A Method for Assigning Weights to Variable Matching in Record Linkage

Salam Abdus[1], Steven C. Hill[1], Marc Roemer[2]

[1]Center for Financing, Access, and Cost Trends, Agency for Healthcare Research and Quality, 5600 Fishers Ln, Mailstop 07W41A, Rockville MD 20857
[2] Centers for Disease Control and Prevention, National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782,

Acknowledgements. This research was conducted while Marc Roemer was employed at AHRQ, prior to becoming an employee at Centers for Disease Control and Prevention (CDC). The views expressed in this paper are those of the authors, and no official endorsement by the Department of Health and Human Services, the Agency for Healthcare Research and Quality, or CDC is intended or should be inferred.

Introduction

The Medical Expenditure Panel Survey (MEPS) is a unique, nationally representative source of micro data on health care use and expenditures for all payers. Household respondents report drugs and the number of times each drug was obtained, while a follow-back survey of pharmacies is the primary source of price and expenditure data. High quality data are critical because the MEPS is widely used for national estimates, behavioral modeling, and policy simulations, including many analyses of prescription drug markets. Ongoing efforts to ensure data quality seek to maintain the currency of the data in rapidly changing pharmaceutical markets (Hill et al. 2011). This project investigates potential improvements in the methods used to match or impute pharmacy data to the drugs reported by households, specifically for household-reported drugs for persons lacking pharmacy data. In particular, we examine a range of regression-based methods of deriving matching weights for imputing pharmacy data. Using a validation sample, we compare predictive accuracy for current and alternative matching weights.

Medical Expenditure Panel Survey Prescription Medicine Data

The MEPS combines two overlapping panels of the U.S. noninstitutionalized civilian population to produce annual estimates (Cohen 1997). Each panel uses five in-person interviews to collect two calendar years of data. During each interview, a single informant typically reports for all household members, regardless of age, and the average recall period is five months. The MEPS asks that this person be the family member most knowledgeable about health and health care use in the household. Respondents report the names of the drugs obtained, the number of times each drug was obtained, and the health conditions associated with the drug. The respondent may consult records such as medicine bottles or receipts, so the reported medication name is often quite specific. However, the information can also be minimal, for example, “pain pills.” When drugs are reported, the drug names are entered in a dynamic roster. The names, addresses, and types of pharmacies that filled each household member’s prescriptions are also requested, along with permission for MEPS to acquire data from these pharmacies.

The Pharmacy Component (PC) of MEPS collects data from the pharmacies for which household sample members signed permission forms. Each pharmacy is asked to provide patient profiles or information about each prescription filled or refilled for each patient during the calendar year. For each fill or refill of a drug, the following information is requested: National Drug Code (NDC), quantity (for example, number of pills), payers, and payments. If the pharmacy does not provide the NDC, the PC asks instead for the medication name, dosage form, strength, and strength unit.

Two general approaches accomplish matching the data reported by pharmacies to the data reported by households. First, for each of a person’s drugs in the HC, an attempt is made to find the same or similar drug obtained by that

person in the PC.[1] In the MEPS data for calendar year 2015, 75.4% of household-reported drugs were for people with pharmacy data. Among these drugs, 80.3% were matched to the person's own pharmacy data.
This project focuses on the imputation process, which is the second approach used when the first match approach fails. Imputation was used for 39.4% of drugs reported by the household; the imputation recipient is the household-reported drug. The imputation process matches a vector of drug details from pharmacy data reported for some other person. The donor pool is all pharmacy-reported drugs, regardless of the persons who obtained them; potential donors are the pharmacy responses combined with the characteristics of the person for whom the pharmacy data were collected. The imputation procedure seeks, for each recipient, a donor with similar characteristics, especially characteristics that could affect the price, out-of-pocket cost, and patent status (brand name or generic) of the drug obtained. These include drug name, months per fill (to better match the number of pills), pharmacies used, personal characteristics (types of health insurance, geography, health status, conditions, demographics), and cumulative number of fills in current and prior rounds, which may be roughly related to out-of-pocket spending due to reaching deductibles and out-of-pocket maxima. In the imputation procedure, these characteristics are assigned matching weights. Each potential donor is assigned a score based on whether the donor and recipient share the same value for each characteristic and the matching weight for each characteristic. The donor is the pharmacy-reported drug with the highest score.

