PharmaSUG 2018 - Paper RW-04

Triangulating Multiple Sources of Contradictory Prescription Data – a Real-World Case Study

Vidhya Parameswaran, MPH

ABSTRACT

Effective clinical decision-making relies on summarizing information from reliable, evidence-based research. With the increasing need for real-world evidence generation, scientists emphasize on developing a comprehensive understanding of their data to ensure generalizability of the observed results. Data triangulation is an efficient information integration technique through which multiple sources of overlapping data are combined to validate and strengthen the quality of the datasets. Merging data from multiple sources representing the same set of variables may give rise to contradictory values, and this may present challenges in the interpretation of the results. In this paper, I use SAS® to study the benefits of triangulating three different sources of prescription information in a retrospective, observational study setting and develop an algorithm to identify and resolve data inconsistencies. In particular, I use a mechanism through which I assign a quality score to each data source based on accuracy, validity, completeness, and uniformity of the observed records, and selectively prune records with lower quality scores. Finally, I compare the results obtained from the analysis of the triangulated data set with the results from the analysis of the data source with the highest quality score, and discuss the role of triangulating data in understanding prescribing patterns of medications among healthcare providers.

INTRODUCTION

The enactment of the 21st Century Cures Act in December 2016 has paved the way for streamlining the drug development and approval process by allowing pharmaceutical companies to provide real-world evidence on new products or new indications on existing therapies instead of clinical trials [1]. Such anecdotal data are extracted from health insurance claims, national registries, FDA's adverse event reporting system (FAERS), and electronic health record systems in routine clinical care settings. Although data from observational studies can offer meaningful insights into the real-world effectiveness of therapeutic products, the quality of data collected from clinical trials can enable scientists to make unbiased causal inferences due to the controlled nature of research environment. Some of the biggest problems associated with analyzing routinely collected real-world clinical information include incomplete records, duplication of existing records, lack of standards to verify the quality of data collected, and occurrence of unstructured narrative text that may be difficult to interpret [2].

With the advent of data science and natural language processing, scientists have developed several new methodologies to make sense of incomplete and unreliable data. Quantitative and qualitative research enquires often employ data triangulation techniques which enable scientists to develop a comprehensive understanding of the problem by integrating information from several overlapping data sources. Apart from cross-verifying and validating existing datasets, triangulation can also help compensate for inadequacies found in one source of data by supplementing information from other sources [3].

A REAL-WORLD EXAMPLE

I was recently tasked with conducting a retrospective database analysis of the real-world effectiveness of a drug in controlling hyperphosphatemia among end-stage renal disease patients undergoing dialysis. Patients who were previously treated with drug A were prescribed to switch to drug B (treatment of interest) as part of routine clinical care. Patients who switched to drug B continued with drug B treatment for 12 months or switched to a new therapy (drug X). The objective of the study was to provide a before and after comparison of changes in biochemical markers of mineral bone disease and to compare the differences in adherence to each treatment regimen by calculating the medication possession ratio. Given that the drug was new to the market, it was not included in the national preferred formulary list and was prescribed when patients failed drug A.

DATA SOURCES

End-stage renal failure patients usually undergo rigorous hemodialysis treatment 3 to 5 times a week at dialysis centers. Medical record generation occurs at the nephrologist's office, dialysis centers, pharmacies, and hospitals in which the patients are hospitalized for acute clinical care. Given the vast amount of information that is generated, it is essential to integrate and merge all the records to ensure that we extract accurate and complete datasets.

The drawback of merging multiple sources of data is the rise of inconsistencies in the information collected. Data which are produced and managed concurrently by different sources often lack consistency without a control scheme, and this can present several challenges to the quality of research produced.

Specialty Pharmacy Prescription Database (S)

Data on prescription fills were procured from a specialty retail pharmacy specializing in providing oral medications and diabetic test supplies for patients with renal failure. All medications were directly shipped to the patients' homes or dialysis centers, and comprehensive monthly medication summaries were prepared and reviewed by pharmacists and sent to their physicians and renal dieticians. With a high rate of patient adherence to prescription (measured by medication possesion ratio), and a low dispensing error rate, records from this data source are expected to be of very high quality.

