Social Media Data Analysis

Raw Data retrieval

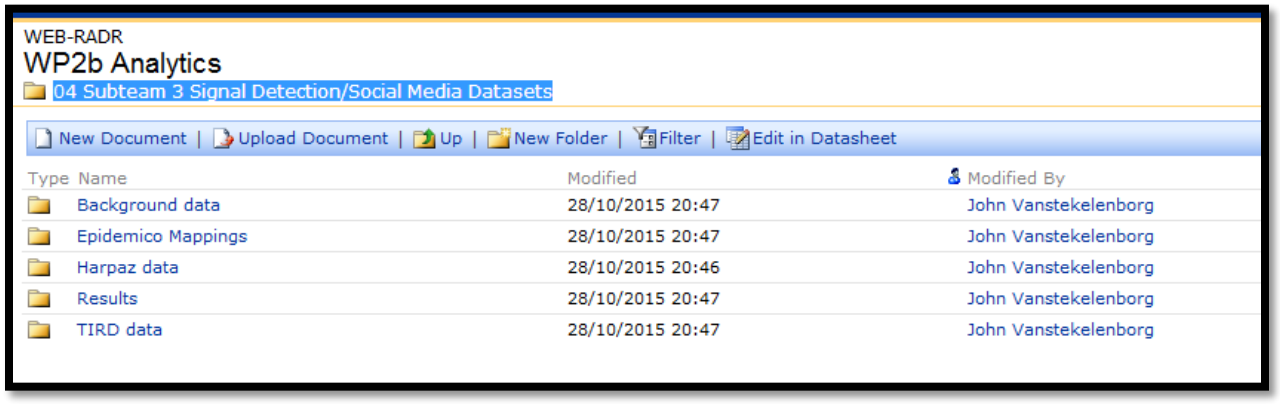
The social media datasets can be retrieved from https://sps-ext.nibsc.ac.uk/MHRA/imi

Navigate to:

- Documents and Lists
- WP2b Analytics
- 04 Subteam 3 Signal Detection/Social Media Datasets

The "raw" social media data files (ie each record is a post) can be found in the folders:

- TIRD data
- Harpaz data



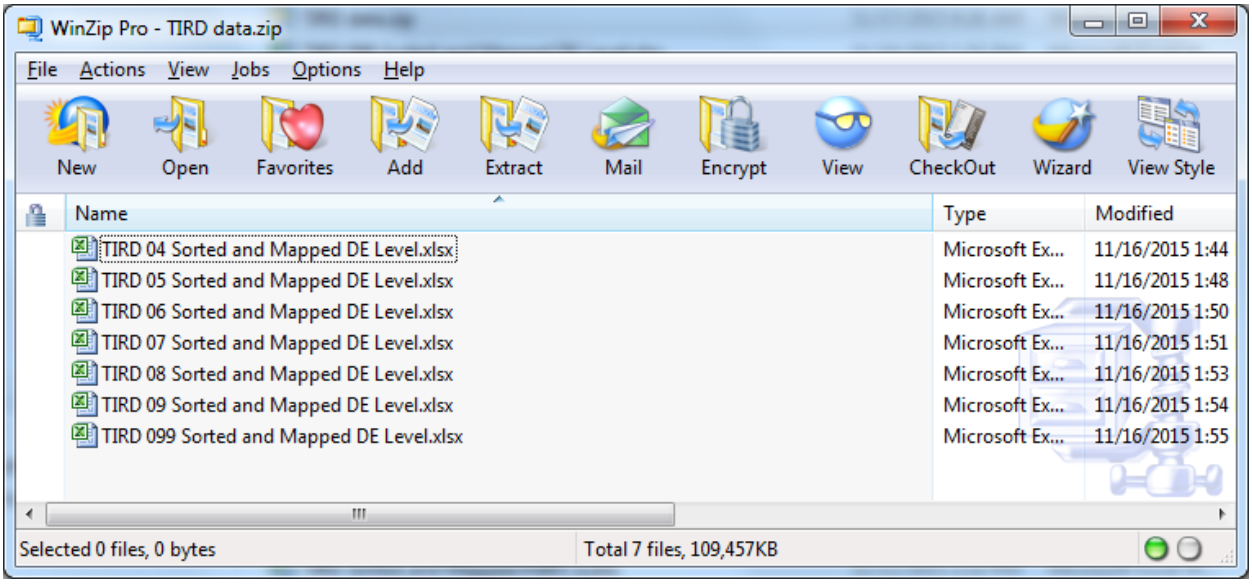
There is NO need to retrieve any Harpaz data, as the analysis will be done centrally by UMC.

The TIRD raw data is stored in a ZIP file that consists of 7 Excel files, see screenshot below. Each file contains social media post information at the unique DRUG/EVENT level ie each row has one drug (which is one of the TIRD drugs), and one PT.

Each file corresponds to a different indicator score cutoff:

- TIRD 04 Sorted and Mapped DE Level.xlsx: indicator score >=0.4
- TIRD 05 Sorted and Mapped DE Level.xlsx: indicator score >=0.5
- TIRD 06 Sorted and Mapped DE Level.xlsx: indicator score >=0.6
- TIRD 07 Sorted and Mapped DE Level.xlsx: indicator score >=0.7
- TIRD 08 Sorted and Mapped DE Level.xlsx: indicator score >=0.8
- TIRD 09 Sorted and Mapped DE Level.xlsx: indicator score >=0.9
- TIRD 099 Sorted and Mapped DE Level.xlsx: indicator score >=0.99

.