The imputation process relies on seven match attempts, where the initial attempt requires exact agreement on many details about the drug, and subsequent match attempts require progressively less agreement on details. . The first attempt requires an exact agreement on the drug's active ingredient, dosage form, and strength. These drug characteristics are coded in Wolters Kluwer's proprietary Generic Product Identifier (GPI), a 14-digit code which identifies pharmaceutically equivalent drugs, such as a brand name drug and its generic competitors. For household-reported drugs, professional coders identify as many digits of the GPI as possible based on the medication name and any other information appended to the name provided by the respondent. Typically, active ingredient and dosage form can be coded for household-reported drug names. In the 2015 calendar year data, among the household-reported drugs eligible for imputation, 63.0% were matched on the first match attempt. The second attempt requires an exact match on active ingredient and dosage form (12-digit GPI). In the 2015 calendar year data, 13.9% of drugs eligible for imputation were matched on the second match attempt. This paper focuses on these first two attempts. Additional information about the MEPS prescription drug data is in Hill et al. (2014).

Validation Study Design

We use a sample with known true matches, specifically 54,443 household-reported drugs from the 2015 MEPS that matched to the person's own pharmacy data. We then create the data sets of potential donors for each recipient as if the person did not have their own pharmacy data, that is, all potential matches for the same drug from different people with pharmacy data. For the first attempt of matching, where exact matching of drug is required at the GPI-14 level, we get a dataset with 26.1 million potential donors, (because each record of pharmacy data could match to any recipient with the same drug). We reduce this data set to 13.5 million records by removing records that do not provide variation relevant to matching or provide minimal variation. Specifically, we removed multiple reports of the same drug for a person, potential donors with identical match variables, and recipients with only one potential donor. For the matching at the GPI-12 level, we get a dataset with 56.0 million potential donors (since the matching is at a lower level, we get more potential donors). We reduce this data set to 12.3 million observations, by randomly selecting half of the recipients and by removing donors that do not provide variation relevant to matching or provide minimal variation.

We randomly split the recipients and their donors into two samples, A and B. Our alternative matching methods are based on regression analyses, and the split sample allows us to test out-of-sample predictive accuracy, which reduces the risk that overfitting on the estimation sample would yield false conclusions about the predictive

[1] More specifically, person-round-drugs are the unit for matching, so the fills in the PC data are aggregated to the person-round-drug level in order to mirror the structure of the HC data. After matching drug name to drug name, the HC and PC drug records are expanded into acquisitions. Each drug name reported in the HC interview is fanned out to the number of fills the household reported. Then each HC fill is paired with a PC fill within the drug-to-drug matched set. If the number of fills differs between the HC and PC, then the number of acquisitions is determined by the HC and some randomization is used to allocate the PC fills to the HC fills.

accuracy of the regression-based approaches.[2] We estimate regressions on sample A and predictions on sample B, and then regressions on sample B and predictions on sample A. We average predictive accuracy across the two samples.

For each imputation method, we assess predictive accuracy in terms of total expenditures for the fill, out-of-pocket spending for the fill, and patent status, that is, brand name or generic drug. We assess predictive accuracy for expenditures with three measures. Overall bias is measured as the mean prediction error between expenditures for the true match and imputed expenditures. We assess accuracy for each observation as the mean absolute prediction error. We also estimate Lin's concordance correlation coefficient between the true and imputed expenditures (Lin 1989). This measure combines accuracy and precision and ranges from −1 to +1, where a larger positive value indicates better accuracy and precision. We also measure the agreement rate of patent status on the true match and imputed match.

Six Methods of Deriving Matching Weights

We use regression methods to develop five alternative sets of matching weights to compare with the current set of matching weights.

Current weights. The weights assign positive points when the recipient and potential donor have the same value for a variable. Scoring systems are used for 3 other variables: drug names, pharmacy names, and health conditions. Drug name helps to match brand name donors to receipts who used the brand name when reporting the drug. Drug and pharmacy names are matched using a Soundex function to account for misspelling and other potential data errors. Up to six health conditions associated with drugs are matched, with conditions coded using the Clinical Classification Software (Elixhauser et al. 2000).