Electronic Prescribing Database (E)

Computer-generated electronic prescriptions created by healthcare providers and sent to any desired pharmacy of the patients' choice were obtained. Some of the advantages of integrating records from this database included capturing information for patients who did not receive medications through the specialty pharmacy, or for those receiving poly-pharmacy prescriptions, and determining whether the dose prescribed by the physician matched the dose dispensed by the pharmacy. The disadvantage of using these electronically prescribed records was the fact that I could not determine if the patients had picked up their prescription fills from the pharmacy.

Medication Reconciliation Database (M)

Electronic medical records consisting of monthly medication reconciliation lists were extracted. Medication reconciliation is a formal process where the nurse practitioners request patients to bring in all their prescription drugs, over-the-counter medications, and supplements to the dialysis clinics, and compile complete and up-to-date lists of their current medications. The quality of data obtained from these lists is sub-optimal because the data are hand-entered and subject to recall and interpretation bias.

ASSESSING QUALITY OF DATA SOURCE

To assess the quality of each data source and assign quality scores, I carried out accuracy assessment tests to determine and compare data quality dimensions among the three different data sources. The accuracy of records are determined by comparing the mean values of significant variables with mean demographic measures recorded in national registries. The validity of records are assessed by examining whether the observed variables comply with the assigned formats. Completeness and uniqueness are determined by studying whether the records have any missing variables or whether they are duplicated.

Data dimension	S	E	M
Number of patients with records in each database	8663	11,523	10,236
Accuracy [§]			
Number of reference variables compared: Patient age, time on dialysis, drug brand name, generic drug name, dose	5	5	5

Difference in mean patient age from US DOPPS [§] practice monitor (%)	10.5	10.8	12.1
Difference in mean time on dialysis from US DOPPS [§] practice monitor (%)	4.8	6.4	6.2
Difference in % of patients on drug A from US DOPPS [§] practice monitor June 2017 (%)	22	18.6	36
Difference in % of patients on drug B from US DOPPS [§] practice monitor (%)	17	32	28
Relative quality score	**3**	**2**	**1**
Validity			
Number of reference variables compared: Patient age, time on dialysis, drug brand name, generic drug name, dose, prescription fill date, prescription end date, number of pills dispensed, NDC code	9	9	9
% of patients with all variables conforming to format	78.7	59.6	32.3
% of patients with ≤ 4 variables conforming to format	5.4	20.3	42.6
% of patients with > 4 variables conforming to format	94.6	79.7	57.4
Relative quality score	**3**	**2**	**1**
Completeness			
Number of reference variables compared: Patient age, time on dialysis, drug brand name, drug generic name, dose, prescription fill date, prescription end date, number of pills dispensed, NDC code, manufacturer name	10	10	10
% of patients with no missing variable values	59.1	43	15.8
% of patients with ≤ 5 non-missing variable values	27.9	25.4	66.9
% of patients with > 5 non-missing variable values	72.1	74.6	33.1
Relative quality score	**3**	**2**	**1**
Uniqueness			
% of patients with non-duplicated records	97.8	96.5	72.3
Relative quality score	**2.5**	**2.5**	**1**
Mean relative quality score	**2.875**	**2.125**	**1**

Table 1: Quality Assessment Tests to Compare Data Quality Dimensions among Three Data Sources

[§] Comparing results from the study of a national sample of US dialysis patients (Dialysis Outcomes and Practice Pattern Study). Accessed 03/15/2018 https://www.dopps.org/dpm/

