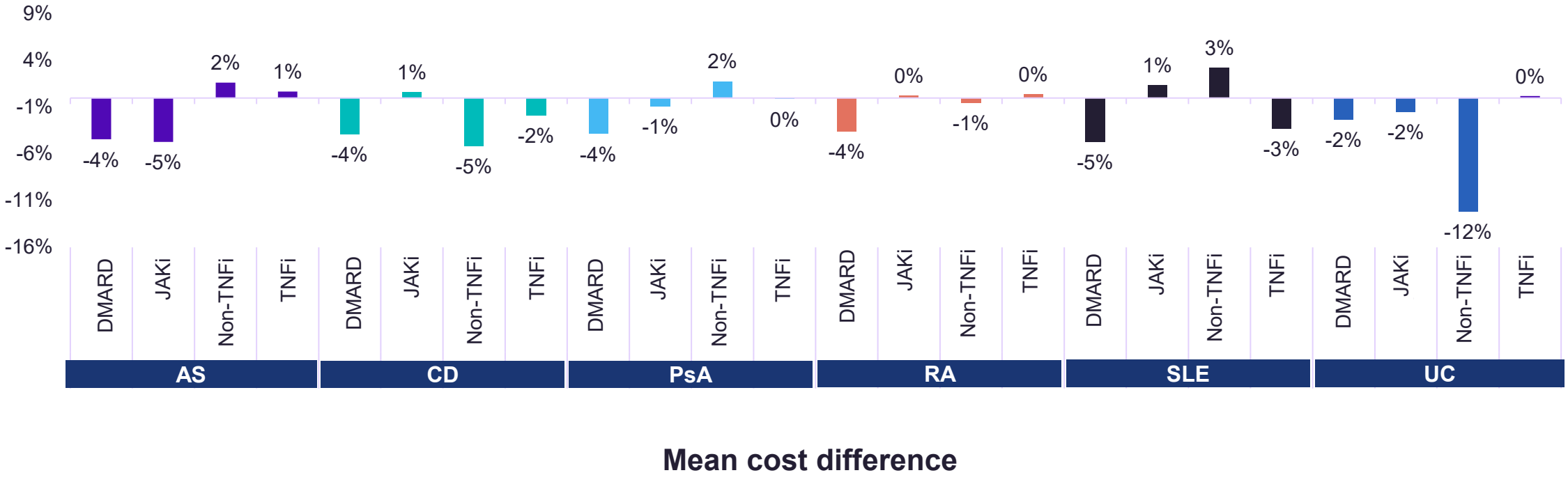
Cohorts: Autoimmune diseases

- Ankylosing spondylitis (AS)
- Crohn's disease (CD)
- Psoriatic arthritis (PsA)
- Rheumatoid arthritis (RA)
- Systemic lupus erythematosus (SLE)
- Ulcerative colitis (UC)

Members with ≥12 months of continuous medical and Rx enrollment pre and post-index date, defined as first diagnosis during 01/01/2013 - 12/31/2021

Cohort size			
Diseases	Using carve-in Rx claims only	Combining with carve-out	Increase
AS	4,069	6,215	53%
CD	9,370	14,584	56%
PsA	14,837	21,947	48%
RA	57,268	80,378	40%
SLE	14,776	21,112	43%
UC	7,982	12,271	54%

Case study validation #2



Abbreviations: ankylosing spondylitis (AS); Crohn's disease (CD); psoriatic arthritis (PsA); rheumatoid arthritis (RA); systemic lupus erythematosus (SLE); ulcerative colitis (UC); disease-modifying anti-rheumatic drugs (DMARD); janus kinase inhibitors (JAKi); tumor necrosis factor inhibitors (TNFi)

Limitations

Imputation quality is negatively affected in certain cases:

- Medications with limited data, including those new to market or with infrequent use.
- Medications with large variability in allowed amounts or predictors (e.g., quantity dispensed).
- Errors in predictor values (e.g., quantity dispensed = 999).

Only total costs can be imputed (unable to distinguish between plan-paid and patient-paid amounts).

Conclusion

Machine learning techniques provided reasonable estimates for missing pharmacy costs using real-world data.

- This approach will allow for use of data that previously could not be utilized for healthcare cost analyses.
- Will increase the quality of research and allow for more robust healthcare cost analysis across different therapeutic areas – improving representativeness and precision of estimates

Next steps:

- Developing separate models for NDCs with larger than optimal differences between imputed and actual values to improve performance.
- Flagging those NDCs with “suboptimal differences” so researchers can take steps to mitigate the impact and/or interpret with caution.
- Identifying more data preprocessing steps to minimize the impact of issues in predictor values.
- Creating new models every year to capture the up-to-date cost trends and new NDCs.

Q&A

Thank you!

Shiva Krishna Vojjala
shiva.vojjala@carelon.com