JRC Scientific and Technical Reports

How to maximise event-based surveillance web-systems: the example of ECDC/JRC collaboration to improve the performance of MedISys

Authors:

Jas Mantero (European Centre for Disease Prevention and Control)
jas.mantero@ecdc.europa.eu

Jenya Belyaeva (Joint Research Centre of the European Commission)
jenya.belyaeva@ext.jrc.ec.europa.eu

Jens P. Linge (Joint Research Centre of the European Commission)
jens.linge@jrc.ec.europa.eu

EUR 24763 EN - 2011

JRC EUROPEAN COMMISSION

ipsc Institute for the Protection and Security of the Citizen

ecdc EUROPEAN CENTRE FOR DISEASE PREVENTION AND CONTROL

The mission of the JRC-IPSC is to provide research results and to support EU policy-makers in their effort towards global security and towards protection of European citizens from accidents, deliberate attacks, fraud and illegal actions against EU policies.

The European Centre for Disease Prevention and Control (ECDC) was established in 2005. It is an EU agency aimed at strengthening Europe's defences against infectious diseases. It is seated in Stockholm, Sweden. According to the Article 3 of the Founding Regulation, ECDC's mission is to identify, assess and communicate current and emerging threats to human health posed by infectious diseases. In order to achieve this mission, ECDC works in partnership with national health protection bodies across Europe to strengthen and develop continent-wide disease surveillance and early warning systems. By working with experts throughout Europe, ECDC pools Europe's health knowledge to develop authoritative scientific opinions about the risks posed by current and emerging infectious diseases.

European Commission
Joint Research Centre
Institute for the Protection and Security of the Citizen

Contact information
Address: Jens P. Linge, JRC IPSC TP267, Via E. Fermi 2749, I-21027 Ispra (VA)
E-mail: jens.linge@jrc.ec.europa.eu
Tel.: +39-0332-78-6446
Fax: +39-0332-78-5154

http://ipsc.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

JRC 63805

EUR 24763 EN
ISBN 978-92-79-19755-0
ISSN 1018-5593
doi:10.2788/69804

Luxembourg: Publications Office of the European Union

Printed in Luxembourg

Table of Contents

Summary

The mission of the European Centre for Diseases prevention and Control (ECDC) is to identify, assess and communicate current and emerging infectious threats to human health within the European Union (EU). The identification of threats is based on the collection and analysis of information from established communicable disease surveillance networks and from unstructured information mostly originating from non-health care sources, e.g. online news sites.

MedISys, an automatic real-time media monitoring and threat detection system developed by the Joint Research Centre (JRC) of the European Commission, is among the tools used by the ECDC for timely identification of potential public health threats from online information sources. In 2008, an ECDC internal analysis indicated that MedISys issued alerts faster than other human-mediated web-based systems. As timely detection is crucial to enable efficient response to public health threats, ECDC decided to further explore the potential of MedISys as an EU early warning tool. We analysed the functionality of the existing system in view of improving the usability of the web site, revising the sources and reducing the amount of irrelevant articles. We provided JRC with practical suggestions for the interface and asked public health experts at national level to assist in the revision of sources. To reduce the number of irrelevant articles, alternative search strategies for fifteen diseases were tested against the existing strategies using the positive predictive value (PPV) and the sensitivity to measure the performance of the system. Our intervention increased the PPV value (from 15.3% to 71.1%) and the sensitivity of the system.

We conclude that the best search strategies use a limited number of keywords weighted as positive (with weights adjusted below the alert thresholds) and an extended list of keywords weighted as negative. We recommend a high number of epidemiological terms within the keyword combinations.

The results indicate that user feedback is crucial to exploit the full potential of event-based surveillance systems such as MedISys. We will improve the detection of other infectious diseases and intend to cover all EU languages. Customized country versions will be set up in collaboration with JRC; ECDC will encourage the use of the system at national level in the EU member states.

Background

Epidemic Intelligence

Epidemic Intelligence (EI) encompasses activities related to the early identification of potential health threats, as well as their verification, assessment and investigation, in order to recommend adequate public health control measures [1]. To identify events that potentially represent a risk to human health, public health authorities should have a systematic approach for analysing data they routinely collect (indicator-based surveillance, IBS) and monitoring information about potential public health events (event-based surveillance, EBS). This comprises official reports from health authorities not included in traditional surveillance systems and unofficial reports (e.g. media reports and any informal communication).