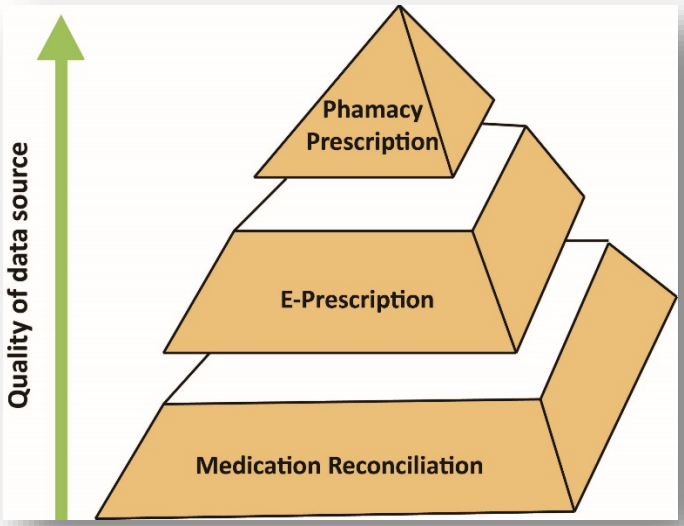

Figure 1. Quality of Data Source Pyramid

IDENTIFYING AND RESOLVING DATA DISCREPANCIES

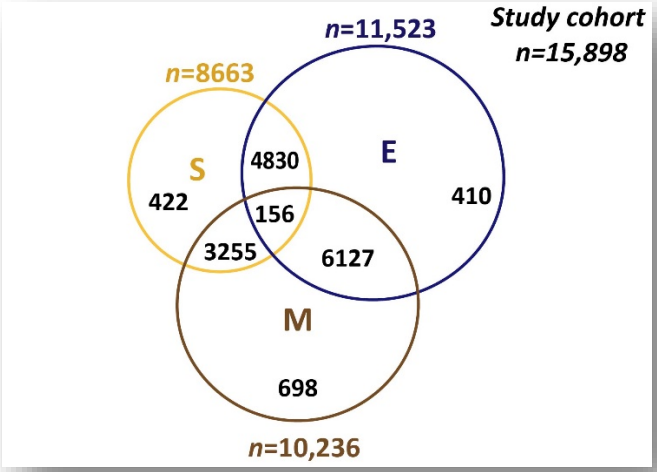

Figure 2. Number of Patients with Records from Each Data Source

Let us assume S represents specialty pharmacy records, E represents electronic prescription records and M represents medication reconciliation lists.

MONO-SOURCE PATIENTS

Each patient has records originating from only one data source. These patients can be represented by (SE'M') U (S'EM'), (S'E'M), and their records are relatively easy to extract and clean.

A sample table representing the overall merged dataset consisting of records from all sources is given below:

ID	Drug Name	Dose	Pills per Day	Prescription Fill Date	Stop Date	Source
133386	Drug A	1500 mg	8	01/02/2015	04/01/2015	M
925983	Drug X	200 mg	11	12/28/2014	03/15/2015	E
1123568	Drug B	500 mg	3	05/09/2016	07/01/2016	S

Table 2. Patients Records from One Data Source

SAS Steps for Triangulation of Records from One Data Source

Step 1: Create separate datasets for each source:

```
DATA Specialty_Rx  EPrescriptions MedReconciliation;
SET all;
BY id;
IF source="Specialty pharmacy records" THEN OUTPUT Sp_Rx;
ELSE IF source="E-Prescribe database" THEN OUTPUT EP_Rx;
ELSE IF source="Medication reconciliation list" THEN OUTPUT MedRec;
RUN;
```

Step 2: Count the number of patients in each data source:

```
PROC FREQ DATA= Sp_Rx NOPRINT; TABLES id /OUT=cn_ Sp_Rx; RUN;
PROC FREQ DATA= EP_Rx  NOPRINT; TABLES id /OUT=cn_ EP_Rx; RUN;
PROC FREQ DATA= MedRec NOPRINT; TABLES id /OUT=cn_ MedRec; RUN;
```

Step 3: Merge all the count datasets:

```
DATA sources;
MERGE cn_ Sp_Rx (in=a) cn_ EP_Rx (in=b) cn_ MedRec (in=c);
BY id;
IF a THEN s=1;
IF b THEN e=1;
IF c THEN m=1;
IF s=1 & e ne 1 & m ne 1 THEN category="Only specialty pharmacy records";
ELSE IF s ne 1 & e = 1 & m ne 1 THEN category="Only E-Prescribe database";
ELSE IF s ne 1 & e ne 1 & m = 1 THEN category="Only medication reconciliation list";
ELSE category="Poly-source";
RUN;
```

