Open Payments data Exercise  
Created by: @MaryJoWebster  
August 2015

This is an exercise using the [Open Payments data](https://www.cms.gov/openpayments/) released in early 2015 by the Centers for Medicare and Medicaid Services (CMS), showing payments from pharmaceutical and medical device companies to doctors. This particular data we’re going to use is a slice of that data with payments to doctors who are licensed to practice in Minnesota (but not necessarily based there) and includes both general payments and those for research. I added a field called “tabletype” that indicates which batch of data it comes from. This data released in early 2015 – representing payments in 2014 -- was the first year that complete data was made available.

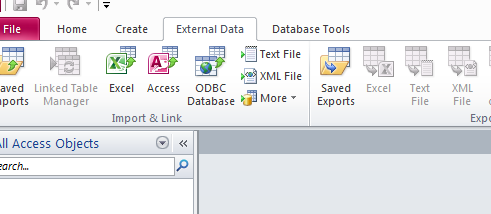
See the separate Excel file for the record layout.

The data is in a comma-delimited file called “MNdocs\_2014.csv”

**Importing the data:**

Open a new, blank Access database.

Go to the File menu and choose “Save Database As”…. Give it a name and save it to a location where you’ll remember (ideally on a network drive that is backed up, not your C: drive)

Back inside the database….Go to the External Data tab and choose “Text file”

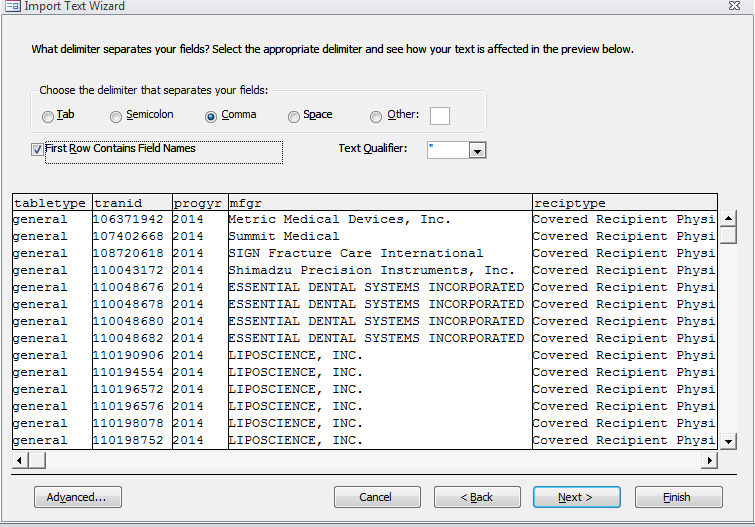
This will bring up the Import Wizard

On the first screen, click the Browse button to find “MNdocs\_2014.csv”

Leave the default option – “Import the source data into a new table…”

Click OK

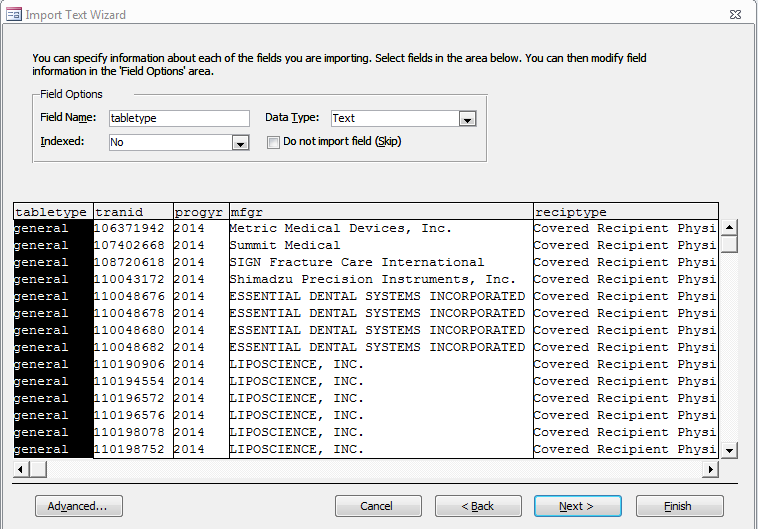
Access should guess that this is a delimited text file – if so, click the Next button.

On the next screen, make sure it has guessed that the delimiter is a “comma” and click the box that says “First Row Contains Field Name” and make sure “Text qualifier” has double quotes selected.

Click Next.

