UCSF
University of California
San Francisco

Using De-Identified Data: 10 practical steps

Jennifer Creasman
CTSI Consultation Services

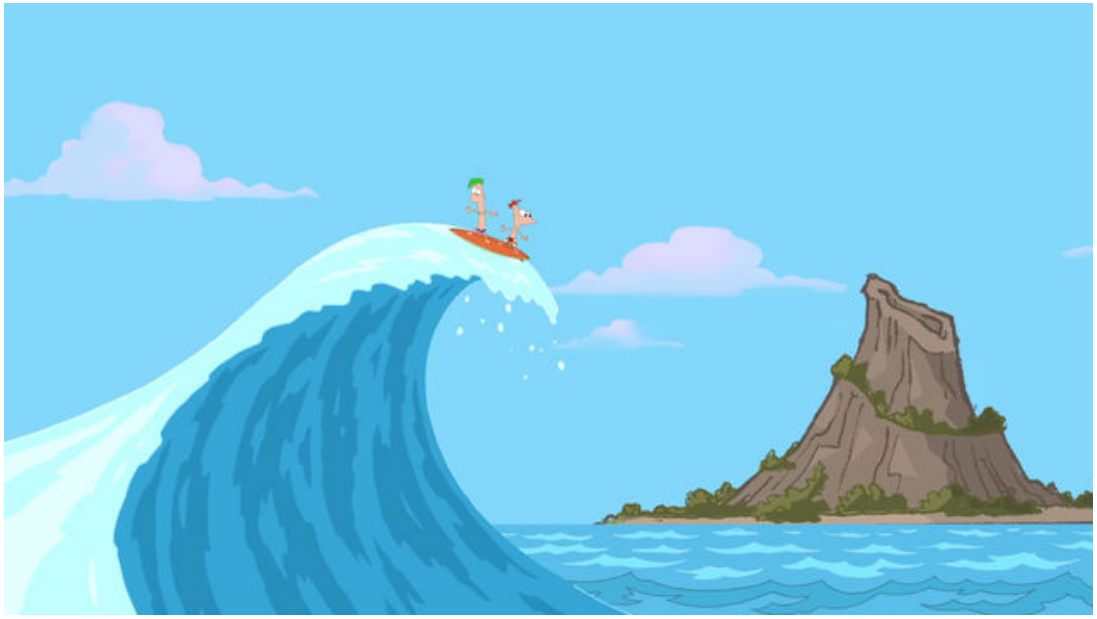
RDB De-identified Flat Files, reduced EHR data but…

Still A Tidal Wave of Data

- Number of Tables: 16
- Total File Size: 209 GB

…and growing

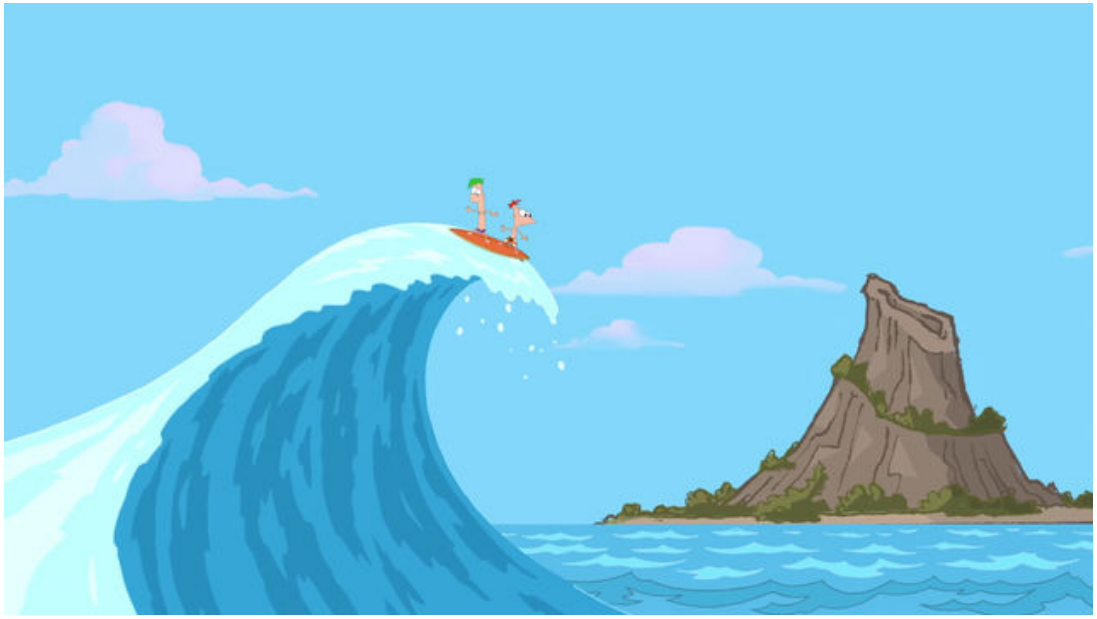
You can't read these into MExcel or MSAccess – you need other tools..