The internet allows informal information-sharing and easy access to a large amount of information useful for EBS [2, 3]. The revision of the International Health Regulations (IHR, 2005) of the World Health Organisation [4, 5] acknowledged the need to include informal reports to the range of information sources used to detect new public health events of international concern. This involves the use of unstructured information and requires a revision of the procedures for health surveillance at national and international level. A better understanding of available technologies will simplify the use of the internet for epidemic detection purposes [6].

European Centre for Disease Prevention and Control (ECDC)

The European Centre for Disease Prevention and Control (ECDC), established by the European Parliament and Council Regulation 851/2004 of 21 April 2004, became operational on 20 May 2005. According to its founding regulation, ECDC's mission is to identify, assess and communicate current and emerging threats to human health posed by infectious diseases [7].

Most EU member states have long-established disease surveillance systems for communicable diseases (IBS), but EBS activities are often not fully standardised across the countries. The ECDC works in partnership with national health protection bodies to strengthen and develop EU-wide disease surveillance.

Event-based surveillance systems

ECDC uses several web-based technologies for early detection of communicable disease threats and monitoring of related web-information (early warning systems). Most of these systems have been developed recently [8, 9]. *Automated systems* such as MedISys [10, 11], HealthMap [12, 13] and Biocaster [14] collect web articles of potential public health interest in real-time with none or very limited human intervention; reports are categorised and

displayed in maps and statistical graphs. These systems rely mostly on information extraction methods recognising specific terms or combinations thereof in the text.

In *moderated systems,* human moderation is needed (to a varying degree) to filter and classify web-information, i.e. trained analysts evaluate the content of reports that are often pre-filtered automatically. GPHIN [15, 16], ProMed [17] and Argus [18] are examples of well-established moderated systems.

Timely retrieval of information about potential public health hazards is crucial to accelerate the implementation of response measures. Automated event-based surveillance systems such as MedISys therefore play an important role. However, these tools present some clear limitations due to the partial or total absence of human analysis in the selection process. Our experience showed that information is often redundant (duplication of relevant items) and that many irrelevant articles are selected (false positive component). In contrast, human-moderated systems offer information of high quality in terms of content, but may present selection bias (i.e. miss relevant information). Human moderation relies on human analysis based on, amongst others, geographical, linguistic and epidemiological criteria to define the nature of a potential public health event. Moreover, the intervention of analysts leads to a time delay in the communication of detected events and additional costs.

Considering the above-mentioned limitations, we explore the benefits of automated early warning systems with free access for the public and the possible role of users to improve performance.

MedISys

MedISys is an event-based surveillance system developed by the European Commission Directorate General for Health and Consumers (DG SANCO) and the Joint Research Centre (JRC) of the European Commission. The system monitors web-based information (media articles and open-source public health reports) about human, animal and plant infectious diseases, chemical, biological, radiological and nuclear (CBRN) threats, and food & feed contaminations. It features a publicly available web site [19] and a restricted access version for public health officials that also covers nuclear and chemical hazards; public health professionals get free access to the restricted version upon request.

MedISys retrieves approximately 100000 news articles from more than 2500 selected media sources per day [20]. In addition, the system screens several hundred specialist and government web sites and twenty commercial news providers. The system evaluates and classifies the information displaying the results in a web interface covering more than 200 public health categories. The interface has been developed to meet the requirements of the main users (DG SANCO, ECDC, EFSA, EU member states). Articles are categorized according to hazard and/or country; maps with alert levels are presented together with statistics for the last month. Articles can be further filtered by

language, news source, and country. Additionally, the system clusters articles about the same topic, identifies duplication of news items and extracts from the articles a description of the events via the Pattern-based Understanding and Learning System (PULS) developed by the University of Helsinki [21].

MedISys evaluates all incoming news items through a **triggering methodology** based on predefined multilingual search terms for each disease and public health topic included in the system (ALERT DEFINITIONS) using:

a) List of weighted terms; and/or

b) Combination of terms.

a) List of weighted terms

This mechanism is based on a list of words with positive and/or negative weights triggering previously established alert acceptance thresholds.