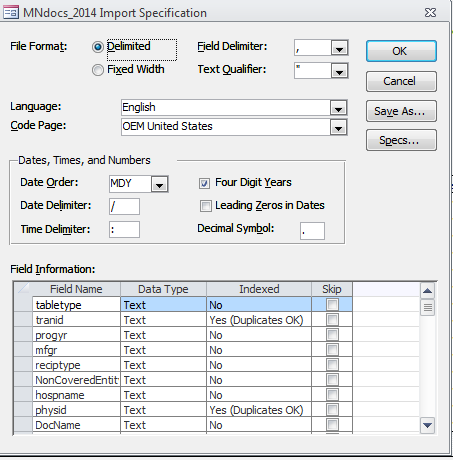
The next screen is the most important. Here is where you set the field names and the data types (text, numeric, date/time, etc). If you get something wrong here, you’ll have problems down the road. It’s a good idea to at least try to view your data in a text editor before importing. Even the free text editors should allow you to open a file you are importing to Access – if it says the file is too big, it’s very likely you’ll also have trouble with Access’ size limitations.

Some important things to keep in mind when setting data types:

* If you have a date field that you want to do analysis on – i.e. day of the week, time elapsed from one date to another, even just sorting chronologically – then you need to have it set as Date/Time.
* If you have any fields that serve as unique identifiers or fields that use numbers, but you won’t do math on them (i.e. zip codes), it’s best to set them as Text. Numeric fields in Access (and Excel does the same) will drop any leading zeros. This can cause huge problems down the road.
* If you have a numeric or currency field that you want to do mathematical calculations on – sum, average, etc. --- it needs to be set as one of the numeric field types. Those are: Integer, long integer, currency, single and double.
* If you have a field with decimal places, set it as double.
* Integer is for very small numbers (i.e. 1 through 10). Long integer should be your default for numbers, unless you need decimal places (then use “double”)
* Use “memo” if you have a text field that is greater than 255 bytes. (these are rare)
* If you have any suspicion that a field might have mixed types of data (some text, some numeric, some dates), then you either need to import it as text or clean it up before bringing it in. This is crucial for a field that you might want to set as a date or currency or numeric. For example, it’s quite common to get data that has “N/A” in fields that might otherwise be used for a date or a number – Access will throw them out if you try to set the field type as number or date.
* It is feasible (although harder) to bring all your fields in as “text” and then do conversions into new fields using Update queries for anything that needs to be numeric or date/time.
* Pay attention to the format of the date/time fields in your text file. Is it stored like: “12/5/2014” or “2014-12-05” or “Dec-05-2014” or “December 5, 2014”? The last two – with “Dec” or “December” – will probably not import as a date field; Access will want to bring it in as text. The one that starts with the year will need to be noted in the import wizard (see steps below for how to do that)

There are two ways you can go through each of your fields, setting names and data types. The first is by clicking through them in this preview window and editing the section called “field options”. To do this you have to click on each field in the preview window, turning it black (like tabletype is in the screenshot), and then editing the field options info for that field before moving on to the next field.

The second is by clicking the Advanced button (this will only be visible when importing text files. It doesn’t show up when importing Excel files)

Here’s what shows up when you click Advanced:

Note the section in the middle dealing with “dates, times and numbers”. Here is where you specify how your dates are stored (month-day-year, year-month-day, etc). If your dates are stored like “2014-12-04”, then you need to choose “YMD” here and it will be applied to all date fields (hopefully all your date fields are the same…usually they are but I’ve encountered datasets where they are not)

In the “field information” section at the bottom, you can edit the field names and data types. And also indicate if you want a field indexed and whether you want to skip a particular field.

This approach is easier if you have a lot of fields that need their names changed or something like that.

NOTE: Access does NOT like field names made up of more than one word, or names that contain symbols or start with numbers. It will let you do them, but you’ll have extra typing to do when it comes time to write queries. Some options that help with field naming: Use camel case (i.e. LastName) or use underscores (i.e. “last\_name”)

**Back to importing open payments data…**

For this dataset, you’ll see the field names are already set to single words and no symbols, etc. But all the fields are coming in as “text” data type. There are only 2 numeric and date fields that need to be converted: “payment” , “paydate”.

The fields called “context” and “studyname” need to be set as “memo” data type – cause they each exceed 255 bytes.

And then make sure the “date order” properties (in the “dates, times and numbers” section) is set to YMD cause the dates in this database are like this: “2014-08-02”

If you want, you can click on “Save As…” to save these import specifications in case you need/want to import this file again (or the same format, but with another year’s worth of data)

The next screen asks if you want Access to add a primary key. This is a unique identifier for each record. All it does is number the records, starting with 1. If you’re only dealing with one table, your need for this is minimal. But sometimes people like to have it. Your choice.

The final screen lets you set the name of the table it’s going to create. Let’s leave this one as the default, which is the same as the name of the text file: “MNdocs\_2014”

**Let’s get to know our data:**

Each record represents one “total” (could be more than one transaction) between a company and a doctor (or hospital) under a particular pay category (“paynature”). So for example, one record might show how much Company A paid Dr. Smith for consulting and another record would show how much that same company paid the same doctor for entertainment.