Again: there is NO need to retrieve any Harpaz data, as the analysis will be done centrally by UMC.

For completeness (and benefit of UMC), the appendix contains a description of the Harpaz raw data sets and the Harpaz results data sets.

SDR results Data retrieval

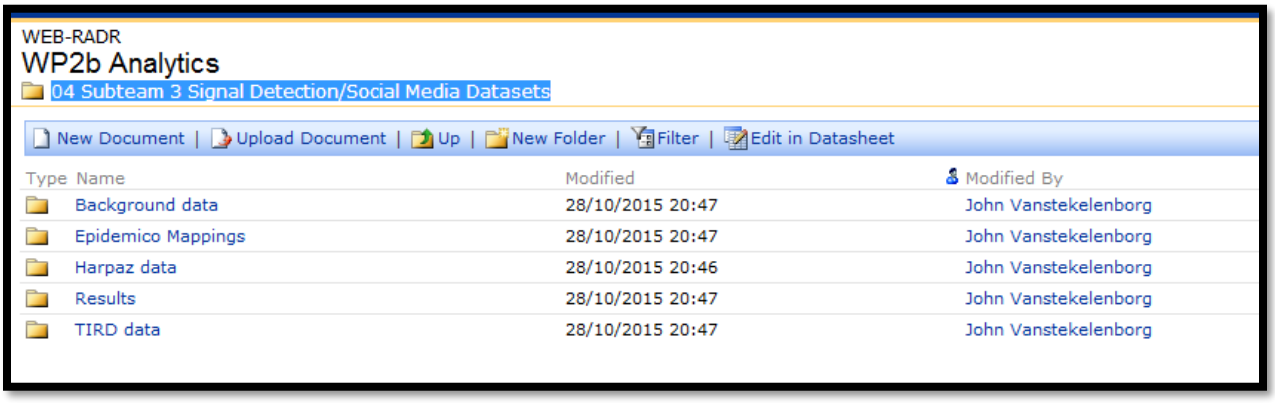
The SDR calculations (PRR, IC etc) can also be retrieved from https://sps-ext.nibsc.ac.uk/MHRA/imi

Navigate to:

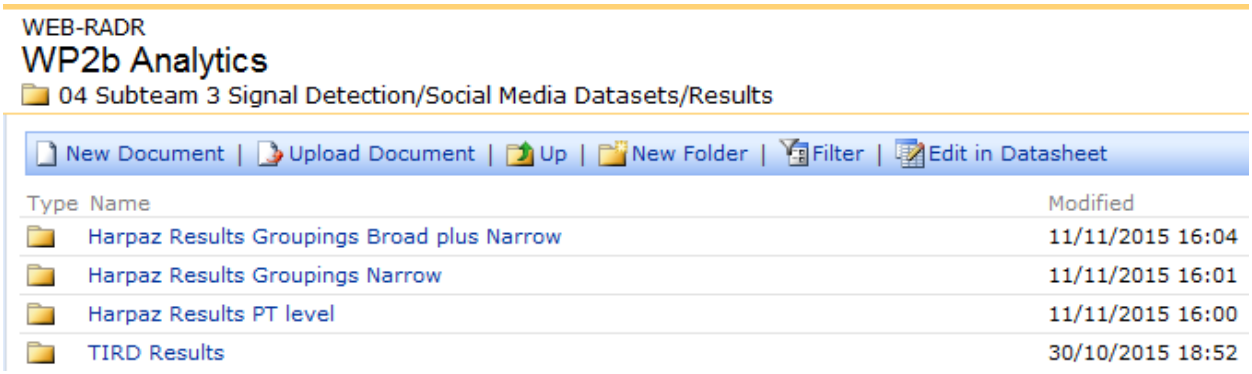
- Documents and Lists
- WP2b Analytics
- 04 Subteam 3 Signal Detection/Social Media Datasets

The results files can be found in (surprise) the folder:

- Results (see screenshot below)