Step 4: Delete patients with multiple sources of prescriptions and clean your dataset for analysis:

```
DATA onesource;
MERGE all (in=a) sources (in=b);
BY id;
IF b and category NOT ="Poly-source";
len0=filldate- index_date;
len1=stopdate-index_date;
IF len0 > index_date + 360 THEN delete; *** prescriptions ending before baseline ***;
IF len1 < index_date - 360 THEN delete; **** prescriptions starting after follow-up ***;
IF len0 < index_date - 360 THEN start= index_date - 360
IF len1 > index_date + 360 THEN stop=index_date+ 360;
ARRAY drugA{*} A1-A721;
ARRAY drugB{*} B1-B721;
ARRAY drugX{*} X1-X721;
ARRAY drugA_Dose{*} A_Dose1-A_Dose721;
ARRAY drugB_Dose{*} B_Dose1-B_Dose721;
ARRAY drugX_Dose{*} X_Dose1-X_Dose721;
DO i=len0 TO len1;
IF drugname="drug A" THEN DO;
drugA{i}=1;
drugA_Dose{i}=Dose_per_day;
END;
IF drugname="drug B" THEN DO;
drugB{i}=1;
drugB_Dose{i}=Dose_per_day;
END;
IF drugname="drug X" THEN DO;
drugX{i}=1;
drugX_Dose{i}=Dose_per_day;
END;
DROP i source;
RUN;
```

Step 5: Calculate daily dose and type of medication used for each patient:

```
PROC MEANS DATA=onesource NOPRINT;
BY id;
VAR A1-A721 A_Dose1-A_Dose721 B1-B721 B_Dose1-B_Dose721 X1-X721 X_Dose1-
X_Dose721;
OUTPUT OUT=onesource_v2 (keep=id A1-A721 A_Dose1-A_Dose721 B1-B721 B_Dose1-
B_Dose721 X1-X721 X_Dose1-X_Dose721) MEAN= A1-A721 A_Dose1-A_Dose721 B1-B721
B_Dose1-B_Dose721 X1-X721 X_Dose1-X_Dose721;
RUN;
```

```
DATA onesource_v3;
SET onesource_v2;
BY id;
ARRAY drugA{*} A1-A721;
ARRAY drugB{*} B1-B721;
ARRAY drugX{*} X1-X721;
ARRAY drugA_Dose{*} A_Dose1-A_Dose721;
ARRAY drugB_Dose{*} B_Dose1-B_Dose721;
ARRAY drugX_Dose{*} X_Dose1-X_Dose721;
ARRAY phosbind{*} $10 bl_pb360-bl_pb1 fu_pb1-fu_pb361;
ARRAY allDose{*} bl_Dose360-bl_Dose1 fu_Dose1-fu_Dose361;
DO i=1 TO len721;
phosbind{i}=""; allDose{i}=0;
IF drugA{i} =1 THEN DO;
phosbind{i}=compbl(phosbind{i} || "A");
allDose{i}=allDose{i} + drugA_Dose{i};
END;
IF drugB{i} =1 THEN DO;
phosbind{i}=compbl(phosbind{i} || "B");
allDose{i}=allDose{i} + drugB_Dose{i};
END;
IF drugX{i} =1 THEN DO;
phosbind{i}=compbl(phosbind{i} || "X");
allDose{i}=allDose{i} + drugX_Dose{i};
END;
RUN;
```

POLY-SOURCE PATIENTS

Patients with records from multiple sources are represented as (S – SE'M') U (E – S'EM') U (M – S'E'M)

A sample table representing the dataset for these poly-source patients consisting of records from three sources is given below:

ID	Drug Name	Dose	Pills per Day	Prescription Fill Date	Stop Date	Source
133386	Drug A	1500 mg	8	01/02/2015	04/01/2015	M
133386	Drug A	200 mg	11	01/04/2015	03/15/2015	E
133386	Drug B	500 mg	3	02/12/2015	04/01/2015	S

Table 3. Patients Records from Multiple Data Sources

Order of priority:

Specialty pharmacy records > electronic prescriptions > medication reconciliation list