Ten Practical Steps to Staying on Top of a Tidal Wave of EHR Data

Assumptions:

- You have a clearly defined research question
- You have clinical expertise in the area of research
- If you decide to work with a Data Scientist /Programmer -- **All steps still apply**



Step 1: Find the Right Tool for You!

Language-oriented Software Programs

- UCSF Licenses: https://it.ucsf.edu/services/licensed-software
- MyResearch Platform
- Single-user license option

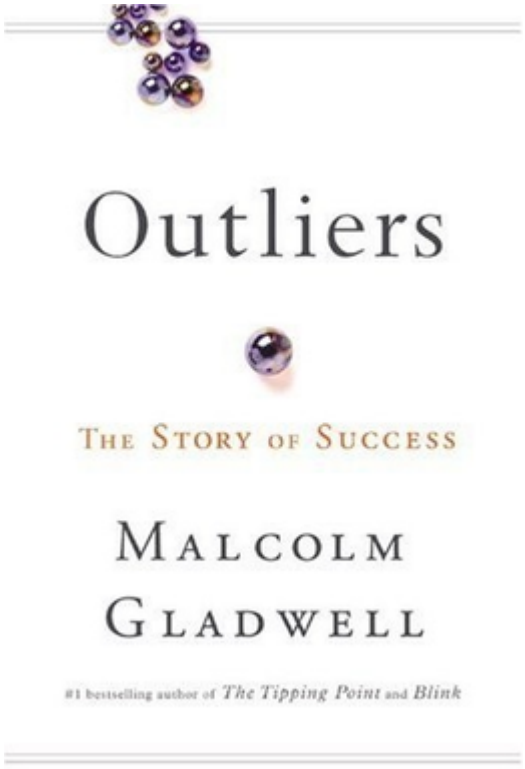
Step 2: Develop Your Programming Skills

How Do I Become a Better Programmer?