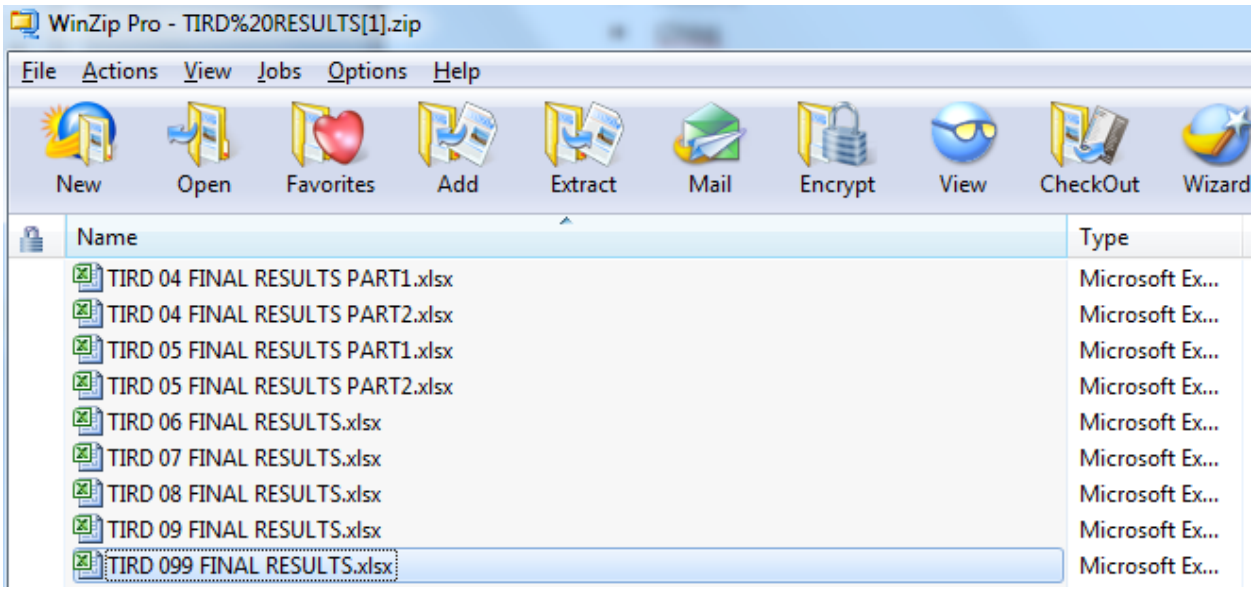


There are 4 results files:



Note that you will ONLY need to retrieve the ***TIRD Results***. You can ignore the other results files.

Each of the directories contains a ZIP file. The TIRD results ZIP file contains:



Where each of the Excel files again corresponds to an indicator cutoff.

NOTES:

- Due to its size, the result files with cutoff 0.5 had to be split into two: TIRD 05 FINAL RESULTS PART1.xlsx and TIRD 05 FINAL RESULTS PART2.xlsx
- Due to its size, the result files with cutoff 0.4 had to be split into two: TIRD 04 FINAL RESULTS PART1.xlsx, and TIRD 04 FINAL RESULTS PART2.xlsx

AGAIN: You will ONLY need to retrieve the *TIRD results* files.

Determination of earliest social media post-date of a Drug/Event combination – TIRD data

Let's assume the TIRD 07 Sorted and Mapped DE Level.xlsx is retrieved. The following attributes are visible in the Excel file:

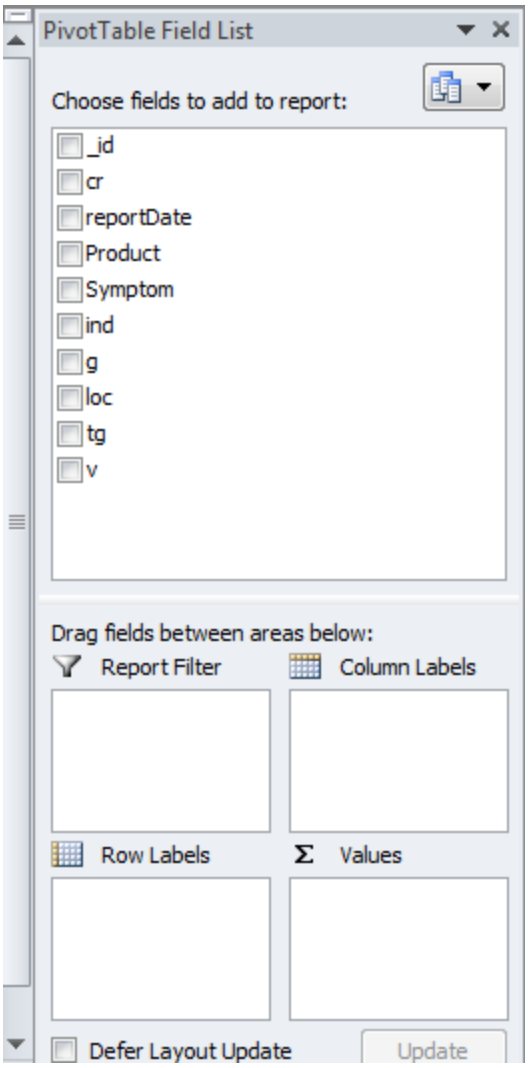
- _id
- cr
- reportDate
- Product
- Symptom
- Ind
- g
- loc
- tg
- v

See below for a screenshot:

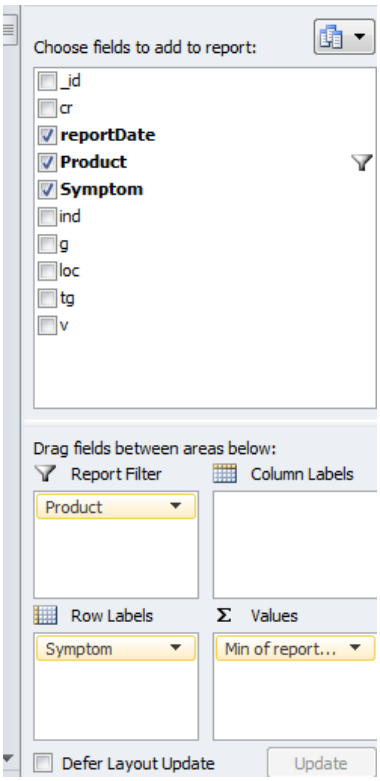
_id	cr	reportDate	Product	Symptom	ind	g	loc	tg	v
175142857689149440	2012-03-01T08:58:05Z	3/1/2012	levetiracetam	Seizure	0.74674		madison. ga	2	FALSE
175143267116126209	2012-03-01T08:59:43Z	3/1/2012	levetiracetam	Loss of consciousness	0.72644		moco.	2	FALSE
175185415303741440	2012-03-01T11:47:12Z	3/1/2012	topiramate	Somnolence	0.88631	m		2	FALSE
175185415303741440	2012-03-01T11:47:12Z	3/1/2012	topiramate	Headache	0.88631	m		2	FALSE
175335173821440000	2012-03-01T21:42:17Z	3/1/2012	diclofenac	Malaise	0.91173	m		2	FALSE
175436256245854208	2012-03-02T04:23:57Z	3/2/2012	topiramate	Drug ineffective	0.73578	m	Lindale, Tx	2	FALSE
812478153_10150830698358154	2012-03-04T00:43:07Z	3/4/2012	topiramate	Drug ineffective	0.79025	f		2	FALSE
812478153_10150830698358154	2012-03-04T00:43:07Z	3/4/2012	topiramate	Dysgeusia	0.79025	f		2	FALSE
812478153_10150830698358154	2012-03-04T00:43:07Z	3/4/2012	topiramate	Nonspecific reaction	0.79025	f		2	FALSE

The goal is to find the EARLIEST reported date of a particular Drug/Event pair for a given TIRD drug. Let's assume we are interested in topiramate, then do the following:

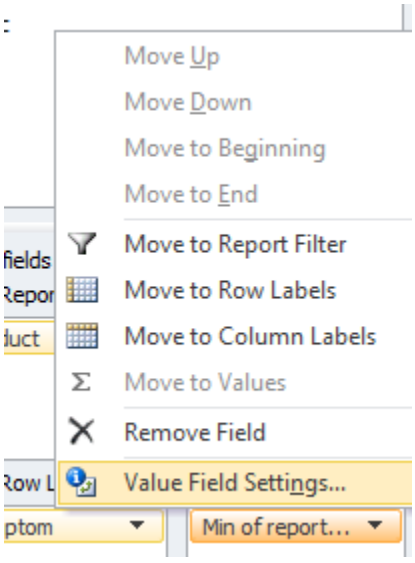
1. Highlight all 10 columns
2. Click INSERT → PivotTable
3. Click NEW WORKSHEET, then OK
4. A new worksheet is created with a pivot table, see screenshot of the pivot table variables:



5. Set up the pivottable by:
    a. Select Product as Report Filter
    b. Select Symptom as Row Label
    c. Select reportDate as Values



6. Then change “count of reportDate” to “Min of reportDate” by clicking on the field, and then clicking “Value Field Settings” and selecting “Min” from the selection box.

a. 

7. The pivotttable will show. Make sure to change Format of the column “min of ReportDate” to Date, and choose dd-mm-yy
8. Then select the drug of interest in the Filter field, eg topiramate:

	A	B
1	Product	topiramate
2		
3	**Row Labels**	**Min of reportDate**
4	Abdominal discomfort	28-Mar-12
5	Abdominal distension	13-Feb-13
6	Abdominal pain	29-Mar-12
7	Abdominal symptom	29-Mar-12
8	Abnormal dreams	23-Mar-12
9	Abortion spontaneous	26-Apr-13
10	Abscess	14-Jan-15
11	Acne	18-Jan-13
12	Adolescence	20-Jan-13
13	Affect lability	1-May-12
14	Ageusia	16-Aug-13
15	Aggression	10-Oct-12
16	Agitation	29-Dec-12

The net result is the earliest date of a social media post for the product and event of interest eg topiramate/Abdominal discomfort: 28-March-2012

These Drug/Event dates can then be compared to the reference standard index dates and to spontaneous data earliest report dates.

Repeat this analysis for the other raw TIRD datasets, corresponding to different indicator score thresholds.

IMPORTANT: PLEASE SAVE THE PIVOTTABLES, AS THEY WILL BE USED FOR FUTURE RESEARCH

Determination of earliest SDR date of a Drug/Event combination – TIRD data

Let's assume the TIRD 07 FINAL RESULTS.xlsx Excel file is retrieved. The following attributes are visible in the Excel file:

- num
- MONTHYEAR
- Drug
- Event
- a
- ab
- ac
- abcd
- dateEventKey
- b
- c
- d
- RRR
- PRR
- PRR025
- PRR975
- Chisq
- IC
- IC025

