PharmaSUG 2016 - Paper HA05

Four "Oops" Moments While Using Electronic Health Records to Identify a Cohort of Medication Users

Steve Ezzy, Optum Epidemiology, Waltham, MA

ABSTRACT

In life, and when using electronic health record (EHR) data, things don't always go the way you plan. Random quirks in software, metadata and clinical data can lead you down rabbit-trails that take hours to discover, diagnose and correct. I present four quirks encountered while identifying a cohort of users of a particular medication within an EHR and describe procedures to help safeguard against them.

1. Use of mapped fields, such as medication codes, can simplify a myriad of local EHR code issues. However detailed knowledge about how the mapping was performed is essential.

2. Determining which dates are best to use from Rx data is complicated --> Issue Date, Patient-Report Date, Action Date, Med Reported Date, Update Date, Discontinue Date, Expiration Date? Choice of date should be determined by whether the research question is aimed at topics such as prescribing behaviors, adherence or drug utilization.

3. Duplicates or not? Some clinical Rx records are complete duplicates of each other, distinct only by an Rx ID field. We can de-duplicate rows ignoring the Rx ID, but must consider the risk of under-reporting of prescriptions.

4. Non-breaking spaces; few programmers know these special characters exist but when they occur in data, they can confound your SAS code if you're not expecting them. For example, you may encounter data selection errors when using IF statements with text strings which include these characters.

INTRODUCTION

When using electronic health record (EHR) data, random quirks in software, metadata and clinical data can lead you down rabbit-trails that take hours to discover, diagnose and correct. Many of these trails are different from those encountered when using more familiar insurance claim data.

Although neither insurance claim data nor EHR data are expressly structured for research purposes, insurance claim data are standardized for the purpose of payment and therefore lend themselves to research purposes more readily than EHR data where the purpose is medical record keeping and where standardization is far less.

I have likened the journey of using EHR data in research to exploring the Wild West where rule of law, well-worn paths and detailed maps are all hard to come by. I present here four trails I've recently encountered while identifying a cohort of users of a particular medication within an EHR and place signposts to help safeguard you on your journey.

TRAIL 1 - USING MAPPED FIELDS

DISTINGUISHING “LOCAL” FIELDS FROM “MAPPED” FIELDS

Mapped fields can eliminate the hassle of dealing with a myriad of local EHR codes, but be sure that the mapping is performed the way you expect. For example, these are the National Drug Code (NDC) fields from our EHR that we used in a recent study:

- Local NDC – The NDC code reported in the EHR data. This field will be set to null out if it contains protected health information (PHI).
- Standardized Local NDC - The field created by standardization of the Local NDC (client's native NDC value) into a standard (padded with leading 0s for each segment) 11 digit, non-dashed derivative. Local NDC values that fail this standardization process (ex. 10 digit NDC without dashes, or a 12 digit NDC) are null in the Standardized Local NDC field.
- Mapped NDC - The Local NDC is verified by executing a query value to confirm the value matches against reference data values. Mapped NDC is imputed based on an underlying mapping logic and is a representative NDC which may or may not match the Local NDC because it is based on the most popular NDC in the EHR for a specific "ingredient" and it does not take route, strength, form or other data elements into account. It was recommended by our database developer that we use Mapped NDC only as a way to identify the "ingredient" of the product and not for product specific analysis.

We attempted to identify users of the extended-release form of a certain drug (e.g. NDC=’11111111111’) by looking at the first non-missing value in this trio of NDC fields in this order: Standardized Local NDC, then Local NDC, and lastly Mapped NDC. Then we selected records with any of the extended-release drug’s NDCs. The COALESCEC function was used to select the first non-missing field for each record:

```
ndc_x = coalescec(trim(standardized_local_ndc), trim(local_ndc), trim(mapped_ndc));
if ndc_x in ('11111111111');
```

However the number of users of the extended-release drug was lower than expected, so we did some research and found that not all was as it seemed.

