SCOPE Work Package 4
ADR Collection

Duplicate Detection:
Best Practice Guide

SCOPE

SCOPE Work Package 4
ADR Collection
Duplicate Detection: Best Practice Guide

Contents

Acknowledgments

Authors

This document was prepared by the members of SCOPE team of the Agency for Medicinal Products and Medical Devices of Croatia (HALMED).

Acknowledgements

The authors gratefully acknowledge the Consumers, Health and Food Executive Agency (CHAFEA) of the European Commission for funding the SCOPE Joint Action and the MS partners for their contribution.

We would like to kindly thank our colleagues across EU MSs for completing the Patient Reporting survey questionnaire of the SCOPE Joint Action WP4 ADR Collection.

We thank the authors of all published papers on patient ADR reporting presented or cited in this white paper for their contribution to the knowledge of patient reporting across EU.

It has been our privilege to collaborate with our colleagues from Agenzia Italiana del Farmaco (AIFA), Lægemiddelstyrelsen (DKMA), Netherlands Pharmacovigilance Centre Lareb and the Medicines and Healthcare Products Regulatory Agency (MHRA), who have actively contributed to the development of this document.

1. Introduction

1.1 Background

The Strengthening Collaboration for Operating Pharmacovigilance in Europe (SCOPE) Joint Action has been created to support operations of pharmacovigilance in Europe following the requirements introduced by the 2010 European pharmacovigilance legislation[1, 2, 3]. Information and expertise on how regulators in EU Member States (MSs) run their national pharmacovigilance systems was gained in order to develop and deliver guidance and training in key aspects of pharmacovigilance, with tools, templates and recommendations. The aim of the SCOPE Joint Action was to support the development of a consistent approach across the EU network for all pharmacovigilance operations, in order to benefit medicines safety monitoring and communications to safeguard public health.

SCOPE was divided into eight separate work packages, with five work packages focusing on pharmacovigilance topics to deliver specific and measureable objectives, ranging from improvements in Adverse Drug Reaction (ADR) reporting to the assessment of quality management systems.

Work Package 4 ADR Collection was focused on national schemes for the spontaneous reporting of ADRs and was aimed to provide National Competent Authorities (NCAs) with a full understanding of good practices within national systems for collecting ADRs. Information was gathered from MSs' institutions [a] to understand their national ADR system, pharmacovigilance IT system capabilities, as well as implementation of patient reporting, types of reporting forms developed, and electronic reporting developments, including those from clinical healthcare systems. This information was used to create best practice guidelines, performance indicators and a media toolkit for raising awareness of ADR reporting systems, which will be supported through delivery of a training course for institutions.

Within WP4 there were five topics. Within the topic Audit of national reporting systems, the information about MSs' practices in ADR reporting was collected through a questionnaire completed by MSs. The questionnaire focused on national ADR reporting systems for medicinal products, vaccines and biologics, covering specific issues, such as additional monitoring. In depth analysis of the retrieved data, as well as subsequent follow-up with relevant stakeholders within the SCOPE project, revealed certain points of interest within the community, bringing out the main problems as well as examples of good practice.

[a] The term Member State's institution refers to the institution responsible for Adverse Drug Reaction (ADR) reporting, collection, processing and analysis within the particular Member State. Therefore, wherever the term 'Institution' is mentioned it does not necessarily refer to the National Competent Authority (NCA), although it will be synonymous in the majority of MSs.

One of the issues important for spontaneous reporting is report duplication. Duplication of cases is an important data quality problem and can pose significant problems for detecting and analysing signals arising from pharmacovigilance databases by misleading clinical assessment or distorting statistical screening, both artificially inflating and masking signals of disproportionate reporting.

The results of the SCOPE questionnaire indicate that over 92% (25/27[b]) of MSs perform duplicate detection; the two MSs that do not perform duplicate detection use Eudravigilance (EV) as their national database, so duplicate detection with regards to their national ADRs is performed in EV. SCOPE results also revealed differences among MSs with regard to the duplicate detection process, ranging from MSs who have more than one duplicate detection system in place (e.g. a computer algorithm and SOPs/written procedures) to MSs without any duplicate detection system in place.

Based on these responses, it was decided to follow-up with the MSs that were considered to have good practice with regard to the duplicate detection process. This mainly refers to MSs with well-defined duplicate detection methods and a good description of the system. In addition, Uppsala Monitoring Centre was consulted for their input on the management of duplicates due to their extensive experience and scientific expertise regarding different aspects of ADR reporting, including the duplicate detection process. During the follow-up, detailed information on the duplicate detection process was gathered, describing key points of the process: identification of potential duplicates, confirmation of duplicates and management of duplicate cases.

It has to be emphasised that this document is not intended as a duplication of already existing applicable EU guidance on the duplicate detection process. The aim of the document is to:

- Provide an overview of current MSs' experiences related to duplicate management, indicating examples of good practice
- Promote the importance of duplicate detection and emphasize key points of the process
- Summarise the applicable EU guidance documents related to the duplicate detection process
- Serve as a starting point for the development of a learning tool on duplicate detection processes specifically aimed at MSs' institutions.