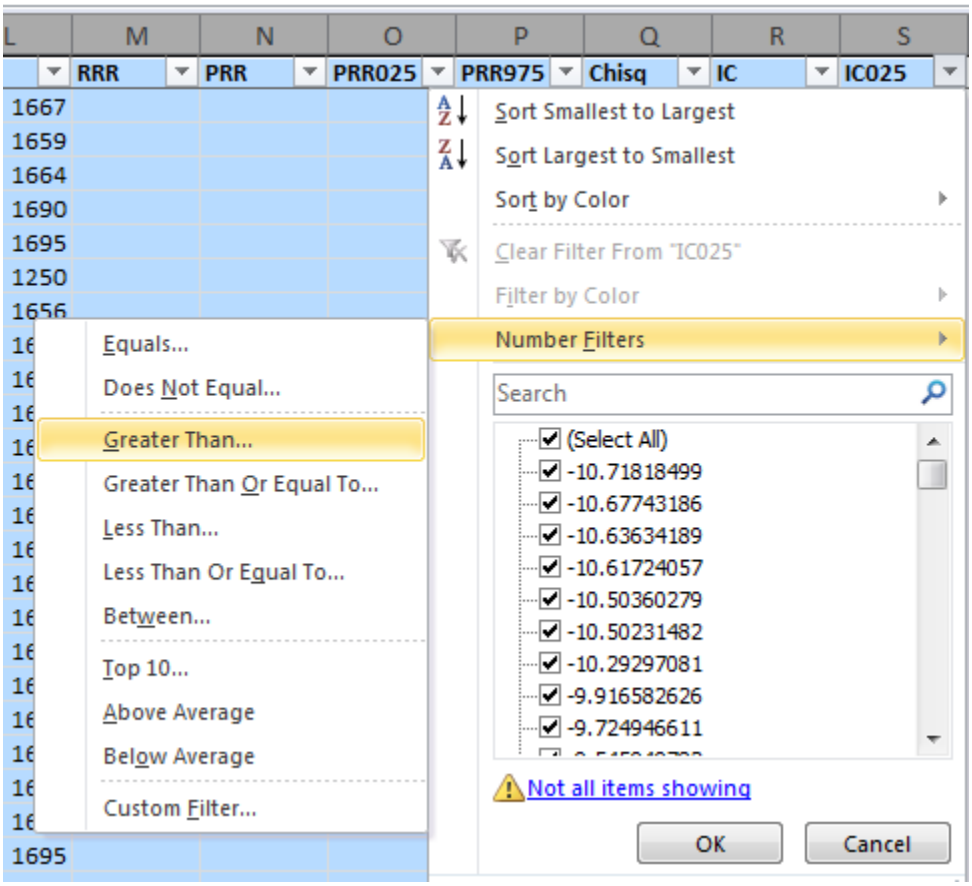
See below for a screenshot:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	num	MONTHYEA	Drug	Event	a	ab	ac	abcd	dateEventK	b	c	d	RRR	PRR	PRR025	PRR975	Chisq	IC	IC025
393565	393564	Sep-13	diclofenac	Injection site pain	34	2667	290	28142	09/2013Inj	2633	256	25219	1.237123	1.268616	0.889052	1.810227	1.724855	0.302039	-0.2238
393566	393565	Sep-13	diclofenac	Increased appetite	9	2667	369	28142	09/2013Inc	2658	360	25115	0.257364	0.238798	0.123394	0.462133	21.58834	-1.9006	-2.99366
393567	393566	Sep-13	diclofenac	Weight decreased	5	2667	511	28142	09/2013We	2662	506	24969	0.103248	0.094387	0.039152	0.227543	43.81602	-3.15313	-4.6806
393568	393567	Sep-13	diclofenac	Mood swings	0	2667	97	28142	09/2013Mo	2667	97	25378							
393569	393568	Sep-13	diclofenac	Sleep disorder	0	2667	11	28142	09/2013Sle	2667	11	25464							
393570	393569	Sep-13	diclofenac	Drug reaction with eosi	1	2667	17	28142	09/2013Dr	2666	16	25459	0.620702	0.596996	0.079203	4.499856	0.256202	-0.49302	-4.29025
393571	393570	Sep-13	diclofenac	Flatulence	7	2667	47	28142	09/2013Fla	2660	40	25435	1.571564	1.671588	0.749582	3.727684	1.610136	0.59825	-0.66199
393572	393571	Sep-13	diclofenac	Bipolar disorder	0	2667	142	28142	09/2013Bip	2667	142	25333							
393573	393572	Sep-13	diclofenac	Abdominal distension	4	2667	21	28142	09/2013Ab	2663	17	25458	2.009892	2.247513	0.756817	6.674423	2.243892	0.853688	-0.88307
393574	393573	Sep-13	diclofenac	Tremor	3	2667	196	28142	09/2013Tre	2664	193	25282	0.161509	0.148476	0.047502	0.464086	14.5277	-2.44624	-4.49668
393575	393574	Sep-13	diclofenac	Hypoaesthesia	44	2667	216	28142	09/2013Hy	2623	172	25303	2.149467	2.443517	1.758974	3.394465	30.1094	1.085466	0.627506
393576	393575	Sep-13	diclofenac	Tinnitus	0	2667	26	28142	09/2013Tin	2667	26	25449							
393577	393576	Sep-13	diclofenac	Weight increased	9	2667	418	28142	09/2013We	2658	409	25066	0.227195	0.210189	0.108716	0.406374	26.52921	-2.07809	-3.17116
393578	393577	Sep-13	diclofenac	Drug dose omission	7	2667	286	28142	09/2013Dr	2660	279	25196	0.258264	0.239654	0.113321	0.506826	16.64216	-1.87992	-3.14015
393579	393578	Sep-13	diclofenac	Feeling hot	0	2667	16	28142	09/2013Fe	2667	16	25459							
393580	393579	Sep-13	diclofenac	Blindness	1	2667	50	28142	09/2013Bli	2666	49	25426	0.211039	0.194937	0.026929	1.41112	3.264094	-1.80418	-5.60141

We will now construct a pivottable, similar to the one constructed for finding the earliest post-date.