Table 1: MedISys triggering algorithm:, 1st mechanism (LIST OF WEIGHTED TERMS)

Keyword	Weight	
dengue	+15 points	**Alert threshold fixed at 30 points**
DHF	+15 points	
concert	-999 points	

Table 1 reports an example of how a "*list of weighted terms"* for dengue could look like. In this case, an article would be selected by the system and displayed in the disease category for dengue if the term "dengue" appears at least twice in the text in order to reach the alert threshold (15x2=30 points). However, if the text includes the term "concert", the article will not be selected (15x2-999= -969 points, thus below the alert threshold). Negative weighted keywords such as "concert" in this example help avoiding the selection of false positives items. In the above example, an article about a concert by a music band called "dengue" would not be selected by the system and would not be considered as relevant for public health.

b) Combination of terms

The second triggering mechanism used by MedISys alerts is represented by combinations of search words using Boolean expressions ("AND", "AND NOT" rules). Articles are triggered by this mechanism if at least two combined keywords are included in the same text ("AND" combination of terms), regardless of how many times the single

terms appear. Negative words, excluding the selection of items ("AND NOT" combinations of terms) can be considered, as in the below example (Table 2).

Table 2: MedISys triggering algorithm, 2nd mechanism (COMBINATION OF TERMS)

Any of the following terms	AND any of the following (AND Combination)	AND NONE of the following (NOT Combination)
dengue	outbreak	concert

In this simple example for a hypothetical "c*ombination of terms"* for dengue, an article would be selected if both the terms "dengue" and "outbreak" were included in the text and if the term "concert" was not included in the same text. The rationale is that the presence of the two terms "dengue" and "outbreak" in the same article is considered a sign of public health relevance, regardless of how many times these words appear in the text. However, the presence of the term "concert" would be interpreted as the text being most probably not about the disease, but a rock band named "dengue". There are currently more than two hundred multilingual alerts definitions in the systems that were created following these procedures. More details about the extraction mechanism are described elsewhere [22].

MedISys was created to support public health professionals in the screening of news media and sites of public health authorities. This activity includes both the early detection of (re)emerging health threats and the follow-up of known public health events. A statistical alert system supports users in the detection of unusual changes in the reporting of diseases and symptoms. The statistical alerts are based on the comparison of the number of articles displayed for each disease and for each country in the latest 24 hours with a two-week period average. The statistical information is represented in graphs and generates automated email alerts to subscribed users.

An ECDC internal analysis included in the 2009 ECDC Annual Epidemiological Report for Communicable Diseases in Europe [23] and referring to the epidemic intelligence activities done during 2008 indicated that MedISys was rarely the first source of information for detection of new potential threats. A retrospective consultation showed that the system timely detected information about the threats followed during the same year and often issued statistical alerts faster than human-moderated web-applications.

Objective

The objective of the study was to assess the functionality and added value of MedISys as an early detection tool for public health threats at EU level. We aimed to identify areas of intervention from a user perspective that could improve the system, initiating a systematic collaboration with the developers (JRC), in order to increase the efficacy of the system.

Methods

ECDC descriptive study on MedISys

In early 2009 we performed a descriptive study on MedISys to identify potential areas of intervention. The main areas identified were the modification/customization of the user interface, the revision of the sources screened at EU level and the functionality of the automatic article selection. We targeted the automatic article selection as the priority area for intervention. We named the existing set of alert definitions **MedISys ZERO** and analysed how these were set up, acknowledging a general lack of standardized procedures. Table 3 describes the main characteristics of MedISys ZERO while Table 4 shows as practical example the existing alert definition for dengue (MedISys ZERO-Dengue)

Table 3: Main aspects of the alert definitions in MedISys ZERO

	Weighted keywords		Combination of terms
MEDISYS ZERO	*Selection of keywords*	Lack of standardized procedures: the selection does not follow defined rules. The positive weighted keywords include generic terms describing the disease and/or the pathogen.	Limited and not standardized use.
	Positive weighted keywords	The values of the positive weights are the same as the threshold alert value.	
	Negative weighted keywords	Limited and not standardized use	

Table 4: MedISys ZERO - keywords used for the dengue alert definition

Keyword	Weight (threshold fixed at 30 points)
dengue	+30 points
DHF	+30 points

ECDC evaluation of the existing alert definitions

The sample used to evaluate the performance of the existing alerts was the set of articles in English screened by the system during a 24 hour period (between 12am of 22 March 2009 and 12am of 23 March 2009). The reference for the analysis was the human evaluation of the content of the same sample of articles in terms of public health relevance. This evaluation was performed by an ECDC EI expert. The process was restricted to articles selected for 15 diseases considered of particular interest at the EU level. The list included anthrax, influenza, avian influenza (H5N1), tuberculosis, measles, dengue fever, chikungunya, ebola, Crimean Congo hemorrhagic fever, rabies, botulism, tularemia, cholera, poliomyelitis and plague. We established that the EI expert evaluating the sample of articles considered the public health relevance for the 15 selected diseases without any limitation in terms of geographical location of events. Following this definition, a total of twenty articles were considered relevant by the human filter in the observed period. We evaluated sensitivity, specificity and positive predictive value of MedISys ZERO (the existing system) according to the following definitions [24, 25, 26]:

- **Sensitivity:** proportion of articles considered relevant by the human filter correctly detected by the system;
- **Specificity:** proportion of articles considered irrelevant by the human filter correctly disregarded by the system;
- **Positive Predictive Value (PPV):** proportion of articles detected by the system that were evaluated as relevant by the human filter.

The PPV entails the quantity of "false positives". It can be expressed as the probability that an article selected and displayed by the system is indeed relevant in terms of public health.

As the PPV is strictly linked to the prevalence, we retrospectively excluded that major outbreaks of the included diseases were occurring during the testing period. The specificity resulted to be already high in the existing system, meaning that almost all the disregarded articles were irrelevant for the human filter. A high level of irrelevant articles was however identified by the human filter among the articles selected by the system.

ECDC proposal of new alert definition

In order to reduce the number of false positives and to improve the sensitivity of the existing system we tested four alternative standardized approaches to set up alert definitions. The process resulted in four different sets of alert definitions for each of the 15 diseases. The proposed alerts included new "combinations of terms" and "lists of weighted terms". The sets were named MedISys I, MedISys II, MedISys III and MedISys IV.

The suggested strategies considered English only and were not exhaustive in view of all the possible alerts that could be created. The rationale about using these four (and not other) alerts was the need of defining basic standardized procedures for alert definitions based on a logical approach against a virtually unlimited inclusion of terms. The main characteristics of the new four alerts are summarized in Table 5.

Table 5: **Main aspects of the definitions of the new sets of alerts (MedISys I, II, III and IV)**

	Weighted keywords		Combination of terms
MEDISYS I	**Selection of keywords**	Limited number of terms (1-2) selected among terms commonly used by media/medical literature to define the disease. No synonyms and no symptoms are included.	Extended use of combination of terms: terms naming the disease combined with epidemiological common terms (e.g. "epidemic", "outbreak"). Not-terms are the same as the negative weighted keywords.
	Positive weighted	The values of the positive weights are lower than the threshold alert (10 points against 30 points)	
	Negative weighted	Extended number (4-5) of negative weighted terms. Three terms are common for all the alerts, the others are disease specific	

	Weighted keywords		Combination of terms
MEDISYS II	**Selection of keywords**	Extended use of terms (5-6 terms) referring to the disease and identified through a semantic analysis (synonyms and non-medical expressions are included e.g. non-scientific terms referring to the disease)	As for MedISys I
	Positive weighted	The values of the positive weights are the same as the threshold alert (30 points against 30 points)	
	Negative weighted	As for MedISys ZERO: limited and not standardized number of terms with negative value.	

	Weighted keywords		Combination of terms
MEDISYS III	**Selection of keywords**	No use of single keywords	As for MedISys I
	Positive weighted	-	
	Negative weighted	-	

	Weighted keywords		Combination of terms
MEDISYS IV	**Selection of keywords**	As for MedISys II	As for MedISys I
	Positive weighted	As for MedISys I	
	Negative weighted	As for MedISys I	

We evaluated the performance of the new sets of alerts using the reference obtained by the human filtering.

MedISys I uses one or two common terms referring to the disease or to the agent as weighted positive keywords. The terms are weighted positively with a lower score than the alert threshold value; hence, a term has to appear more than once in a text to trigger an article.

In the example of the dengue alert (see Table 6), a threshold of 30 points is used and the single positive terms each have a weight of 10 points. Thus, positive terms (the same one or different ones) have to appear at least three times in the text in order to select an article. "Dengue" is the name of a rock band while DHF, the acronym for dengue hemorrhagic fever, is not only the acronym for dengue hemorrhagic fever but also the acronym used by the Danish Handball Federation. For these reasons a negative value was assigned to terms as "handball" and "concert" (see table 7): if these words appear in a text including the term "dengue" the article will not be selected.

Table 6: MedISys I - single keywords used for the "dengue alert definition"

Keyword	Weight (threshold alert established as 30 points)
dengue	+10 points
DHF	+10 points
concert	-999 points
handball	-999 points

The alert definitions of MedISys I incorporate, in addition to single weighted keywords, combinations of terms referring to the diseases with an extended epidemiological terminology, e.g. "outbreak", "epidemic" (see table 7). This means that if the term "dengue" appears in the same text of the term "outbreak" the system selects the article, regardless of the number of times that "dengue" appears.