- 10,000 hours rule: Just do it!
- Data Science Courses coursera

  https:// www.coursera.org/data-science
- Join a group Meetup

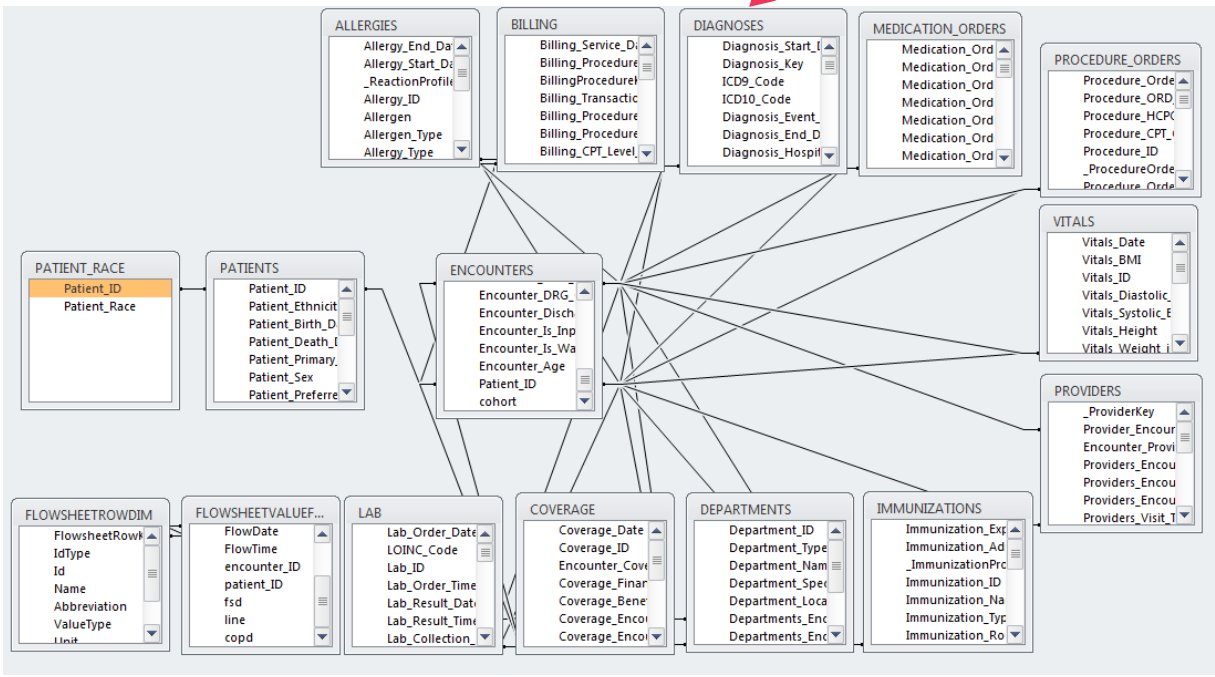
  https://www.meetup.com/topics/computer-programming/
- Become friends with other programmers. Eat lunch with them.



Step 3: Understand Your data

Working with RDB De-identified Flat Files

- Review the RDB documentation
- Study the relationships between the tables
- Open the files
- **Look at the data**



Observations: One record per patient

De-identified: Patient File

	Patient_ID	Patient_MRN	Patient_Ethnicity	Patient_Birth_Date	Patient_Death_Date	Patient_Primary_Care	Patient_Sex
1	346768627874553	426004796288908	*Unspecified	12/13/1950	.	9.5279584E14	Female
2	944827931467444	872320204973221	Not Hispanic or Latino	12/26/1956	.	1.2129073E14	Male
3	841613798402250	500256611499935	Not Hispanic or Latino	07/03/1949	.	3.5206031E14	Female
4	830560461618006	765218106564134	Not Hispanic or Latino	01/21/1937	.	3.312821E14	Female
5	651626786217094	341816220432520	Not Hispanic or Latino	06/01/1937	.	8.8845505E14	Male
6	672923571430147	679128048010171	Not Hispanic or Latino	03/03/1933	.	.	Male
7	922736370004714	756049859337509	Hispanic or Latino	06/06/1938	.	3.9810897E14	Male
8	568691051099449	928630440961570	Not Hispanic or Latino	04/09/1947	.	7.892076E14	Male
9	270648140925914	350340276490897	Hispanic or Latino	11/23/1944	.	4.5368604E14	Male
10	501155952923000	333686850965023	Not Hispanic or Latino	05/31/1926	11/29/2013	2.6485083E14	Male
11	482537743635476	783389108721167	Not Hispanic or Latino	10/11/1955	.	3.1674394E14	Male
12	2655150834471	831416873726994	Not Hispanic or Latino	10/19/1942	.	.	Female
13	656081004999578	359737460035831	Not Hispanic or Latino	08/30/1927	.	9.5279584E14	Male
14	10526411235332	399729480501264	Not Hispanic or Latino	04/06/1945	.	4.4532877E14	Female
15	109764303546399	562666151206940	Not Hispanic or Latino	12/03/1933	07/02/2012	8.4738619E14	Male
16	574953557923436	134469735901803	Not Hispanic or Latino	05/31/1926	10/29/2012	6.2523861E14	Female
17	910693597048521	101702373940498	Not Hispanic or Latino	05/31/1926	.	.	Female
18	807758098933846	290767986793071	Not Hispanic or Latino	08/06/1959	.	.	Male
19	22446092218161	729220696259290	Not Hispanic or Latino	07/23/1937	.	9.9968639E14	Male
20	681325161363930	680498836096376	Unknown/Declined	10/10/1950	.	.	Male
21	282220148947090	711024511605501	Not Hispanic or Latino	11/22/1954	.	5.3899846E14	Female