HOW SPECIFICITY MAY BE LOST IN THE MAPPING PROCESS

Per the process mentioned above, the Local NDC of the extended-release form of the drug (‘11111111111’) had been mapped to the Mapped NDC field as the most popular form of the drug with the same ingredient (e.g. ‘22222222222’), which was not the extended-release form.

Then in certain records where PHI was noted within the value for Local NDC (and therefore the Standardized Local NDC as well), those fields’ values were by practice nulled out, leaving only the Mapped NDC field with any NDC value. Therefore when we searched the ndc_x field for ‘11111111111’ we did not capture those records.

DEVELOPING TECHNIQUES FOR REGAINING SPECIFICITY

Even though Local NDC had been nulled out in these records, the values for Local Dose Strength Per Unit had been preserved. From our knowledge about the regular and extended-release forms of the drug, we could see that in these records, the value for Local Dose Strength Per Unit was consistent with the extended-release form of the drug and was much too high to be associated with the regular form of the drug. So augmenting the selection criterion like so:

```
if ndc_x in ('11111111111') or
  (ndc_x in ('22222222222') and strength_per_unit = '5 MG');
```

… we were able to detect records with the extended-release form which had actually been mapped to the regular form. The number of patients in our study group then increased to expected levels.

TRAIL 2 - SELECTING DATE FIELDS

THOROUGH KNOWLEDGE OF THE DEFINITION OF EACH DATE IS ESSENTIAL.

It is important to be familiar with the range of dates available within each type of prescription (Rx) data.

	Prescription Drug	**Inpatient Administered Drug**	**Patient-Reported Drug**	**Definition of Date Field**
Action Date			X	Date/Time when provider made a recommendation about this medication, telling a patient to continue or stop taking it, without writing an Rx for it.
Administration Date		X		Date/Time when the patient received the medication, if known.
Discontinue Date	X		X	Date/Time the medication is discontinued
Dispensed Date		x		Date/Time the medication was dispensed from Hospital Dispensary to nurse for administration to patient, if known.

Expiration Date	X			Date until which the prescription ordered is “valid”. Not the expiration date according to the manufacturer. Not well populated and of different meanings depending on the data source.
Issue Date	X			Date the medication was ordered.
Med Reported Date			x	Date/Time at which the patient mentioned being on the med.
Update Date	X	x	x	Date that the update to this record occurred.

Table 1. EHR Rx Data – Date Variables

RESEARCH QUESTION DETERMINES THE CHOICE OF DATE(S)

Choice of date(s) should be determined by whether the research question is aimed at prescribing behaviors, adherence or drug utilization. Issue Date would be applicable to all three areas. Most likely, questions about prescribing behaviors would also focus on Action Date, whereas Adherence questions would involve Discontinue Date and Med Reported Date. Drug utilization studies may take into account Administration Date, Dispensed Date and Rx Administered Date.

TRAIL 3 - DEALING WITH DUPLICATE RX RECORDS

EXTENT OF DUPLICATION DEPENDS ON NUMBER OF VARIABLES SELECTED

In a sample of 20,000,000 records from our EHR Rx data, all were unique when considering all 66 variables including a unique Rx ID. But if Rx ID is not considered, 8,915 observations (0.04%) were found to be complete duplicates (i.e. values for all the other 65 variables were identical). The percentage of duplicates increased if even fewer variables were considered.

REASONS FOR DUPLICATION

Duplication of Rx records from EHR data may occur for these and other reasons:

- During a hospitalization, there may be multiple prescriptions written on the same day so all the details would be the same, especially if issue times were not recorded.
- Data for the same prescription can sometimes come from multiple sources (e.g. inpatient medication orders, discharge meds, prescriptions - which you would know only if you had a data source variable to consider). There are differences in the way source EHR systems capture information.

AVOIDING DUPLICATES VS. UNDER-REPORTING

If your data can be filtered by order type (prescription vs. inpatient administration vs. patient-reported), you may want to do so since inpatient administrations are more likely to contain multiple rows (e.g. because of standing orders). Data systems which contribute data to the overall EHR may capture data differently. In one source, a single Rx order can result in multiple Rx administrations; for another, multiple Rx orders can result in multiple Rx administrations. Consider using Order Status Field (completed, cancelled, etc.) to further refine the study inclusion criteria.

