WEB-RADR WP2b – STUDY SUMMARY

Analytics - Record linkage

Type of document:	Study summary
Version:	1.0
Date:	2014-11-28
Authors:	Tomas Bergvall, Johan Ellenius, Ingrid Johansson, Niklas Norén, Sara Gama, Jonas Fransson

Table of contents

1 Introduction

1.1 Purpose of the document

The purpose of this document is to serve as the Work Package 2b – Record Linkage's operational plan for how to organise and realise the deliverables of this WP as described in the Description of Work (DoW).

1.2 Version history

Version	Date	Comments	Authors
0.1	2014-11-13	First draft created	TB
0.2	2014-11-27	Comments incorporated from project members	TB,IJ,JE
0.3	2014-11-28	Comments incorporated from project members	TB,IJ,JE,SG,NN
1.0	2014-11-28	First official version of the protocol	TB,IJ,JE,SG,NN,JF

1.3 Definitions and abbreviations

For definitions and abbreviations in this document, please refer to the WEB-RADR Definitions and Abbreviations Document.

2 Rationale & Study Proposal

2.1 Background

Reporting by healthcare professionals (HCPs) and patients by means of mobile devices as well as the strengthening of safety signals from social media platforms to National Competent Authorities (NCAs) and marketing authorisation holders (MAHs), with subsequent transmission to regulatory authorities is a new and unexplored concept. In addition, the vast amount of information generated through social media requires a well-defined approach as regards monitoring, reporting, analysing and evaluating potential adverse reactions, signals and other medical insights related to medicines. The automatic identification of pharmacovigilance data such as medication names and adverse events in "free text", which is the format of tweets, blog posts etc. in social media, is a necessary first step in order to apply methods capable of identifying safety signals in such data. Another important step is to identify information concerning the same event, both to reduce duplication and retrieve additional information concerning the event of interest.

The record linkage problem is not a new concept and was described as early as 1946 by Dunn [1]. Dunn describes the problems facing the "registrar" trying to collect information to write the "book of life" for an individual. Dunn describes the problem of collecting information from disparate sources and attributing them to a single person much like the aim of this project. The theory of record linkage was developed by Fellegi and Sunter in 1969 [2]. The record linkage problem also goes by many names e.g. co-reference/entity/identity/name/record resolution, duplicate detection, conflation etc. We will however adhere to record linkage in the rest of this document for conformity. Reuter et.al [3] developed an algorithm that link images from Flickr related to the same event using record linkage techniques. The authors use the same features as Becker et al. [4] but show the importance of intelligent blocking of the pairwise comparisons to reduce the number of pairs considered.

In an evaluation across three different European countries in the IMI PROTECT project, led by the MHRA, vigiMatch as implemented by the Uppsala Monitoring Centre (UMC) was found to yield few false positives and at the same time identify duplicates that had gone undetected by current methods in use for the national data sets [5]. An adaptation to run vigiMatch directly on national

data is currently explored in a pilot study, with the aim of further improving the opportunity for effective duplicate detection in pharmacovigilance. vigiMatch can be adapted to identify duplicated or otherwise related content on social media.

2.2 Purpose

The overall aim of WP2b is to develop and link new and existing analytical tools for the analysis of social media content for pharmacovigilance purposes. WP2b therefore focuses on improving the quality and relevance of data extracted from social media sources and their subsequent use in the detection of safety signals. In order to achieve its objectives, WP2b will focus on three areas: 1) Suspected ADR identification; 2) Record linkage and 3) Signal detection. The objective of the present sub work package is "2) Record linkage". The remaining two objectives are accounted for elsewhere.

3 Study Objectives & Timelines

3.1 Study Objectives

The primary objective of this sub work package is to develop and evaluate record linkage algorithms to automatically detect duplicated posts and link users across social media platforms.

3.2 Milestones & Timescales

	Milestone	Estimated date of completion
M2B.4	Establish training data for record linkage in social media	M12
M2B.5	Implementation of vigiMatch for social media	M21
M2B.6	Evaluation of performance of vigiMatch for social media	M27

3.3 Deliverables

Deliverable No.	Deliverable description	Nature (R, P or O)	Expected delivery date
D2B.2	Technical report describing implementation and evaluation of record linkage in social media	R	M32

3.4 Responsibility and Coordination

This study is under the responsibility of:

- Study supervisor: Niklas Norén (UMC)
- Study coordinator: Tomas Bergvall (UMC)

3.5 Participants

The following will also be participating in the project:

Name	Organization	Mail
Tomas Bergvall (lead)	UMC	Tomas.Bergvall@who-umc.org
Sara Gama	Novartis	sara.gama@novartis.com
Letitia Jiri	Amgen	ljiri@amgen.com

