115632 – WEB-RADR

WEB- Recognising Adverse Drug Reactions

WP2B – Analytics

D2B.3 Technical report describing implementation and evaluation of safety signal detection in social media

Lead contributor	John van Stekelenborg (#13 – JPNV)
	JVanstek@its.jnj.com

Due date	Month 37
Delivery date	Month 37
Deliverable type	Report
Dissemination level	PP[1]

Description of Work	Version	Date
	V1.0	Month 37

Document History

Version	Date	Description
V0.1	Month 18	First Draft
V1.0	Month 37	Final Version

[1] Please choose the appropriate reference and delete the rest:
PU = Public
PP = Restricted to other programme participants (including the Commission Services)
RE = Restricted to a group specified by the consortium (including the Commission Services)
CO = Confidential, only for members of the consortium (including the Commission Services)

Summary

The two reference datasets used in assessment of signal detection algorithms are the published “Harpaz dataset” and the Time Indexed Reference Dataset (TIRD), which comprises a subset of 38 products proposed by the EFPIA Partners in WEB-RADR.

The Harpaz dataset comes with its own published set of positive and negative controls, i.e. product-event combinations (PECs) that are considered associated and not associated, respectively.

Protocol for the analysis of WEB-RADR social media data

The two reference datasets used in assessment of signal detection algorithms are the published "Harpaz dataset" and the Time Indexed Reference Dataset (TIRD), which comprises a subset of 38 products proposed by the EFPIA Partners in WEB-RADR.
The Harpaz dataset comes with its own published set of positive and negative controls, i.e. product-event combinations (PECs) that are considered associated and not associated, respectively.
==========
For the TIRD data, the following definitions for positive controls (and their index dates) and negative controls are proposed:

1. Each EFPIA company needs to generate a list of POSITIVE and NEGATIVE CONTROLs for its TIRD product(s).

   Note: Controls (positive or negative) must have at least two posts in the Epidemico data extraction at index score 0.7 OR at least two reports in VigiBase.

2. POSITIVE CONTROL:
   - **Definition:** ***PEC (on PT level) that has been identified by the EFPIA company as validated signal the first time in the period between 01-Mar-2012 and 31-Mar-2015, and that has at least two posts in the Epidemico data extraction at index score 0.7 OR at least two reports in VigiBase.***
   - **Notes:**
     - 'Validated signals' (ie signals that require a further in-depth evaluation) are those PECs that are of significant concern and that are subsequently investigated further to determine whether these events should be added to the label, RMP, or other action. There is some variability in process among the EFPIA companies, but the concept of a 'validated signal' is common.
     - For the case/post count requirement:
       - At least 2 social media posts exist in the Epidemico dataset at a cutoff of 0.7.
         (Note: It is sufficient to look at the final month only to determine this.)

       or

       - At least 2 cases exist in VigiBase dataset provided by UMC.
         (Note: It is sufficient to look at the final month only to determine this.)

3. INDEX DATE of a POSITIVE CONTROL:
   - **Definition:** ***Date on which the POSITIVE CONTROL was declared a 'VALIDATED SIGNAL' regardless of the date and type of trigger; the date must fall between 01-Mar-2012 and 31-Mar-2015.***

4. NEGATIVE CONTROL:
   - **Definition:** ***PEC (on PT level) that is not contained in HLTs linked to POSITIVE-CONTROL-PTs or listed/labelled PTs, and that has at least two posts in the Epidemico data extraction at index score 0.7 OR at least two reports in VigiBase.***
   - Steps to create the list of NEGATIVE CONTROLS: For each TIRD product,
     - Retrieve a list of PTs that have
       - At least 2 social media posts exist in the Epidemico dataset at a cutoff of 0.7.
         (Note: It is sufficient to look at the final month only to determine this.)

       or

       - At least 2 cases exist in VigiBase dataset provided by UMC.
         (Note: It is sufficient to look at the final month only to determine this.)

- From the resulting list, delete all PTs that are positive controls and all neighbouring PTs-within-HLT (i.e. PTs in HLTs that are linked to positive-control-PTs).
- From the remaining list, delete all listed/labelled PTs and all neighbouring PTs-within-HLT (i.e. PTs in HLTs that are linked to listed/labelled PTs).
- Keep all PTs resulting from this approach, even if the number of NEGATIVE CONTROLs may be much higher than the number of POSITIVE CONTROLs.

Description of Analyses

The analyses are carried out at the single case/post level and at the level of SDRs (signals of disproportionate reporting):

Single case/post comparisons (only applies to the TIRD data):

a) For those spontaneous cases for which the first report was received between the min/max dates, calculate what percentage was also present in social media as a post (aka proto-AE), in effect using the existence of a spontaneous case as a positive control at the case level. This would be a type of PPV.

b) Calculate what percentage of social media proto-AEs were NEVER reported as a spontaneous case. I hesitate to call these "social media false positives", as it implies that they should indeed never have been posted and the absence from a spontaneous database implies a true negative control. However, we will probably see a significant increase of these pseudo-false-positives for lower thresholds.

c) Time gain at the single case/post level:

Time gain can only be calculated for those PECs that are POSITIVE CONTROLs ***and*** have been first reported in the company safety database and in social media in the period of interest (01 Mar 2012 to 31 Mar 2015).