Most of the current set of matching weights was created in the late 1990's. Some modifications were made starting with the 2008 data after testing incremental changes in the weights on selected variables and manually reviewing the resulting matches. Less weight was put on medication name in order to account for households reporting brand names when referring to the generics they purchased. An indicator for using mail-order pharmacies and additional insurance-related variables were added. More weight was put on insurance variables.

True Match Regression. We use regression analysis to determine the relative importance of match variables in distinguishing true matches from non-matches. For this regression, each record is a potential donor, plus a record for the true donor. The outcome variable is an indicator for the true donor. Explanatory variables are the same as those used in current matching methods. The coefficients reflect the relative importance of each characteristic in selecting a good match from the pool of potential donors, so the potential donor with the highest match score is the best match.

Expenditure Regressions. Correctly measuring drug expenditures is an important goal of the MEPS. Weighting characteristics according to their ability to accurately predict payments for drugs would facilitate meeting that goal. We estimated four regressions, each with a distinct outcome variable:

1. Total payments
2. Square root of total payments
3. Out-of-pocket payments
4. Square root of out-of-pocket payments

We use square root transformations, because they have been found to predict payments well for other types of health care events. Using the sample of potential donors, we regress each outcome on the characteristics currently used in matching, with two exceptions, pharmacy name and drug name, for which it would be difficult to create explanatory variables. We include fixed effects for each drug, so that the coefficients reflect the effects of the characteristics after requiring exact agreement on drug. We also use fixed effects for the pharmacy chains, and fixed effects for conditions coded using the Clinical Classification Software. The drug name character variable, however, could not be parsimoniously included in these regressions. Using the coefficients, we predict expenditures for the recipient and all the potential donors. For each of the four regressions, the best match is the potential donor with predicted expenditures closest to the predicted expenditures of the recipient (minimum absolute predicted error).

[2] When a regression fits the idiosyncrasies of a test sample so well that it predicts poorly out-of-sample, especially yielding extreme predicted values, then it is said to overfit the data.

Results

Table 1 shows the predictive accuracy of the six imputation methods for total payments when drugs are matched by active ingredient, dosage form, and strength (i.e., at the GPI-14 level). The mean prediction error for total payments using the current match weights was less than a dollar (-$0.3). Mean prediction errors for total payments using the other five methods were larger than that from the current weights, but still were relatively small, ranging from -$2.1 with match weights from the true match regression to -$3.7 with weights from the total payments regression. Match weights from the total payments and the square root of total payments regressions yielded lower mean absolute prediction errors for total payments than the other four methods, $63.4 and $62.1, respectively. The mean absolute prediction errors for total payments from other methods were, however, not much higher, ranging from $64.9 in case of current weights to $68.0 with match weights from the square root of out-of-pocket payments regression. Lin's concordance correlation coefficients, which measure accuracy and precision, differed for total payments, ranging from 0.536 using the current weights to 0.734 using weights from the square root of total payments regression.

Table 2 shows the predictive accuracy of the six imputation methods for out-of-pocket payments and for patent status when drugs are matched at the GPI-14 level. The mean prediction errors for out-of-pocket payment were less than a dollar (negative or positive) in all six imputation methods. The mean absolute prediction error was the lowest using the current match weights, at $12.6. Mean absolute prediction errors in other imputation methods were also not very different, ranging from $12.8 when using weights from the of-pocket payments regression to $15.3 when using weights from the square root of total payments regression. Lin's concordance correlation coefficients were similarly low across all six methods, ranging from 0.094 using weights from the square root of payment regression method to 0.179 using weights from the out-of-pocket payments regression method. All regression methods predicted patent status quite well, ranging from 97.7% agreement with the four expenditure regression methods to 98.2% agreement using current weights.

Tables 3 and 4 show predictive accuracy of the six imputation methods when drugs are matched by active ingredient and dosage form only, that is, at the GPI-12 level. Table 3 shows the results for total payments, and Table 4 shows the results for out-of-pocket payments and patent status. Similar to the GPI-14 level matching, mean prediction error for total payment was the lowest, $1.2, when using the current weights. Mean prediction errors for total payments for weights from other methods, again, were relatively small, ranging from -$1.9 for weights derived from the true match regression to -$4.4 for weights from the total payments regression. The match weights from the total payments and the square root of total payments regressions again yielded lower mean absolute prediction errors for total payments than the weights from the other four methods, $66.0 and $65.7, respectively. The mean absolute prediction errors for total payments from other methods were, again, not much higher, ranging from $67.2 for the weights from the true match regression to $70.3 for weights from the square root of out-of-pocket payments regression. Lin's concordance correlation coefficients differed for total payments across all six methods, ranging from 0.473 for weights from the square root of out-of-pocket payments regression to 0.633 for weights from the total payments regression.