IC025 table

1. First hit the FILTER button
2. Then click on the filter arrow in the IC025 column, click on “Number Filters”, then “Greater Than”



Enter the threshold of IC025 > 0, click OK

3. Now select all filtered data, and copy to another sheet, see screenshot:

	num	MONTHYEA	Drug	Event	a	ab	ac	abcd	dateEventK	b	c	d	RRR	PRR	PRR025	PRR975	Chisq	IC	IC025
2	1155	Mar-12	atenolol	Fatigue	4	18	44	1698	03/2012Fa	14	40	1640	8.575758	9.333333	3.73107	23.34749	27.77603	2.219186	0.482429
3	1723	Mar-12	Baclofen	Muscle spa	3	60	5	1698	03/2012Mu	57	2	1636	16.98	40.95	6.971303	240.5436	46.90743	2.370813	0.320369
4	2881	Mar-12	carbamaze	Pain	21	160	105	1698	03/2012Pa	139	84	1454	2.1225	2.403125	1.532894	3.76739	14.67059	1.048587	0.365695
5	2908	Mar-12	carbamaze	Bipolar dis	6	160	11	1698	03/2012Bi	154	5	1533	5.788636	11.535	3.560101	37.37428	26.41208	2.080779	0.70479
6	2950	Mar-12	carbamaze	Dyspnoea	5	160	12	1698	03/2012Dy	155	7	1531	4.421875	6.866071	2.204635	21.38356	14.72154	1.753903	0.226437
7	2969	Mar-12	carbamaze	Neuropathy	4	160	4	1698	03/2012Ne	156	0	1538	10.6125				38.54079	2.359418	0.62266
8	4614	Mar-12	clozapine	Weight inc	7	23	46	1698	03/2012We	16	39	1636	11.2344	13.07135	6.547428	26.09577	68.00231	2.739422	1.479187
9	5687	Mar-12	diclofenac	Toothache	4	92	9	1698	03/2012To	88	5	1601	8.202899	13.96522	3.814018	51.13434	26.89103	2.187879	0.451121
10	5710	Mar-12	diclofenac	Alcohol use	4	92	11	1698	03/2012Al	88	7	1599	6.711462	9.975155	2.973461	33.46394	20.68961	2.037683	0.300926
11	5716	Mar-12	diclofenac	Pain	25	92	105	1698	03/2012Pa	67	80	1526	4.39441	5.455163	3.668058	8.112958	73.87238	2.042708	1.421733
12	9694	Mar-12	insulin gla	Therapy ch	3	20	20	1698	03/2012Th	17	17	1661	12.735	14.80588	4.709391	46.5483	33.21849	2.250418	0.199975

4. Move to this new sheet with the filtered data only.

5. This new sheet now only contains SDRs that have an IC025>0 , ie these are “real alerts”. Click INSERT → PivotTable
6. Click NEW WORKSHEET, then OK
7. A new worksheet is created with a pivot table
8. Set up the pivottable by:
    a. Select Drug as Report Filter
    b. Select Event as Row Label
    c. Select MONTHYEAR as Values
9. Then change “count of MONTHYEAR” to “Min of MONTHYEAR” by clicking on the field, and then clicking “Value Field Settings” and selecting “Min” from the selection box.
10. The pivotttable will show. Make sure to change Format of the column “min of MONTHYEAR” to Date, and choose March-01
11. Then select the drug of interest in the Filter field, eg topiramate. The output table will look something like:

	A	B
1	Drug	topiramate
2		
3	**Row Labels**	**Min of MONTHYEAR**
4	Abdominal discomfort	November-12
5	Abdominal symptom	February-14
6	Abortion spontaneous	February-14
7	Adolescence	February-14
8	Affect lability	March-14
9	Ageusia	December-14
10	Aggression	January-15
11	Agitation	July-14
12	Alcohol use	February-14
13	Alopecia	June-12
14	Amnesia	June-12

Ie Abdominal discomfort first had an IC025>0 for topiramate in November of 2012.

PRR tables

The following "PRR" alert flavors will need to be constructed:

- PRR>=2; N>=3
- PRR>= 2; N>=3; Chisq>=4
- Lower 95% CI of PRR >=1; N>=3

Outlined below is the construction for the "PRR>= 2; N>=3; Chisq>=4" scenario. The other two scenarios are similar.