Table 7: MedISys I - combination of terms in "dengue alert definition"

ANY OF	AND ANY OF	BUT NONE OF
dengue	epidemic	concert
DHF	outbreak	handball

MedISys II considers a new extended set of positive weighted terms. The proposed terms include non-medical expressions commonly used in the media to describe disease (e.g. "dandy fever" for dengue or bird flu for H5N1).
The positive weighted keywords present the same value as the alert threshold (see table 8). This means that, as for MedISys ZERO, articles are selected each time the terms appear in the text. Negative weighted terms are not used for this alert; the new set of keywords is also used for the combinations of terms, where negative words are again not included (see table 9).

Table 8: MedISys II - single keywords used in *dengue alert definition*

Single weighted keyword	Weight (threshold alert established as 30 points)
dengue	+30 points
DHF	+30 points
dandy fever	+30 points
break bone fever	+30 points

Table 9: MedISys II - combination of terms in *dengue alert definition*

ANY OF	AND ANY OF	BUT NONE OF
dengue	cluster	
DHF	epidemic	
dandy fever		
break bone fever		

The keywords used for the alerts in **MedISys III** are the same as in MedISys II, but they are used only in the combination of terms together with epidemiological terms (no single keywords list). The alert definition for "dengue" in MedISys III follows in Table 10.

Table 10: MedISys III combination of terms in *dengue alert definition*

ANY OF	AND ANY OF	BUT NONE OF
dengue	cluster	concert
DHF	epidemic	handball
dandy fever		
break bone fever		

MedISys IV is a combination of the previous strategies. The terms are the same as for MedISys II and MedISys III but the positive value of the weighted keywords is reduced, as for MedISys I. In addition, an extended use of negative weighted keywords is included. No combination of terms is used this time (see Table 11).

Table 11: MedISys IV - single keywords used in *dengue alert definition*

Single weighted keyword	Weight (threshold alert established as 30 points)
dengue	+30 points
DHF	+30 points
dandy fever	+30 points
break bone fever	+30 points
concert	-999 points
handball	-999 points

Sensitivity, specificity and PPV were calculated for the four sets of alerts. We considered as the best performing set of alerts the one with the highest PPV value able to increase the initial value of the sensitivity (75%).

Once the best set had been identified, we further refined the fifteen alert definitions using an extended set of additional negative weighed keywords. These were based on the experience gained during the testing periodand on the false positive component of articles identified during this period. Six months after the final revision we recalculated sensitivity, specificity and PPV of the alerts **(MEDISYS V - FINAL VERSION)**, employing the same evaluation methodology involving a human EI expert as reference.

Results

During the testing period, MedISys ZERO scanned 13662 articles in English, identifying 98 relevant articles for the fifteen diseases considered in the study. The human filter assessed the same sample, considering 19 articles as relevant for the selected diseases. The EI expert considered 83 of the articles identified by MedISys ZERO as irrelevant (a false positives component of 84.7%). Four articles evaluated as relevant by the human filter were not triggered by the system (false negative component). The results are summarized below in Table 12.

Table 12: Articles selected by MEDISYS ZERO versus human filter evaluation

	Human filter (+)	**Human filter (-)**	**N**
MedISys ZERO (+)	15	83	98
MedISys ZERO (-)	4	13,560	13,564
Total (n)	19	13,643	13,662

The sensitivity and specificity of the existing system (MedISys ZERO) were 79% and 99.4%, respectively; the PPV was 15.3%. The sensitivity, specificity and PPV of the four new search strategies were calculated and compared with the existing set (Table 13).

Table 13: Comparison of MedISys ZERO, I, II, II and IV in terms of sensitivity, specificity, presence of false positives and positive predictive value (PPV)

	MedISys ZERO	MedISys I	MedISys II	MedISys III	MedISys IV
Sensitivity (%)	79.0	**84.2**	84.2	90.0	90.0
Specificity (%)	99.4	**99.6**	99.2	98.8	99.2
False positives (%)	84.7	**75.4**	85.8	89.8	83.5
PPV (%)	15.3	**24.6**	14.2	10.2	16.5

MedISys I was able to increase the PPV from 15.3% to 24.6%. This strategy was able to maintain high sensitivity (84.2% versus 75% of the initial system) and specificity (99.6% versus 99.4%) reducing the false positive component at the same time.