The mean prediction errors for out-of-pocket payments were less than a dollar (negative or positive) in all six imputation methods (Table 4). The mean absolute prediction error in out-of-pocket payments was again the lowest, $12.7, using the current weights. Mean absolute prediction errors in out-of-pocket payments in other imputation methods were also not very different, ranging from $12.9 using weights derived from the out-of-pocket payment regression to $15.6 using weights from the total payments regression methods. Lin's concordance correlation coefficients were similarly low across all six methods, ranging from 0.104 when using weights based on the square root of total payment regression to 0.184 for weights based on either the out-of-pocket payments regression or the true match regression. All regression methods predicted patent status quite well, ranging from 97.2% to 97.9% agreement.

Discussion

In this study we have examined a range of regression-based methods in order to improve the methods in the MEPS used to match or impute pharmacy data to the drugs reported by households, specifically for household-reported drugs for persons lacking pharmacy data. The relative ranking of the sets of matching weights in terms of predictive accuracy depends upon the specific measure of predictive accuracy, but a few findings from our study stand out. First, all methods, including the current weights, predict total and out-of-pocket payments fairly well, insofar as mean prediction errors were small for all sets of matching weights. Second, the current weights actually perform slightly better in predicting total payment, on average; mean prediction errors for total payments were slightly

smaller using the current weights compared with other matching weights. Third, the matching weights from the total payment regressions (linear and square-root) tend to yield somewhat more accurate total payments (measured by mean absolute error) and tend to yield more precise total payments (measured by higher concordance correlation coefficient). Fourth, some weights derived from regressions perform better in terms of other measures such as concordance correlation coefficient. Fifth, weights from none of the methods yield imputations with high levels of precision and accuracy in out-of-pocket payment imputations, although weights based on out-of-pocket payment regression perform better. Lastly, all methods predict patent status with a high degree of accuracy.

There are three possible reasons the sets of matching weights derived from regressions did not exhibit greater predictive accuracy. First, the character variable for drug name could not be included in the payment regressions. For drugs that are off-patent, the drug name reflects brand versus generic manufacturers, which is correlated with price. Indeed, the agreement rates for patent status are slightly higher for imputations using the current weights and the set of weights from the true match regression, and drug name is a match variable in both sets of weights. Second, the large number of fixed effects, some for less common drugs and health conditions, likely caused the payment regressions to overfit the estimation sample, leading to poor out-of-sample predictions. Third, the true match regression does not ensure that the characteristics with the greatest matching weights are the characteristics most strongly correlated with the payment outcomes we assessed.

Our future analysis will investigate matching/imputation at lower levels, such as where agreement is required at only the active ingredient level, or at the drug group level or none. The results, so far, however, suggests there may be trade-offs between lowering mean prediction error and achieving greater predictive accuracy. Both metrics are important: the data should not be biased in aggregate, and analyses of subgroups and classes of drugs are facilitated by accurate and precise imputations. Because no method of developing weights dominated the others, there is no clear reason to change current imputation methods.

References

Cohen, J. 1997. Design and Methods of the Medical Expenditure Panel Survey Household Component. Rockville, Md.: Agency for Health Care Policy and Research. http://www.meps.ahrq.gov/mepsweb/data_files/publications/mr1/mr1.pdf

Elixhauser, A., Steiner, C.A., Whittington, C.A., McCarthy, E. 2000. Clinical Classifications for health policy research: Hospital inpatient statistics, 1995. Healthcare Cost and Utilization project, HCUP-3 Research Note. Rockville, MD: Agency for Healthcare Research and Quality AHCPR Pub. No. 98-0049.