Ingrid Johansson	UMC	Ingrid.Johansson@who-umc.org
Johan Ellenius	UMC	Johan.Ellenius@who-umc.org
Jonas Fransson	UMC	Jonas.Fransson@who-umc.org
Niklas Norén	UMC	Niklas.Noren@who-umc.org
Nabarun Dasgupta	Epidemico	nabarund@gmail.com
Carrie Pierce	Epidemico	carrie@epidemico.com

4 Methodology

This subpackage contains 5 tasks:

Task No.	Task Title / Description
T2B.2.1	Prepare social media for record linkage First focus on structured information available from the social media information then use the non-structured data. Drug names and adverse drug reaction will be extracted using existing methods from Epidemico and possibly refined by methods from the ADR identification subpackage. Lead: UMC (M6-18) Participants: UMC, Epidemico, and EFPIA partners.
T2B.2.2	Establish training and test data for record linkage Identify pairs of social media items that are known, or assumed, to refer to the same suspected case of an adverse reaction Lead: UMC (M6-12) Participants: UMC, Epidemico, and EFPIA partners.
T2B.2.3	Implement record linkage For execution against social media content processed according to the first step Lead: UMC (M6-21) Participants: UMC, and EFPIA partners.
T2B.2.4	Evaluate record linkage performance Against training data derived under 2B.2.2. Lead: UMC (M15-21)

	Participants: UMC, and EFPIA partners.

4.1 Prepare social media for record linkage

Data from Facebook posts, tweets and selected online patient forums/communities will be collected in WP2a. The data extraction process is handled by a tool developed by Epidemico. In order to start the development activities as soon as possible, a simple system to extract, display and store the required information will be developed.

In order to link records between different social media platforms we have recognized a need to also link users across social media platforms. There are sites with open access APIs like about.me that allow users to link their own social media accounts which will be exploited also for record linkage.

Data from WP2a and the suspected ADR identification subpackage will be utilized in this task to develop a dataset including parameters describing the content of the posts to be used in the record linkage algorithm. These parameters could include e.g. age, gender, timestamps, drug names and event names.

4.2 Establish training and test data for record linkage

In order to establish training and test data for record linkage manual inspection of the data will be performed with the aim to classify posts in the following categories: “unrelated”, “same user”, “duplicate post”, “description of same event” and “inconclusive”. Due to ethical considerations we will only be able to classify suspected instances of the above mentioned categories since we will not be able to go back to the source of the posts. This will pose a limitation to the study. The creation of gold standard data is a crucial step in order to get reliable method performance. A more detailed plan will be created based upon what the available data looks like. To extend the training dataset additional automatic classifications will be performed using information from about.me and the notion that users can automatically post their tweets to Facebook. One part of the training dataset will be withheld as a test set in order to get an independent estimation of the method performance.

4.3 Implement record linkage

The aim of this task is to implement record linkage algorithms in order to evaluate them in task 4.4. The vigiMatch method will be adjusted and used on social media data. A challenge is that social media is anticipated to consist of largely unstructured free text. Therefore, we plan to both apply vigiMatch to the original social media text without any pre-processing, and it will also be applied to certain identified data elements extracted from the text, such as drug names, ADRs, dates and location information. A thorough literature review will give further information about other state-of-the-art research in the area that might be useful. Special focus will be placed on feature extraction methods that can improve classification performance.

To start with we will focus on finding related posts within a social media platform and link users across social media platforms. If we see promising results we will extend the scope to find related posts also across social media platforms.

4.4 Evaluate record linkage performance

An analysis of the errors made by the methods on the test dataset will be performed in order to gain understanding of the relative strengths and weaknesses of the algorithms. The analysis will highlight areas for improvements of the algorithms. The method will also be evaluated against data not included either in the training or test datasets. The highlighted clusters will be manually reviewed to gain further insights into method performance.

5 Publication of study protocol and results

6 References

Ref No	Citation
1	Dunn HL. Record Linkage. Am J Public Health Nations Health. 1946;36(12):1412-6.
2	Fellegi IP, Sunter AB. A theory for record linkage. Journal of the American Statistical Association. 1969;64(328):1183-210.
3	Reuter T, Cimiano P, Drumond L, Buza K, Schmidt-Thieme L, editors. Scalable Event-Based Clustering of Social Media Via Record Linkage Techniques. ICWSM; 2011.
4	Becker H, Naaman M, Gravano L, editors. Learning similarity metrics for event identification in social media. Proceedings of the third ACM international conference on Web search and data mining; 2010: ACM.
5	Tregunno PM, Fink DB, Fernandez-Fernandez C, Lazaro-Bengoa E, Noren GN. Performance of probabilistic method to detect duplicate individual case safety reports. Drug Saf. 2014;37(4):249-58. doi:10.1007/s40264-014-0146-y.