SAS Steps for Triangulation of Records from Multiple Data Sources

Step 1: Obtain records for patients with multiple sources of prescription and remove duplicated records:

```
DATA polysource;
MERGE all (in=a) sources (in=b);
BY id;
IF b and category ="Poly-source";
RUN;

PROC SORT DATA=polysource NODUPKEY OUT=nodu;
BY id filldate stopdate drugname dose_per_day;
RUN;
```

Step 2: Assigning quality scores:

```
DATA polysource3;
RETAIN id filldate stopdate drugname Dose_per_day source index_date;
SET nodu;
BY id;
SELECT (source);
WHEN ("Specialty pharmacy records ") weight=3;
WHEN ("E-Prescribe DATAbase ") weight=2;
WHEN ("Medication reconciliation list ") weight=1;
OTHERWISE;
END;
prevstart=lag(filldate);
prevstop=lag(stopdate);
prevpb=lag(drugname);
prevppd=lag(Dose_per_day);
prevwt=lag(weight);
IF first.id THEN DO;
prevstart=.; prevstop=.; prevpb="";
prevppd=.; prevwt=.;
END;
FORMAT prevst: date9.;
RUN;

DATA polycource4;
```

```
SET polysource3;
BY id;
IF filldate=prevstart & stopdate=prevstop & drugname=prevpb & prevppd=. THEN delete;
DROP prev:;
RUN;
```

Step 3: Weighing by quality scores:

```
PROC SORT DATA=multsources4; BY id filldate stopdate drugname weight; RUN;

DATA polysource5;
SET polysource4;
BY id filldate stopdate drugname weight;
prevstart=lag(filldate);
prevstop=lag(stopdate);
prevpb=lag(drugname);
prevppd=lag(Dose_per_day);
prevwt=lag(weight);
IF first.id THEN DO;
prevstart=.; prevstop=.; prevpb="";
prevppd=.; prevwt=.;
END;
FORMAT prevst: date9.;
RUN;

DATA polysource6;
SET polysource5;
BY id;
IF filldate=prevstart & stopdate=prevstop & drugname=prevpb & Dose_per_day ne prevppd
THEN DO;
        IF weight > prevwt THEN cat="Lesser value";
        ELSE IF weight=prevwt THEN cat="Same value";
        ELSE IF weight < prevwt THEN cat="More value";
END;
IF cat="Lesser value" THEN delete;
DROP prev:;
RUN;

PROC MEANS DATA=polysource6 noprint; ***same value patients***;
VAR Dose_per_day;
BY id filldate stopdate drugname weight;
OUTPUT OUT=polysource7 (keep=id filldate stopdate drugname weight Dose_per_day)
MEAN=Dose_per_day;
RUN;
```