Time gain is relative to the INDEX DATE. As an example, assume a POSITIVE CONTROL has an INDEX DATE of January 2014. If the first spontaneous report comes in January2013, then that's -12 months relative to this date; if the first social media post comes in March 2014, then that's +2 months; (same applies to SDRs, see below). This convention establishes a standard 'objective' reference point.

Note: many (or all) positive controls for a product might have been first reported in the company safety database before the period of interest (ie before March 2012). Therefore, the set of PECs that satisfy the condition that the PEC was reported for the 1st time in the Company DB and Social Media may be very small (or non-existent).

For those PECs that are POSITIVE CONTROLs and have been first reported in the company safety database and in social media in the period of interest (01-Mar-2012 to 31-Mar-2015), characterize the time difference (by ***min***, ***max***, ***mean***, ***median***, ***interquartile range***, and ***standard deviation***) between the PECs' INDEX DATEs and

- The date of the PECs' first case (from the company database)
- The date of the PECs' first social media post (for each indicator score)
- The calculations will be performed by UMC.
- **Each EFPIA company has to provide intermediate results to UMC in the following table format (values shown are just examples):**

Company ID	PEC ID	Data source	Data type	Difference between first occurrence and INDEX DATE [in months]
Company X	PEC001	Company	Report/post	-18
Company X	PEC001	Social04	Report/post	-12

Company ID	PEC ID	Data source	Data type	Difference between first occurrence and INDEX DATE [in months]
Company X	PEC001	Social05	Report/post	-6
Company X	PEC001	Social06	Report/post	2
…	…	…	…	…

Social media SDR analysis (applies to the TIRD <u>and</u> Harpaz data):

a) Sensitivity/Specificity:

- Extract the social media data for the PECs that are either POSITIVE or NEGATIVE CONTROLs (ignoring time of first occurrence).
  - Do this for each month of the study period
  - Do this for each indicator score cutoff (0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99)
  - Provide the index date, data source, and standard contingency table counts (a, ab, ac, abcd), see table below.
- Extract the spontaneous reporting data for the PECs that are either POSITIVE or NEGATIVE CONTROLs (ignoring time of first occurrence).
  - Do this for each month of the study period
  - For this exercise, use data from VigiBase
  - Provide the index date, data source, and standard contingency table counts (a, ab, ac, abcd), see table below.

<u>Notes:</u>

- EFPIA partners provide monthly raw contingency table counts for all PECs. Using these counts, UMC will be able to compute sensitivity and specificity for all signal detection methods of interest.
- The net result is a value for sensitivity and specificity for each signal detection method and cutoff. As there are currently 4 established SDR methods, and 7 social media datasets (corresponding to 7 different indicator score cutoffs), there will be a total of 28 values for sensitivity/specificity for the social media. For spontaneous data, there will be 4 values for sensitivity/specificity.

b) For characterizing time gain at the SDR level:

<u>Note:</u> Time gain is relative to the INDEX DATE. As an example, assume a POSITIVE CONTROL has an INDEX DATE of January 2014. If the first SDR from spontaneous reports occurs in January 2013, then that's -12 months relative to this date; if the first SDR from social media comes in March 2012, then that's -22 months; (same applies to single cases/posts, see above). This convention establishes a standard 'objective' reference point.

- Since all first occurrences of SDRs with the different signal detection methods will be possible to compute by UMC based on the monthly contingency table counts, EFPIA partners do not need to provide any additional data for this than that already provided for the purpose of sensitivity/specificity calculations
- Company-specific results (timing of first SDRs with different methods for the positive controls) can be obtained from UMC upon request

---

Each EFPIA company has to provide intermediate results to UMC in the following table format (values shown are just examples). This table will allow UMC to complete calculations for sensitivity/specificity AND time-gain:

Company ID	PEC ID[1]	Positive / Negative Control	Index date[2]	Data source[3]	Year-Month[4]	a[5]	ab[6]	ac[7]	abcd[8]
Company X	PEC001	Positive	201306	vigibase	201203	10	4,000	250	10,000,000
Company X	PEC001	Positive	201306	social04	201203	9	360	45	450,000
Company X	PEC001	Positive	201306	social05	201203	7	300	40	380,000
Company X	PEC001	Positive	201306	social06	201203	5	280	39	360,000
…	…	…	…	…	…	…	…	…	…

[1] Remember to keep a local lookup table so that a combination of your company name and a PEC ID can be back-translated to an actual product-event combination
[2] For positive controls, insert the index date in the format of YYYYMM; for negative controls, leave this blank
[3] Use the names 'social04', 'social05', 'social06', 'social07', 'social08', 'social09', and 'social099' for the thresholds 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.99, respectively
[4] You should include all months of the study period using the YYYYMM format, i.e. from 201203 to 201503
[5] The number of posts (or reports) on the PEC in the specified year-month, for the specified data source
[6] The number of posts (or reports) on the product in the specified year-month, for the specified data source
[7] The number of posts (or reports) on the event in the specified year-month, for the specified data source
[8] The total number of posts (or reports) in the specified year-month, for the specified data source

NOTES

1. The SDRs (and their dates of first occurrence) from spontaneous reports will be determined from VigiBase.
2. The receipt of a single spontaneous report (and its date) will be determined from the individual company spontaneous report databases.