A Method for Assigning Weights to Variable Matching in Record Linkage

Salam Abdus, Steven C. Hill, Marc Roemer

Agency for Healthcare Research and Quality

The findings and conclusions in this presentation are those of the authors and do not necessarily represent the views of the Department of Health and Human Services or the Agency for Healthcare Research and Quality.

Research Question

- For MEPS sample members without pharmacy data from their own pharmacies, details about their prescription drug purchases are imputed from other sample members' pharmacy data.
- Can we improve the matching weights used in the imputation process?
- Using a validation sample, we compare predictive accuracy for current matching methods and alternative methods.

Outline

- Sources of MEPS drug data
- Data matching and imputation in the MEPS drug editing process
- Validation study design
- Use regression coefficients as alternative match weights
- Predictive accuracy of current and alternative methods

Household Component (HC)

For each sample member in each round, MEPS asks:

- Names of drugs obtained
  - Typically typed into CAPI from pill bottle labels
- Number of times each drug was obtained
- Pharmacy names
- Permission to contact pharmacies

Pharmacy Component (PC)

For each fill or refill, responding pharmacies provide:

- National Drug Code (NDC), drug name, dosage form, strength
- Quantity (for example, # pills, ml)
- Payers
- Payments

Matching Household and Pharmacy Data for 2015 MEPS

- 75.4% of household-reported drugs were for people with pharmacy data
  - Among these drugs, 80.3% matched to the person's own pharmacy data
- Imputed vector of pharmacy data for 39.4% of drugs
  - Donor pool is all drugs reported by all responding pharmacies

Multiple Match Attempts to Impute PC Data to HC, 2015

Match (Imputation) Type	Number	Percent
Same drug	52,780	91.4
Active ingredient, dosage form, and strength	36,336	63.0
Active ingredient, dosage form	8,010	13.9
Active ingredient	8,431	14.6
Same therapeutic subclass	1,798	3.1
Same therapeutic class	1,318	2.3
Same therapeutic group	1,228	2.1
Other	595	1.0

Current Imputation Method: Conditional on the Same Drug, Impute PC Data from a Similar Person

- Similarity based on potentially important characteristics
  - Drug name
  - Insurance / sources of payment
  - Names of Pharmacies
  - Geography
  - Months per fill
  - Demographics, health status and conditions
  - Cumulative fills in current and prior rounds in CY
- Currently imputed by matching score using weighted agreement of characteristics
  - Most characteristic weights developed in 1990's
  - Some weights modified starting with 2008 data

Use Validation Sample to Assess Matching Weights

- Use sample persons with their own PC data to identify true matches
- Create data set with all potential matches (all possible pairs of HC/PC person-drug pairs) as if the person did not have PC data
- How similar is the imputed drug match to the true match:
  - Total price per fill
  - Out-of-pocket spending
  - Patent status

Data Set of Potential Donors

- 2015 Household and Pharmacy data
- 54,443 recipients for matching on ingredient, dosage form and strength
- 13.5 million recipient-potential donor pairs
- Split into 2 samples A & B:
  - Run regressions on A and predictions on B
  - Run regressions on B and predictions on A
  - Average predictions

4 Sets of Regressions Coefficients to Use as Alternative Matching Weights

- Outcomes:
  1. Total payments
  2. Square root of total payments
  3. Out-of-pocket payments
  4. Square root of out-of-pocket payments
- OLS regressions on characteristics
  - ► All characteristics except pharmacy & drug name
  - ► Drug and health condition fixed effects
- Predict payments for recipient and potential donors
- For each recipient take the closest donor (absolute difference in predicted payments)

5th Set of Regression Coefficients to Use as Alternative Matching Weights

- Outcome: True match
  - Add the true match to the sample as a potential donor
  - Indicator for true match v. other potential donors
  - Linear probability model
- Explanatory variables are match/similarity of characteristics between donor and recipients
  - Indicators for whether the recipient and potential donor have the same characteristics
  - Match scores for health conditions, drug names, and pharmacy names
- Coefficients directly predict best match among potential donors

Comparing Predictive Accuracy for Matched Donors on the Validation Sample

- Overall bias: mean prediction error =
  - Mean (true − imputed payments)
  - Total, out-of-pocket payments
- Accuracy for each observation: mean absolute prediction error =
  - Mean absolute (true − imputed payments)
  - Total, out-of-pocket payments
- Accuracy & precision: Lin's concordance correlation coefficient between true and imputed payments
  - Total, out-of-pocket payments
- % with same patent status as the recipient

Predictive Power (Adjusted R^2) for regression-based models

- Total payments
  - Linear .642
  - Square root .716
- Out-of-pocket payments
  - Linear .337
  - Square root .388
- True Match: .189

Predictive Accuracy for Total Payments

Match Method	Mean		Lin's concor-dance
	Error	Absolute Error	
Current weights	−$0.3	$64.9	0.536
Regression-based weights			
Total Payments	−$5.1	$65.4	0.617
SqRt of Total Payments	−$5.1	$67.1	0.386
Out-of-pocket Payments	−$4.2	$66.2	0.677
SqRt of Out-of-pocket Payments	−$3.4	$67.5	0.624
True Match	-$2.1	$65.3	0.627

Predictive Accuracy for Out-of-Pocket Payments & Patent Status

Match Method	Mean		Lin's concor-dance	Patent Status Agree
	Error	Absolute Error		
Current weights	−$0.5	$12.6	.125	98.2%
Regression-based weights				
Total Payments	−$0.0	$14.9	.139	97.8%
SqRt of Total Payments	−$0.3	$14.9	.085	97.7%
Out-of-pocket Payments	−$0.5	$12.9	.185	97.7%
SqRt of Out-of-pocket Payments	$0.2	$13.4	.155	97.8%
True Match	−$0.4	$14.0	.164	97.9%

Conclusions

- All methods produce good predictions
- Current imputation weights yield the best predictions on several measures, but not on Lin's concordance.
- Expenditure regression methods may not live up to their potential due to overfitting to less common health conditions and drugs
- Expenditure regression methods might perform better if they could incorporate name matching

Further Research

- Remaining match attempts may yield different predictive results
- Can we better optimize weights across multiple targets and multiple measures of accuracy?
  - When we get to the lower-rung matching attempts, more outcomes are relevant, for example, can we predict the right drug?