MedISys II performed closely to MedISys I in terms of sensitivity and specificity, but lead to a low PPV, even lower than the original system (14.2%).

MedISys III and IV were able to increase the sensitivity to 90%, but the false positive component significantly increased, with a PPV of 10.2% and 16.5%, respectively.

MedISys I presented the highest PPV and specificity; the sensitivity was higher in MedISys III and IV but these two strategies did not perform well as MedISys I in terms of PPV. In addition an analysis of the results according to disease revealed that the small difference identified in sensitivity was related to one specific alert ("influenza").

Based on these considerations, the alert definition strategy used for MedISys I was considered the best. The fifteen alerts included in MedISys I were thus further developed, refined and tested by ECDC experts. In particular, negative keywords were added (see tables 14 and 15 for Dengue alert). After revision, the set of alerts was named "MedISys V – FINAL REVISION" (Table 16).

Table 14: MedISys FINAL REVISION - single keywords used for "dengue alert definition"

Keyword	Weight (threshold alert established as 30 points)
dengue	+10 points
DHF	+10 points
band	-999 points
music	-999 points
movie	-999 points
book	-999 points
handball	-999 points
football	-999 points
concert	-999 points

Table 15: MedISys FINAL REVISION - combination of terms for "dengue alert definition"

ANY OF	AND ANY OF	BUT NONE OF
dengue	epidemic	band
DHF	outbreak	music
	cluster	movie
	death	book
		handball
		football
		concert

Table 16 summarises the results of the analysis of MedISys V in terms of sensitivity, specificity and PPV (the same methodology used to evaluate the previous versions was used).

Table 16: Articles selected by MEDISYS V-FINAL REVISION versus human evaluation

	Human filter (+)	Human filter (-)	N
MedISys (+)	97	28	125
MedISys (-)	5	13069	13074
Total (n)	102	13097	13199

The sensitivity of the revised set of alerts resulted to be 95%; the PPV was 77.6%, meaning that for every ten articles selected as relevant and displayed on the system almost eight resulted to be relevant by the human filter.

The false positive component dropped from 84.7% of the initial version to 22.4%. Table 17 summarises how the triggering methodology of MedISys V - FINAL VERSION compared to the human filter.

Table 17: MEDISYS V- FINAL REVISION (sensitivity, specificity and PPV)

	MedISys V Final Version
Sensitivity (%)	95.0
Specificity (%)	99.8
False positives (%)	22.4
PPV (%)	77.6

The basic strategy for MedISys V includes a very limited use of terms referring to the disease (1 to 2 terms) with a positive weight below the threshold value. Negatively weighted terms are used extensively.

Discussion

The descriptive analysis of MedISys by ECDC identified as main limitation the considerable false positive rate of articles (i.e. many items with limited or no public health relevance were displayed).

Our evaluation of the triggering methodology showed that the alert definitions had been set up without standardized linguistic and/or epidemiological procedures. In the tested subset of the existing system, only fifteen out of hundred selected articles resulted to be relevant for the human filter. The two automatic mechanisms resulted to be based on the inclusion of generic terms describing the diseases with no identified procedures in terms of choice of keywords and their respective weights. Moreover, we noticed that few epidemiological terms and negatively weighted keywords had been used in the initial set-up.

Our alternative standardised alert definitions were the result of our attempt to reduce the "false positive component" without affecting the high level of sensitivity. Balancing high sensitivity and low false positive rate, we identified a basic alert definition strategy that was further elaborated in the following months. The results of this complex revision showed that user intervention can dramatically improve the ability of the system in triggering articles relevant for public health purposes.

Our study illustrates how structured collaboration between users and developers can greatly improve the functionality of automatic early alerting tools. This is especially important in the field of text mining and threat detection for public health because of the clear need to establish standard alert definitions.

Conclusion

ECDC was able to improve the capacity of MedISys in the detection of relevant public health information by focusing on the revision of the search patterns in a systematic approach. A comprehensive review of the included news sources was completed with the support of users at EU national level. We classified the sources as unofficial (media) and official (public health authorities) in order to facilitate the work of EI officers in terms of validation. Furthermore, we identified the need of user capacity building (training, active participation and collaboration); this will be reinforced at EU level through electronic tutorials created in collaboration with DG SANCO and JRC.