Observations: Multiple encounters per patient

De-identified: Encounter File

	Patient_ID	_EncounterKey	_EncounterDate	Encounter_ID	Encounter_Type	Visit_Type	VisitKey	Visit_Length
1	952476149890572	2.5404496E14	10-01-2011	27688734699041	Office Visit	NEW PATIENT	4.1109812E14	60
2	381854549050331	8.2730267E14	09-13-2011	976751514710486	Refill		.	.
3	5071964600384876	8.9935783E13	12-08-2011	982034848071635	Hospital Encounter		4.34137E14	.
4	347782866097987	8.4048817E14	12-06-2011	126061215065420	Document Conversion		.	.
5	347782866097987	2.1545478E14	12-06-2011	66062292549759	Hospital Encounter		6.1210455E13	.
6	802[illegible]912	7.8379866E14	08-14-2011	74731607455760	Refill		.	.
7	811443735379726	8.8837624E14	07-14-2011	676893275231123	Appointment	ICD CHECK	2.7399035E14	30
8	811443735379726	8.6470242E14	07-14-2011	322352526243776	Appointment	ECHOCARDIOGRAPHY	7.1233426E14	60
9	811443735379726	7.3175855E14	07-14-2011	341634161770344	Appointment	BLOOD DRAW	4.8784699E14	15
10	10167426895350	2.4133022E14	07-19-2011	896496478933841	Refill		.	.
11	551847187336534	9.7694616E14	05-20-2012	89419770985842	Patient Email		.	.
12	402067835442722	4.4074398E14	11-20-2011	690918767824769	Hospital Encounter		3.0939208E14	.
13	536267309915274	9.2410316E14	03-09-2012	214549112599343	Refill		.	.
14	817412601318210	6.9421571E14	03-24-2012	206992296967655	Hospital Encounter		.	.
15	617436406668276	9.4356879E14	07-02-2011	167957765981555	Orders Only		.	.
16	992631756700575	1.3853278E14	04-25-2012	242374859750271	Appointment	FOLLOW UP 20	2.2442142E14	30
17	372432052623481	5.0635851E14	03-18-2012	85713618900627	Office Visit	FOLLOW UP 30	6.1939718E14	30
18	372432052623481	1.9162134E14	03-18-2012	35740454681218	History		.	.
19	372432052623481	5.3862351E13	03-18-2012	327159833628684	Appointment	PFT 60 MINUTE	8.276393E14	60
20	873292727395892	7.616849E14	07-07-2011	176528786774725	Anti-coag visit		.	.
21	546086992602795	2.6193195E14	01-27-2012	739505285862833	Office Visit	FOLLOW UP 20	1.456386E13	20