If duplicate Rx records are from the same data source then they are likely to be true duplicates. Rows may be de-duplicated simply ignoring any unique Rx id. Some database administrators advise pulling the variables of interest for your purpose and de-duplicate at that level, but the risk of under-reporting increases with a decrease in the number of variables considered. It is best to use sound judgment, to use real world experience and to talk to subject matter experts to determine the best approach for your purposes.

TRAIL 4 - HANDLING NON-BREAKING SPACES

WHAT ARE NON-BREAKING SPACES AND HOW DO THEY DIFFER FROM COMMON SPACES?

A non-breaking space (NBSP) is a space character that prevents an automatic line break at its position. In the following example, the first sentence uses a common space between “January” and “30”, whereas the second sentence uses an NBSP between “January” and “30”:

The history books say that former president Franklin Delano Roosevelt celebrated his birthday on January 30.

The history books say that former president Franklin Delano Roosevelt celebrated his birthday on January 30.

Keeping certain word groups together on a line is sometimes desirable in word-processing and in such cases the use of NBSPs is required. (This behavior can also be achieved using a similar character, the non-breaking hyphen.) However, the unexpected presence of these characters in text data can introduce programming complications (EHR-related or not).

Here is a comparison of the two types of spaces:

Character	Microsoft Windows Entry Method	Hex Code	Effect of STRIP or COMPRESS Functions (Default Settings)
Space	<Space>	20	Removed
Non-Breaking Space	Alt+0+1+6+0 (on the keypad)	A0	Not removed

Table 2. Comparison of the Space Character and the Non-Breaking Space Character

HOW DO WE REMOVE NON-BREAKING SPACES?

Care must be taken when attempting to remove NBSPs. In the data lines of the following example, a common space separates the pair of 'space' words and an NBSP separates the pair of 'nbsp' words:

```
data test;
    input @1 uncompressed $char11.;
    compressed_default       = compress(uncompressed);
    compressed_w_s_modifier  = compress(uncompressed,,'s');
    compressed_w_a0_hex_code = compress(uncompressed,'A0'x);
    datalines;
space space
nbsp nbsp
;
run;

proc print noobs data=test;
run;
```

Below we see that neither the default setting nor the inclusion of the modifier to remove spaces (‘s’) enables the COMPRESS function to remove NBSPs. Only an explicit mention of the hex value for NBSP ('A0'x) in the list of characters accomplishes this.

```
                                               compressed_
                compressed_    compressed_      w_a0_hex_
uncompressed      default      w_s_modifier       code

space space     spacespace      spacespace     space space
nbsp nbsp       nbsp nbsp       nbsp nbsp      nbspnbsp
```

Output 1. Results from Compressing Spaces and NBSPs Using Various COMPRESS Parameters.

PRACTICAL IMPLICATIONS OF NON-BREAKING SPACES IN TEXTUAL DATA

We attempted to use a set of NDC codes to select Rx records, but were unaware of the presence of NBSPs trailing some of the NDCs in our list. This erroneously limited our selection of Rx records. Once the NBSPs were discovered and eliminated, a complete set of Rx records was selected.

CONCLUSION

As we stumble over and recover from EHR issues like these, we learn more and more. And as this body of knowledge grows, we will be able to design better programming and research tools and more efficiently use this incredibly rich data resource to benefit public health.

ACKNOWLEDGMENTS

I thank Kathleen Mortimer and Robert Gately for their kind review of this paper.

CONTACT INFORMATION

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

Name:	Stephen Ezzy
Enterprise:	Optum Epidemiology
Address:	950 Winter Street, Suite 3800
City, State ZIP:	Waltham, MA 02451
Work Phone:	781-419-8498
Fax:	781-472-8464
E-mail:	stephen.ezzy@optum.com