We will extend our approach to other diseases included in the system and considered relevant for EI purposes at EU level (e.g. Q fever). Translations for further languages will be included for all alerts with the support of users at EU national level e.g. public health professionals involved in EI activities. It is unlikely that the system will present the same precision for the new languages from the beginning, as every language needs a slightly different approach. However, corrections and adjustments will be done in a dynamic and standardized way. The revision process will be continuous and based on a defined collaboration between users at EU member state and ECDC level and the developers at JRC. ECDC will facilitate this process.

The next steps in the ECDC/JRC and DG SANCO collaboration for strengthening MedISys will lead to customized versions of the tool for national needs, i.e. MedISys Country Editions. These national versions will maintain all the system functions with the option of pre-filtering national sources per country and presenting information through translated and customized user interfaces. A close collaboration between the developers and EI experts at national level will be facilitated by ECDC coordination of the project.

We strongly suggest that public health experts involved in daily epidemic intelligence activities at national and international level consider formalizing their collaborations with developers of event-based surveillance systems to monitor the performance of the systems. According to our experience, users should report problems and limitations indentified during daily EI activities to developers in a standardized manner, as this enabled JRC developers to solve most of the problems identified in our analysis. In the authors' opinion, a close collaboration between users and developers will be key for the further development of event-based surveillance systems.

Acknowledgments

We thank Mireia Cid (kalispera.net) for her work on translations and sources.

References

1. Paquet C, Coulombier D, Kaiser R, Ciotti M. Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. Euro Surveill. 2006;11(12):pii=665. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=665

2. Grein TW, Kande-Bure O. Kamara, Rodier G et al " Rumors of Disease in the Global Village: Outbreak Verification Emerging Infectious Disease Vol.:97-102 March-April 2000

3. Woodall "Official versus unofficial outbreak reporting through the Internet". Intern Journal of Medical Informatics 47 (1997), 31-34

4. World Health Organization .International Health Regulations (2005), Articles 5-9 http://whqlibdoc.who.int/publications/2008/9789241580410_eng.pdf

5. Baker MG, Forsyth AM. The new International Health Regulations: a revolutionary change in global health security. NZ Med.J 2007 Dec 14:120(1267):U2872 N Z Med J. 2007 Dec 14;120(1267):U2872.

6. Weekly Epidemiological Records WEEKLY EPIDEMIOLOGICAL RECORD, NO. 1, 7 JANUARY 2000 WHO.s Weekly Epidemiological Record, Vol. 75, No. 1 (2000). an integrated approach to communicable diseases surveillance

7. Regulation (EC) No 851/2004 of the European Parliament and of the Council of 21 April 2004 establishing the European Centre for Disease Prevention and Control. Available from: http://www.ecdc.europa.eu/About_us/Key_Documents/ecdc_regulations.pdf

8. Linge JP, Steinberger R, Weber TP, Yangarber R, van der Goot E, Al Khudhairy DH, Stilianakis NI. Internet surveillance systems for early alerting of health threats. Euro Surveill. 2009;14(13):pii=19162. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19162

9. Hartley D, Nelson N, Walters R, Arthur R, Yangarber R, Madoff L, Linge J, Mawudeku A, Collier N, Brownstein J, Thinus G & Lightfoot N. The Landscape of International Event-based Biosurveillance. Emerging Health Threats Journal 2010, 3. doi: 10.3134/ehtj.10.003

10. Steinberger R, Fuart F, Pouliquen B & van der Goot E. MedISys: A Multilingual Media Monitoring Tool for Medical Intelligence and Early Warning. In: Proceedings of the International Disaster and Risk Conference (IDRC2008), pp. 612-614, Davos, Switzerland

11. Europe Media Monitor (EMM), information on MedISys available online from http://emm.newsbrief.eu/overview.html#MedISys

12. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J Am Med Inform Assoc. 2008 Mar-Apr;15(2):150-7

13. Brownstein JS, Freifeld CC, Reis BY, Mandl KD (2008) Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. PLoS Med 5(7): e151. doi:10.1371/journal.pmed.0050151

14. Collier, N. Doan, S., Kawazoe, A., Matsuda Goodwin, R., Conway, M., Tateno, Y., Ngo, Q., Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M. and Taniguchi, K. (2008),"BioCaster: detecting public health rumours with a Web-based text mining system", Bioinformatics, 24(24):2940-2941, Oxford University Press, DOI: 10.1093/bioinformatics/btn534.