There are some records that are payments toward research projects – these generally have different fields of information (off the end of the table) – and will be labeled “research” in the paynature field. We can see what state the “payer” (company) is located in and what state the doctor or hospital is located in. The fields “lic1”, “lic2”, etc. indicate which state(s) the doctor is licensed to practice in.

The fields “assoc1”, “assoc2”, etc., indicate the name of a pharmaceutical that the payment is related to. The fields “market1”, “market2”, etc. indicate the name of a medical device(s) that the payment is related to. You’ll see that record typically have one or the other filled in, but not both. Some don’t have any filled in.

**Let’s ask some questions…**

1. Which Minnesota doctor received the most money overall and what is his/her specialty?

Daniel Berry, orthopaedic surgery, $2.5 million

SELECT MNdocs\_2014.DocName, spec1, Sum(MNdocs\_2014.payment) AS SumOfpayment  
FROM MNdocs\_2014  
GROUP BY MNdocs\_2014.DocName, spec1  
ORDER BY Sum(MNdocs\_2014.payment) DESC;

1. What were doctors paid for the most? (“paynature” field)

Royalty or license, $17.7 million

SELECT MNdocs\_2014.paynature, Sum(MNdocs\_2014.payment) AS SumOfpayment  
FROM MNdocs\_2014  
GROUP BY MNdocs\_2014.paynature  
ORDER BY Sum(MNdocs\_2014.payment) DESC;

1. Which MN doctor received the most for consulting?

Jack Bert, orthopedic surgeon, $365,770

SELECT MNdocs\_2014.DocName, MNdocs\_2014.spec1, Sum(MNdocs\_2014.payment) AS SumOfpayment  
FROM MNdocs\_2014  
WHERE (((MNdocs\_2014.paynature) Like "consulting\*"))  
GROUP BY MNdocs\_2014.DocName, MNdocs\_2014.spec1  
ORDER BY Sum(MNdocs\_2014.payment) DESC;

1. Which company gave the most to MN doctors?

DePuy Synthes Products, $3.8 million

SELECT MNdocs\_2014.Payer, Sum(MNdocs\_2014.payment) AS SumOfpayment  
FROM MNdocs\_2014  
GROUP BY MNdocs\_2014.Payer  
ORDER BY Sum(MNdocs\_2014.payment) DESC;

1. What product was most of that money related to?(from the company identified in question 3)

Market1 field just says “orthopedics” -- $3.6 million

SELECT MNdocs\_2014.market1, Sum(MNdocs\_2014.payment) AS SumOfpayment  
FROM MNdocs\_2014  
WHERE (((MNdocs\_2014.Payer) Like "DePuy Synthes Prod\*"))  
GROUP BY MNdocs\_2014.market1  
ORDER BY Sum(MNdocs\_2014.payment) DESC;

1. Are there any variations of the name of that company? If so, how do you think you should deal with that?

DePuy Synthes Products LLC, DePuy Orthopaedics Inc. and DePuy Synthes Sales Inc.

SELECT MNdocs\_2014.Payer, Sum(MNdocs\_2014.payment) AS SumOfpayment  
FROM MNdocs\_2014  
Where MNdocs\_2014.Payer Like "DePuy\*"  
GROUP BY MNdocs\_2014.Payer  
ORDER BY Sum(MNdocs\_2014.payment) DESC;

1. Are there any drugs or devices that are associated with a significant amount of all the payments to MN docs (not just those from DePuy)? Is the data standardized well enough to be able to say anything from these results?

A lot of records – totally $6.5 million – don’t have anything filled in for market1 or assoc1.

Orthopedics, $3.8 million, comes up second. But the data looks very unreliable. Some refer to specific products, others like the DePuy payments seem to be more general/vague.

1. What percentage of the money paid to MN doctors came from MN-based companies? (hint: this will require a couple of steps and possibly a little work in Excel)

15%

SELECT MNdocs\_2014.payerstate, Sum(MNdocs\_2014.payment) AS SumOfpayment  
FROM MNdocs\_2014  
GROUP BY MNdocs\_2014.payerstate;

1. Medtronic is the biggest player in Minnesota’s medical device company – who did they pay the most? (keep in mind these will be only MN docs they paid; it won’t include non-MN docs they paid) Hint: First spend some time looking at how Medtronic’s name appears in the “payer” field.

Ensor Transfeldt, $728,887

SELECT MNdocs\_2014.DocName, Sum(MNdocs\_2014.payment) AS SumOfpayment  
FROM MNdocs\_2014  
WHERE (((MNdocs\_2014.Payer) Like "medtronic\*"))  
GROUP BY MNdocs\_2014.DocName  
ORDER BY Sum(MNdocs\_2014.payment) DESC;

1. Ask your own question and find the answer.
2. Write a paragraph using your key findings.