Observations: Multiple values per encounter

De-identified: Flowsheet Data

	patient_ID	encounter_ID	FlowDate	fsd	line	FlowsheetRowKey	Value	Occurre	Count	FlowTime	cop
1	836950261146	2.4592397E13	06/04/2014	6210641	1	33748	No	-1	1	12:58:00.000	1
2	836950261146	2.4592397E13	06/04/2014	6210641	2	9514	Back	-1	1	13:01:00.000	1
3	836950261146	2.4592397E13	06/04/2014	6210641	3	9258	7	-1	1	13:01:00.000	1
4	836950261146	2.4592397E13	06/04/2014	6210641	4	1960	71	-1	1	13:01:00.000	1
5	836950261146	2.4592397E13	06/04/2014	6210641	5	5106	4144	-1	1	13:01:00.000	1
6	836950261146	2.4592397E13	06/04/2014	6210641	6	2	97	-1	1	13:01:00.000	1
7	836950261146	2.4592397E13	06/04/2014	6210641	7	35440	Oral	-1	1	13:01:00.000	1
8	836950261146	2.4592397E13	06/04/2014	6210641	8	34432	95.9	-1	1	13:01:00.000	1
9	836950261146	2.4592397E13	06/04/2014	6210641	9	39413	17	-1	1	13:01:00.000	1
10	836950261146	2.4592397E13	06/04/2014	6210641	10	38524	79	-1	1	13:01:00.000	1
11	836950261146	2.4592397E13	06/04/2014	6210641	11	32710	146/88	-1	1	13:01:00.000	1
12	836950261146	2.4592397E13	06/04/2014	6210641	12	31786	0	-1	1	13:01:00.000	1
13	836950261146	2.4592397E13	06/04/2014	6210641	13	16457	2.43	-1	1	13:01:00.000	1
14	836950261146	2.4592397E13	06/04/2014	6210641	14	16458	36.2	-1	1	13:01:00.000	1
15	836950261146	2.4592397E13	06/04/2014	6210641	15	36269	75.3	-1	1	13:01:00.000	1
16	836950261146	2.4592397E13	06/04/2014	6210641	16	36273	70.8	-1	1	13:01:00.000	1
17	836950261146	2.4592397E13	06/04/2014	6210641	17	19416	0	-1	1	13:01:00.000	1
18	836950261146	2.4592397E13	06/04/2014	6210641	18	19678	0	-1	1	13:01:00.000	1
19	836950261146	2.4592397E13	06/04/2014	6210641	19	19945	0	-1	1	13:01:00.000	1
20	836950261146	2.4592397E13	06/04/2014	6210641	20	20426	75.3	-1	1	13:01:00.000	1
21	836950261146	2.4592397E13	06/04/2014	6210641	21	24476	2.35	-1	1	13:01:00.000	1

Step 4: Don't Drown in Your Data!

Reduce the files to only include your study cohort

- SAS program examples are available in RDB folder
  - Easy to use if identifying patients using ICD9/10 codes and/or patient's demographic data
  - Otherwise, complex programming might be required
- Start with a smaller set of variables

Step 5: Get Dirty!

Own and drive the research

- Review the literature: What variables are expected to be there?
- Develop an explicit instructions on how outcomes and covariates should be derived from the available data
- Provide an explicit and exact recipe for new *calculated* variables
- Understand limitations of the data

Step 6: Develop Project Specific Documentation

Plan Ahead

- Create a mock Table 1
- Start with word descriptions
- Specify variable names, don't leave anything open to interpretation

Table 1. Characteristics of inpatient PAD populations with and without depression

Variable	D	No D	P
Age [Patient_Age]			
Gender [Patient_Sex]			
BMI [Vitals_BMI]			
Race [Patient_Ethnicity]			
Smoking [Patient_smoking]			

Calculated Variable	Description
reinsertion	whether a foley catheter was reinserted during the hospitlization for any reason with any time period
new foley during hospitalization	placement time is after hospital admission time for patients that we admitted without a foley catheter
qmonth change	new placement time minus previous placement time >25 days