1. Go back to the original data, again hit the FILTER button
2. Click on the filter arrow in the PRR column, click on "Number Filters", then "Greater Than OR Equal To", then enter "2" as the threshold
3. Click on the filter arrow in the 'a' column, click on "Number Filters", then "Greater Than OR Equal To", then enter "3" as the threshold
4. Click on the filter arrow in the Chisq column, click on "Number Filters", then "Greater Than OR Equal To", then enter "4" as the threshold
5. Copy the filtered data to a new sheet. Move to this new sheet.
6. Then follow the pivottable directions above

The net result should again be something like:

Drug	diclofenac
Row Labels	**Min of MONTHYEAR**
Abdominal discomfort	March-13
Abdominal distension	June-13
Abdominal pain	May-12
Abdominal pain upper	December-14
Abdominal symptom	April-12
Alcohol use	March-12
Arthralgia	May-12
Arthritis	October-12
Back pain	March-12
Burning Sensation	March-14
Childhood	December-14
Confusional state	July-14

Repeat this analysis for the other raw TIRD datasets, corresponding to different indicator score thresholds.

IMPORTANT: PLEASE SAVE THE PIVOTTABLES, AS THEY WILL BE USED FOR FUTURE RESEARCH

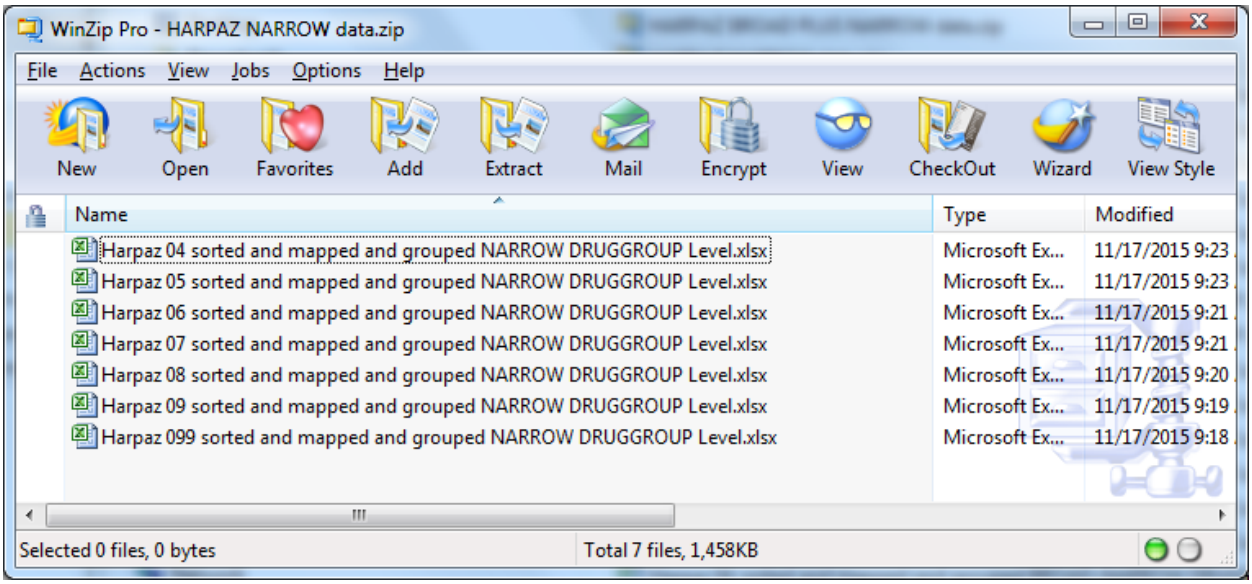
Appendix: Harpaz data sets

NOTE: EFPIA partners do not need to retrieve or analyze any of the Harpaz data. UMC will perform the analysis.

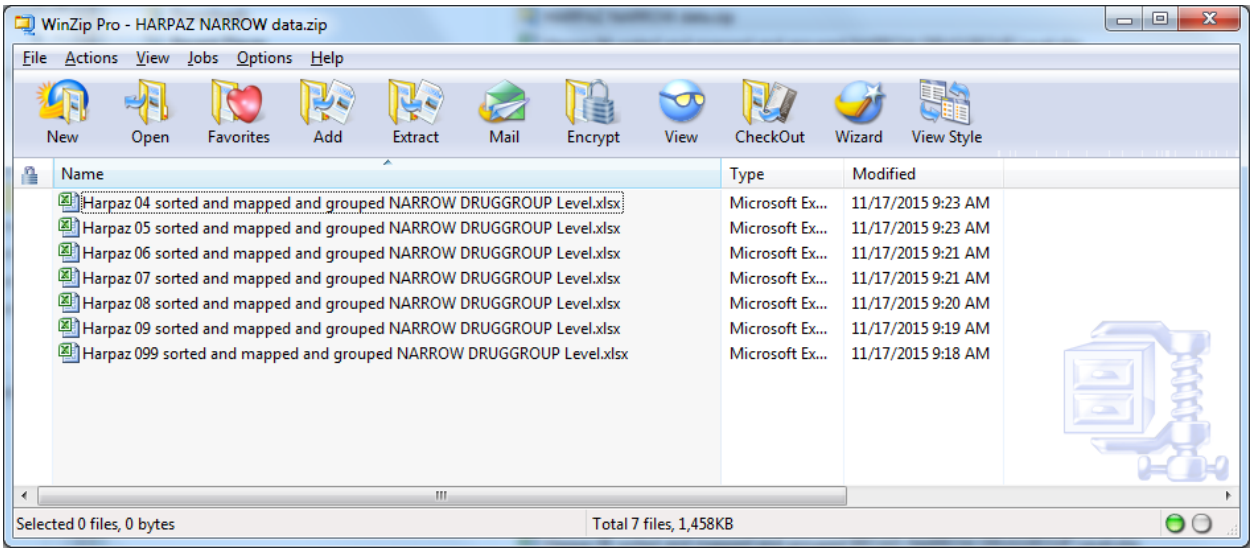
Raw data

There are TWO separate Harpaz raw data sets , each one stored in a ZIP file with 7 Excels.