Hill, S.C., Roemer, M., Stagnitti, M.N. 2014. Outpatient Prescription Drugs: Data Collection and Editing in the 2011 Medical Expenditure Panel Survey. Methodology Report #29. Agency for Healthcare Research and Quality, Rockville, MD. http://meps.ahrq.gov/mepsweb/data_files/publications/mr29/mr29.pdf

Hill, S.C., Zuvekas, S. H., Zodet, M.W. 2011. Implications of the Accuracy of MEPS Prescription Drug Data for Health Services Research. Inquiry 48(3):242-259.

Lin, L. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics 45(1):255–268.

Table 1: Predictive accuracy for total payments by current matching weights and alternative regression-based weights, when drug is imputed within active ingredient, dosage form, and strength

Sets of Matching Weights	Mean		Lin's concordance[a]
	Prediction Error	Absolute Prediction Error	
Current Weights	-$0.3	$64.9	0.536
Regression-based Weights			
Total Payments	-$3.7	$63.4	0.641
Square Root of Total Payments	-$2.7	$62.1	0.734
Out-of-Pocket Payments	-$3.3	$66.1	0.679
Square Root of Out-of-Pocket Payments	-$2.6	$68.0	0.622
True Match	-$2.1	$65.3	0.627

Source: Authors' calculation from Medical Expenditure Panel Survey Household Component and Pharmacy Component, 2015.
[a] Lin's concordance correlation coefficient (Lin 1989) combines accuracy and precision and ranges from −1 to +1, where a larger value indicates better accuracy and precision.

Table 2: Predictive accuracy for out-of-pocket payments and patent status by current matching weights and alternative regression-based weights, when drug imputed within active ingredient, dosage form, and strength

Sets of Matching Weights	Mean		Lin's concordance[a]	Patent Status Agreement
	Prediction Error	Absolute Prediction Error		
Current Weights	-$0.5	$12.6	0.125	98.2%
Regression-based Weights				
Total Payments	-$0.1	$14.9	0.113	97.7%
Square Root of Total Payments	$0.1	$15.3	0.094	97.7%
Out-of-Pocket Payments	-$0.6	$12.8	0.179	97.7%
Square Root of Out-of-Pocket Payments	-$0.2	$13.1	0.144	97.7%
True Match	-$0.4	$14.0	0.164	97.9%

Source: Authors' Calculation from Medical Expenditure Panel Survey Household Component and Pharmacy Component, 2015.
[a] Lin's concordance correlation coefficient (Lin 1989) combines accuracy and precision and ranges from −1 to +1, where a larger value indicates better accuracy and precision.

Table 3: Predictive accuracy for total payments by current method and alternative regression-based methods, when drug is imputed within active ingredient and dosage form only

Sets of Matching Weights	Mean		Lin's concordance[a]
	Prediction Error	Absolute Prediction Error	
Current Weights	$1.2	$68.0	0.498
Regression-based Weights			
Total Payments	-$4.4	$66.0	0.633
Square Root of Total Payments	-$4.0	$65.7	0.485
Out-of-Pocket Payments	-$2.5	$70.1	0.477
Square Root of Out-of-Pocket Payments	-$2.8	$70.3	0.473
True Match	-$1.9	$67.2	0.624

Source: Authors' Calculation from Medical Expenditure Panel Survey Household Component and Pharmacy Component, 2015.
[a] Lin's concordance correlation coefficient (Lin 1989) combines accuracy and precision and ranges from −1 to +1, where a larger value indicates better accuracy and precision.

Table 4: Predictive accuracy for out-of-pocket payments and patent status by current matching weights and alternative regression-based weights, when drug imputed within active ingredient and dosage form only

Sets of Matching Weights	Mean		Lin's concordance[a]	Patent Status Agreement
	Prediction Error	Absolute Prediction Error		
Current Weights	-$0.2	$12.7	0.162	97.9%
Regression-based weights				
Total Payments	$0.1	$15.6	0.086	97.2%
Square Root of Total Payments	-$0.2	$15.2	0.104	97.3%
Out-of-Pocket Payments	-$0.4	$12.9	0.184	97.3%
Square Root of Out-of-Pocket Payments	-$0.3	$13.3	0.154	97.2%
True Match	-$0.5	$13.9	0.184	97.4%

Source: Authors' Calculation from Medical Expenditure Panel Survey Household Component and Pharmacy Component, 2015.
[a] Lin's concordance correlation coefficient (Lin 1989) combines accuracy and precision and ranges from −1 to +1, where a larger value indicates better accuracy and precision.