Step 6: Develop Project Specific Documentation

Each variable might require a separate calculation

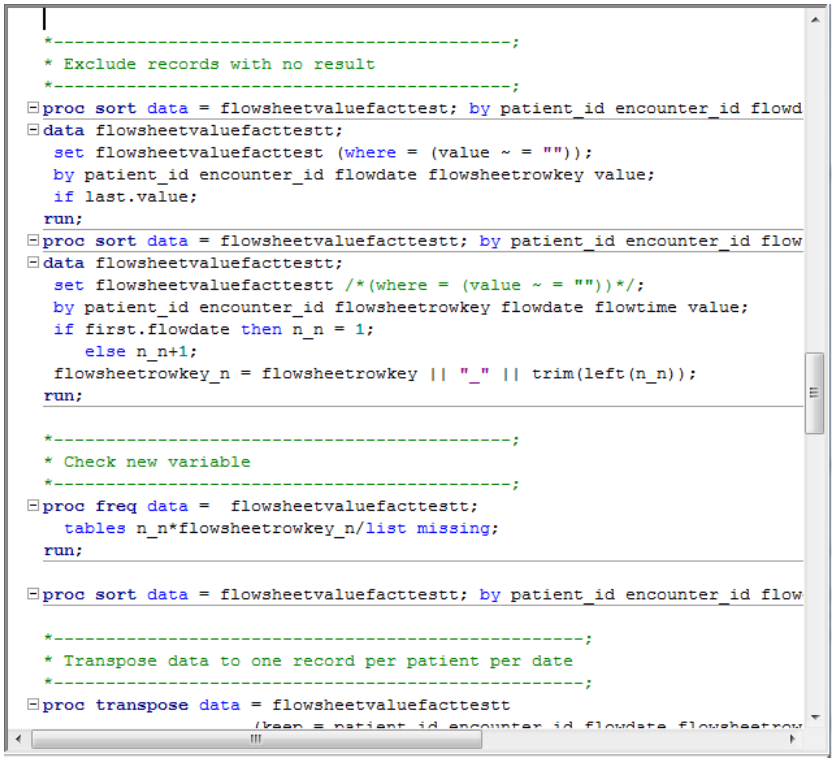
```
*-------------------------------------------------------;
* Calculate # foley re-insertions for NEW foley
*-------------------------------------------------------;
proc sort data = newfoleyno; by pat_mrn_id hosp_admsn_time placement_inst
data reinsertion;
   format lagplacement lagremoval hosp_admsn_time placement_instant mmddyy
   set newfoleyno;
   by pat_mrn_id hosp_admsn_time placement_instant format_removal placemen
   lagplacement = lag(placement_instant);
   lagremoval = lag(format_removal);
   if first.pat_mrn_id  then do;
         lagplacement = .;
         reinsertion = 0;
         lagremoval = .;
         end;
   if lagplacement~=. then reinsertion = 1;
   keep pat_mrn_id hosp_admsn_time placement_instant format_removal placem
 run;

proc sort data = reinsertion; by pat_mrn_id hosp_admsn_time placement_ins
data reinsertion2;
   set reinsertion;
   by pat_mrn_id hosp_admsn_time placement_instant format_removal placemen

   if reinsertion = 0 then osh = -1;
      else if hosp_admsn_time>placement_instant then osh = 1;
      else if hosp_admsn_time<=placement_instant then osh = 0;

  if hosp_admsn_time>placement_instant then newfoley = 0;
      else if hosp_admsn_time<=placement_instant then newfoley = 1;
```

Calculated Variable	Description
reinsertion	whether a foley catheter was reinserted during the hospitlization for any reason with any time period
new foley during hospitalization	placement time is after hospital admission time for patients that we admitted without a foley catheter
qmonth change	new placement time minus previous placement time >25 days