Step 4: Selectively pruning records based on quality score and finding average daily dose of drug:

```
DATA j1;
SET polysource7;
BY id;
SELECT (weight);
WHEN (3) source="Specialty Pharmacy";
WHEN (2) source="E-Prescription";
WHEN (1) source="Medication Reconciliations";
OTHERWISE;
END;
len0= filldate-index_date;
len1= stopdate-index_date;
IF len0 > index_date + 360 THEN delete; *** prescriptions ENDing before baseline ***;
IF len1 < index_date - 360 THEN delete; **** prescriptions starting after follow-up ***;
IF len0 < index_date - 360 THEN start= index_date - 360
IF len1 > index_date + 360 THEN stop=SO_start + 360;
ARRAY drugA{*} A1-A721;
ARRAY drugB{*} B1-B721;
ARRAY drugX{*} X1-X721;
ARRAY drugA_Dose{*} A_Dose1-A_Dose721;
ARRAY drugB_Dose{*} B_Dose1-B_Dose721;
ARRAY drugX_Dose{*} X_Dose1-X_Dose721;
DO i=len0 to len1;
IF drugname="drug A" THEN DO;
drugA{i}=1;
drugA_Dose{i}=Dose_per_day;
END;
IF drugname="drug B" THEN DO;
drugB{i}=1;
drugB_Dose{i}=Dose_per_day;
END;
IF drugname="drug X" THEN DO;
drugX{i}=1;
drugX_Dose{i}=Dose_per_day;
END;
DROP i source;
RUN;


DATA Sp_Rx2 EP_Rx2 MedRec2;
SET j1;
BY id;
IF source="Specialty Pharmacy" THEN OUTput Sp_Rx2;
```

```
ELSE IF source="E-Prescription" THEN OUTput EP_Rx2;
ELSE IF source="Medication Reconciliation" THEN OUTput MedRec2;
RUN;

%MACRO everysource(dts=, prefix=);
PROC MEANS DATA=&dts noprint;
VAR pill1-pill182;
BY id;
OUTPUT OUT=&dts.a (keep=id &prefix.1 - &prefix.182) mean= &prefix.1 - &prefix.182;
RUN;
%MEND;

%everysource(dts=sp_rx2, prefix=sp_rx);
%everysource(dts=ep_rx2, prefix=ep);
%everysource(dts=medrec2, prefix=mr);

DATA j2;
MERGE sp_rx2a (in=a) mr2a (in=b) ep_rx2a (in=c);
BY id;
IF a or b or c;
ARRAY Dose{*} bl_Dose360-bl_Dose1 fu_Dose1 fu_Dose361;
ARRAY wt1{*} sprx1-sprx721;
ARRAY wt2{*} ep1-ep182;
ARRAY wt3{*} ss1-ss182;
ARRAY wt4{*} mr1-mr182;
DO i = 1 to 721;
IF wt1{i} ne . THEN Dose{i}=wt1{i};
ELSE IF wt2{i} ne . THEN Dose{i}=wt2{i};
ELSE IF wt3{i} ne . THEN Dose{i}=wt3{i};
ELSE IF wt4{i} ne . THEN Dose{i}=wt4{i};
ELSE Dose{i}=.;
END;
DROP sprx: ep: mr: i;
RUN;
```

COMPARING OUTCOMES FROM THE TRIANGULATED DATASET WITH SPECIALTY PHARMACY PRESCRIPTION DATASET

To assess the comparative advantage of triangulating multiple data sources, I compared the outcomes observed between the analysis of the triangulated dataset and the non-triangulated dataset with the highest quality score (specialty pharmacy prescription database)

Metric	Triangulated dataset	Specialty pharmacy dataset
Number of patients with treatment records	15,898	8,663
Average time on drug A during baseline (months)	5.7	4.8
Average time on drug B during follow-up (months)	11.6	8.9
Number (%) of patients completing 12 months follow-up on drug B	96.2	68.4
Average drug A dose during baseline	6880 mg/day	6800 mg/day
Average drug B dose during baseline	2400 mg/day	1950 mg/day
Mean medication possession ratio of drug B	86%	73%

Table 4. Comparing Results from Triangulated Dataset and Specialty Pharmacy Records

Comparing the results from the triangulated dataset with results from the specialty pharmacy dataset, we can observe that there are notable differences in average follow-up time on drug B, number of patients completing one year of drug B therapy, and mean treatment adherence measure (medication possession ratio). It is evident that triangulation of pharmacy dataset with other sources of data are useful in helping scientists carry out comparative effectiveness research.

CONCLUSION

Merging data from multiple sources representing the same set of variables may give rise to contradictory values. Triangulating multiple datasets with overlapping variables helps ensure the quality and completeness of information collected for research purposes. The SAS MACROS presented in this paper can be easily modified and adapted to address different research questions.

REFERENCES

1. Nutbeam, D., 2000. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. *Health promotion international*, *15*(3), pp.259-267.
2. Luepker, R.V., 2005. Observational studies in clinical research. *The Journal of laboratory and clinical medicine*, *146*(1), pp.9-12.
3. Leser, U., & Freytag, J. C. (2004, June). Mining for patterns in contradictory data. In *Proceedings of the 2004 international workshop on Information quality in information systems* (pp. 51-58). ACM.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Vidhya Parameswaran, MPH
Email: vpmwar@bu.edu

LinkedIn: https://www.linkedin.com/in/vidhya-parameswaran-48930192