15. Keller M, Blench M, Tolentino H, Freifeld CC, Mandl KD, Mawudeku A, et al. Use of unstructured event-based reports for global infectious disease surveillance. Emerg Infect Dis. 2009 May

16. Mykhalovskiy E, Weir L. The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. Can J Public Health. 2006 Jan-Feb;97(1):42-4.

17. Madoff CL, . ProMed mail: an early warning system for emerging diseases; clinical Inf dis 2004:39:227-32

18. ARGUS, http://biodefense.georgetown.edu/projects/argus.aspx

19. MedISys, public version available online at: http://medisys.newsbrief.eu/

20. Steinberger R, Pouliquen B & van der Goot E(2009). An Introduction to the Europe Media Monitor Family of Applications. In: Fredric Gey, Noriko Kando & Jussi Karlgren (eds.): Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009), pp. 1-8. Boston, USA. 23 July 2009.

21. PULS project, University of Helsinki. Information available at: http://puls.cs.helsinki.fi/medical/

22. Steinberg R, Fuart F, van der Goot E, Best C, von Etter P & Yangarber R. "Text mining from the web for Medical Intelligence". From Fogelman-Soulie F et al "Mining Massive Data Sets for Security". IOS Press 2008, available online at http://langtech.jrc.it/Documents/2008_MMDSS_Medical-Intelligence.pdf

23. ECDC Annual epidemiological report on communicable diseases in Europe 2009 (revised edition): analysis of threats monitored 2005-2009. Available online from: http://www.ecdc.europa.eu/en/publications/Publications/0910_SUR_Annual_Epidemiological_Report_on_Communicable_Diseases_in_Europe.pdf

24. Updated Guidelines for Evaluating Public Health Surveillance Systems – Recommendations from the guidelines Working Group, July 27, 2001 / 50 (RR13):1-35. Available online at http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5013a1.htm

25. Framework and Tools for Evaluating Health Surveillance, Health Surveillance Coordinating Committee (HSCC), Population and Public Health Branch Health Canada. March 2004 – Available online at http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5013a1.htm

26. German R, "Sensitivity and Predictive Value Positive Measurements for Public Health Surveillance Systems". Epidemiology Nov 2000, Vol.11 N.6

European Commission

EUR 24763 EN – Joint Research Centre – Institute for the Protection and Security of the Citizen
Title: How to maximise event-based surveillance web-systems: the example of ECDC/JRC collaboration to improve the performance of MedISys
Author(s): Mantero J. & Belyaeva J.
Luxembourg: Publications Office of the European Union
2011– 22 pp. – 21 x 29.7 cm
EUR – Scientific and Technical Research series – ISSN 1018-5593
ISBN 978-92-79-19755-0
doi:10.2788/69804

Abstract
The mission of the European Centre for Diseases prevention and Control (ECDC) is to identify, assess and communicate current and emerging infectious threats to human health within the European Union (EU). The identification of threats is based on the collection and analysis of information from established communicable disease surveillance networks and from unstructured information mostly originating from non-health care sources, e.g. online news sites.
MedISys, an automatic real-time media monitoring and threat detection system developed by the Joint Research Centre (JRC) of the European Commission, is among the tools used by the ECDC for timely identification of potential public health threats from online information sources. In 2008, an ECDC internal analysis indicated that MedISys issued alerts faster than other human-mediated web-based systems. As timely detection is crucial to enable efficient response to public health threats, ECDC decided to further explore the potential of MedISys as an EU early warning tool. We analysed the functionality of the existing system in view of improving the usability of the web site, revising the sources and reducing the amount of irrelevant articles. We provided JRC with practical suggestions for the interface and asked public health experts at national level to assist in the revision of sources. To reduce the number of irrelevant articles, alternative search strategies for fifteen diseases were tested against the existing strategies using the positive predictive value (PPV) and the sensitivity to measure the performance of the system. Our intervention increased the PPV value (from 15.3% to 71.1%) and the sensitivity of the system.
We conclude that the best search strategies use a limited number of keywords weighted as positive (with weights adjusted below the alert thresholds) and an extended list of keywords weighted as negative. We recommend a high number of epidemiological terms within the keyword combinations.
The results indicate that user feedback is crucial to exploit the full potential of event-based surveillance systems such as MedISys. We will improve the detection of other infectious diseases and intend to cover all EU languages. Customized country versions will be set up in collaboration with JRC; ECDC will encourage the use of the system at national level in the EU member states.

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

JRC
EUROPEAN COMMISSION

LB-NA-24763-EN-C

ISBN 978-92-79-19755-0

Publications Office