Step 7: Create Reproducible Code

Tips for Easy Reproducibility

- Read raw de-identified files directly into programming software
- Use only the program to manipulate the data
- Add text to describe the purpose of the program at a high level and imbed it in the code to describe the specific task
- Have an organized file structure
  /Data
  /Documents
  /Programs



```
*-------------------------------------------;
* Exclude records with no result
*-------------------------------------------;
proc sort data = flowsheetvaluefacttest; by patient_id encounter_id flowd
data flowsheetvaluefacttestt;
  set flowsheetvaluefacttest (where = (value ~ = ""));
  by patient_id encounter_id flowdate flowsheetrowkey value;
  if last.value;
run;
proc sort data = flowsheetvaluefacttestt; by patient_id encounter_id flow
data flowsheetvaluefacttestt;
  set flowsheetvaluefacttestt /*(where = (value ~ = ""))*/;
  by patient_id encounter_id flowsheetrowkey flowdate flowtime value;
  if first.flowdate then n_n = 1;
     else n_n+1;
  flowsheetrowkey_n = flowsheetrowkey || "_" || trim(left(n_n));
run;

*-------------------------------------------;
* Check new variable
*-------------------------------------------;
proc freq data =  flowsheetvaluefacttestt;
  tables n_n*flowsheetrowkey_n/list missing;
run;

proc sort data = flowsheetvaluefacttestt; by patient_id encounter_id flow

*-----------------------------------------------------;
* Transpose data to one record per patient per date
*-----------------------------------------------------;
proc transpose data = flowsheetvaluefacttestt
```

Step 8: Ask for Help

UCSF Resources

- Cores at UCSF
- UCSF Library https://www.library.ucsf.edu/
  - Workshops & Special Events
  - Data Science Training Opportunities
- CTSI's Consultation Services
  - Study design experts to help define covariates and outcomes from EHR
  - Programming experts to help translate your recipe into code
- And more …

Image credit: springcreekanimalhospital.com

Step 9: Prepare an Analytical Dataset and Codebook (see Step 6)

Merge study variables into one nice data frame

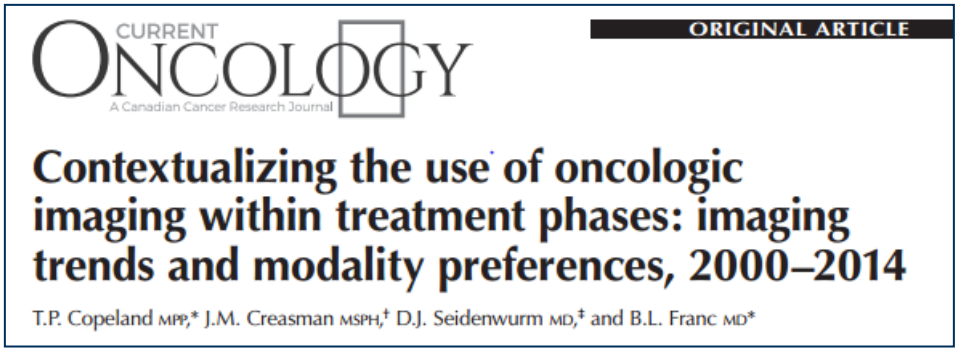
- Each row represents a different a observation, with potentially repeated observations for individual patients
- Each column represents a different variable in your dataset