The first ZIP file is called: HARPAZ NARROW data.zip. The “NARROW” refers to the narrow definitions of medical concepts in the Harpaz paper.



The second ZIP file is called: HARPAZ BROAD PLUS NARROW data.zip. The “BROAD PLUS NARROW” refers to the definitions of medical concepts in the Harpaz paper which include both the ‘broad” terms and “narrow” terms.



The Harpaz 04, 05 etc designations refer to the indicator threshold used.

Determination of earliest post-date of a Drug/Event combination – Harpaz data

For the Harpaz data, the same steps can be followed as in the TIRD analysis with the following exception:

- The "event" concept in the Harpaz data is defined at the MEDICAL CONCEPT level. When you open one of the Harpaz raw data sets, an additional column is visible. For example, in the *Harpaz 07 sorted and mapped and grouped NARROW DRUGGROUP Level.xlsx* dataset, the column "GROUP NARROW" is visible.
- All actions outlined above for the TIRD data, should be repeated with the "GROUP NARROW" taking the place of "Symptom", ie we want to find the earliest report Date for each of the Narro medical concepts.

	A	B	C	D	E	F	G	H	I	J	K
1	_id	cr	reportDate	Product	Symptom	ind	g	loc	tg	v	GROUP NARROW
2	180402114	2012-03-15T21:16:30Z	3/15/2012	levocetirizine	Dyspnoea	0.89209	f	Tompkinsvi	2	FALSE	Dyspnea
3	715974916	2012-03-20T06:18:27Z	3/20/2012	levetiracetam	Drug reaction with e	0.91898	f		2	FALSE	Drug reaction with eosinophilia and systemic symptoms (DRESS)
4	183049703	2012-03-23T04:37:04Z	3/23/2012	levocetirizine	Pruritus	0.91355			2	FALSE	Pruritis
5	183440676	2012-03-24T06:30:40Z	3/24/2012	anastrozole	Pruritus	0.71228		SC	2	FALSE	Pruritis
6	185963508	2012-03-31T05:35:30Z	3/31/2012	levocetirizine	Urticaria	0.72262		Los Feliz, Lc	2	FALSE	Urticaria
7	186024677	2012-03-31T09:38:34Z	3/31/2012	levocetirizine	Pruritus	0.89131		Da Nola/ye	2	FALSE	Pruritis
8	188860433	2012-04-08T05:26:50Z	4/8/2012	levocetirizine	Urticaria	0.75924	f	Bahrain !	2	FALSE	Urticaria
9	188929617	2012-04-08T10:01:45Z	4/8/2012	ketoconazole	Pruritus	0.94792		CAN/PHL	2	FALSE	Pruritis
10	189463944	2012-04-09T21:24:59Z	4/9/2012	levocetirizine	Anaphylactic shock	0.82352		Brooklyn, N	2	FALSE	Anaphylaxis
11	189587546	2012-04-10T05:36:08Z	4/10/2012	levocetirizine	Pruritus	0.81061	f		2	FALSE	Pruritis
12	190069723	2012-04-11T13:32:08Z	4/11/2012	clozapine	Pruritus	0.9335	m		2	FALSE	Pruritis
13	192318202	2012-04-17T18:26:47Z	4/17/2012	levocetirizine	Urticaria	0.83993		Baltimore M	2	FALSE	Urticaria
14	193198310	2012-04-20T04:44:01Z	4/20/2012	levocetirizine	Dyspnoea	0.70408	m		2	FALSE	Dyspnea
15	193456886	2012-04-20T21:51:30Z	4/20/2012	lisinopril	Raynaud's phenome	0.80122		Madison W	2	FALSE	Peripheral vascular disorder

The output of the exercise should look like:

Product	methylphenidate
Row Labels	**Min of reportDate**
Dyspnea	29-Jan-15
Hallucinations	2-Mar-13
Hematopoietic disorders	6-Mar-13

UMC should do this analysis only for the "HARPAZ NARROW data" files, and NOT for the "HARPAZ BROAD PLUS NARROW data"

Determination of earliest SDR date of a Drug/GROUP combination – Harpaz data

The procedure is exactly the same as fr the TIRD data. The Harpaz SDR analysis was done at the PT level, Narrow group level , and Broad+Narrow level. As discussed before, PT level analysis and the Broad+Narrow Analysis can be ignored.

For the Narrow results files, the data format is exactly the same as for the TIRD data. The "EVENT" column, however, does not show individual PTs, but the medical concept group names that were used to calculate the SDRs.