	A	B	C	D	E	F	G	H	I	J	K
1	arrival_dt	arrival_time	admission	random_n	random_e	age_at_er	shifted_e	shifted_in	er_time	hx_asthm	hx_bronch
2	9/13/2014	13SEP14:16:30:00	13SEP14:2	136	64441589	61	13SEP14:1	13SEP14:2	0	0	0
3	8/4/2016	04AUG16:14:20:00	04AUG16:	7146	12583	35	04AUG16:	04AUG16:	0	1	0
4	12/5/2014	05DEC14:10:22:00	05DEC14:1	7303	50842310	60	05DEC14:1	05DEC14:1	0	0	0
5	11/30/2015	30NOV15:16:00:0(	30NOV15:	16677	66702476	64	30NOV15:	30NOV15:	0	0	0
6	2/27/2014	27FEB14:14:26:00	27FEB14:2	17111	80507157	67	27FEB14:1	27FEB14:2	0	0	0
7	10/20/2012	20OCT12:22:32:00	21OCT12:(	17297	40976111	84	20OCT12:2	21OCT12:(	1	0	0
8	10/9/2012	09OCT12:09:26:00	09OCT12:1	17297	54708429	84	09OCT12:(	09OCT12:1	0	0	0
9	3/3/2012	03MAR12:17:55:0(	03MAR12:	17297	66144483	84	03MAR12:	03MAR12:	0	0	0
10	11/15/2013	15NOV13:20:28:0(	15NOV13:	18230	26858826	36	15NOV13:	15NOV13:	0	0	0
11	10/5/2015	05OCT15:09:45:00	05OCT15:1	19879	8078692	69	05OCT15:(	05OCT15:1	0	0	0
12	9/16/2015	16SEP15:17:18:00	16SEP15:2	19879	80654919	69	16SEP15:1	16SEP15:2	0	0	0
13	8/18/2016	18AUG16:19:48:00	18AUG16:	20049	20628032	38	18AUG16:	18AUG16:	0	0	0
14	8/3/2016	03AUG16:16:28:00	04AUG16:(	22791	2820757	67	03AUG16:	04AUG16:(	1	0	0
15	4/14/2016	14APR16:14:09:00	15APR16:2	22791	35613000	67	14APR16:1	15APR16:2	1	0	0
16	8/1/2013	01AUG13:19:25:00	02AUG13:(	23587	2447391	92	01AUG13:	02AUG13:(	1	0	1
17	1/5/2015	05JAN15:06:23:00	05JAN15:1	23587	24285348	94	05JAN15:(	05JAN15:1	0	0	1
18	5/6/2013	06MAY13:19:59:0(	06MAY13:	23587	39156222	92	06MAY13:	06MAY13:	0	0	1
19	6/13/2016	13JUN16:14:57:00	13JUN16:2	23587	51344676	95	13JUN16:1	13JUN16:2	0	1	1

Step 10: Analyze and Publish Your Results

Easier said than done!

- Project started 1/2015, submitted 4/2016 and published 4/2017

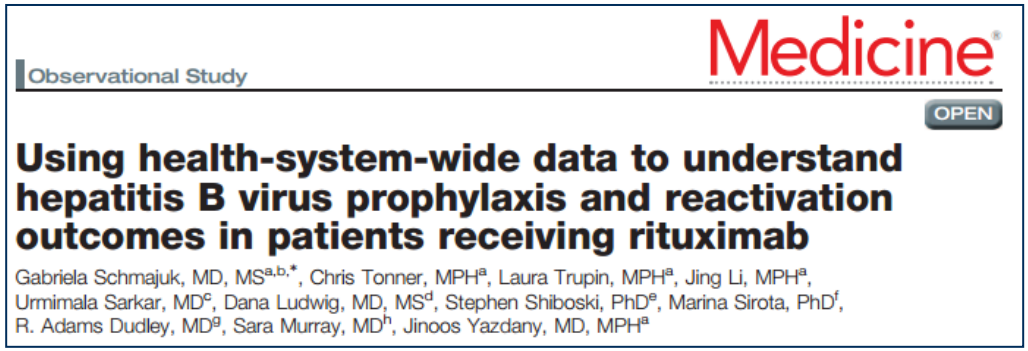

CURRENT ONCOLOGY
A Canadian Cancer Research Journal

ORIGINAL ARTICLE

Contextualizing the use of oncologic imaging within treatment phases: imaging trends and modality preferences, 2000–2014

T.P. Copeland MPP,* J.M. Creasman MSPH,† D.J. Seidenwurm MD,‡ and B.L. Franc MD*

- Project started 6/2016 and was published 3/2017


Observational Study

Medicine

OPEN

Using health-system-wide data to understand hepatitis B virus prophylaxis and reactivation outcomes in patients receiving rituximab

Gabriela Schmajuk, MD, MS[a,b,*], Chris Tonner, MPH[a], Laura Trupin, MPH[a], Jing Li, MPH[a], Urmimala Sarkar, MD[c], Dana Ludwig, MD, MS[d], Stephen Shiboski, PhD[e], Marina Sirota, PhD[f], R. Adams Dudley, MD[g], Sara Murray, MD[h], Jinoos Yazdany, MD, MPH[a]

Real-time Feedback

On your phone, tablet, laptop – Go to:

slido.com

Enter event code:

clinicaldata