Due to the inherent differences in the MSs reporting systems, ADR databases, number and type of ADRs received, the information presented in this document cannot be applicable for all settings and is intended to provide ideas that can be used for adjusting MSs' duplicate detection process.

[b] total number of respondents

1.2 Definitions and abbreviations

Terminology	Description
ADR	Adverse Drug Reaction
ATC	Anatomical Therapeutic Chemical Classification
ANSM	Agence Nationale de Sécurité du Médicament et des Produits de Santé, French Medicines Agency'
DEC	Drug Event Combinations
DKMA	Danish Medicines Agency
EMA	European Medicines Agency
HCP	Healthcare Professional
GVP	Good Vigilance Practice
ICSR	Individual Case Safety Report
Lareb	The Netherlands Pharmacovigilance Centre
MAH	Market Authorisation Holder
MHRA	Medicines and Healthcare Products Regulatory Agency
MS	Member State
NCA	National Competent Authority
PRAC	Pharmacovigilance Risk Assessment Committee
PT	MedDRA Preferred Term
SCOPE	Strengthening Collaboration for Operating Pharmacovigilance in Europe
SUKL	Státní ústav pro kontrolu lé iv, Czech Medicines Agency
UMC	Uppsala Monitoring Centre
WP	Work Package
WHO	World Health Organization

2. The duplicate detection process

The strength of spontaneous reporting systems is that they cover all types of legal medicinal products used in any setting. In addition to this, the reporting systems are built to obtain information specifically on potential ADRs and the data collection concentrates on variables relevant to this objective and directs reporters towards careful coding and communication of all aspects of an ADR. The increase in systematic collection of ICSRs in large electronic databases has allowed the application of data mining and statistical techniques for the detection of safety signals. There are known limitations to spontaneous ADR reporting systems, which include limitations embedded in the concept of voluntary reporting, whereby known or unknown external factors may influence the reporting rate and data quality. ICSRs may be limited in their utility by a lack of data for an accurate quantification of the frequency of events or the identification of possible risk factors for their occurrence. For these reasons, any signal from spontaneous reports needs to be verified clinically before further communication.[4]

Duplication of cases is an important data quality problem and can pose significant problems for analysing signals arising from pharmacovigilance databases by misleading clinical assessment or distorting statistical screening, both artificially inflating and masking signals of disproportionate reporting.[7] Duplicates are separate and unlinked records that refer to one and the same case of a suspected ADR.[9]

The potential sources of duplication of cases in the ADR databases are different reporters submitting a report for the same patient, the same reporter reporting to the NCA as well as to the MAH, and situations where multiple pharmaceutical companies hold a product licence and all have a requirement to forward the report to the appropriate regulatory authority.[4] Sometimes a duplicate might be the result of the MS receiving follow-up cases that had not been linked to the original report.

ADR reports can be, in many instances, dissimilar despite being duplicates: different terms may have been used to code the same incident, patient information may have a different level of specificity due to differences in the personal data protection rules in MSs, etc. Regardless of the system used for collecting and collating ICSRs, there should always be an appropriate mechanism in place for identifying duplicates. When a single suspected ADR incident yields several reports, this may divert the signal detection process and analysis. Some studies indicate that duplicates may account for as large a proportion as 5% of all reports.[7] Suspected report duplication appears not to be evenly spread in the data set, but whereas most reports have no suspected duplicates, a small minority have several. Higher rates of suspected duplicates are observed for literature reports (11%) and reports with fatal outcome (5%), whereas a lower rate was observed for reports from consumers and non-health professionals (0.5%).[9]

Handling duplicate reports generally involves three steps: detection of duplicates, confirmation of duplicates and management of duplicates. High-level business process maps and process descriptions in relation to the quality review of ICSRs and the detection and management of duplicate ICSRs are provided in Good Vigilance Practice Module VI Appendix 6 and Appendix 7. Further guidance on the detection of duplicate ICSRs is available in the CHMP Guideline on Detection and Management of Duplicate Individual Cases and Individual Case Safety Reports (ICSRs). Management of duplicates during EMA literature detection is described in the document Process description for managing duplicates in the context of the Medical Literature Monitoring in (MLM) service.

2.1 Detection of duplicate cases

The first step in duplicate management is to identify possible ADR report duplicates. Screening for duplicates is usually carried out when a new report arrives in the database, i.e. during data entry or during the process of loading ICSRs that have been received electronically. Duplicates can be detected even before entering the case into the database – that is usually the case during manual data entry of cases received by paper, fax or phone. Duplicates can also be detected during periodic data review and the signal management process when detailed analysis of cases is performed.

Potential duplicates can be detected by individual case review of all reports or by computerised duplicate detection algorithm, depending of the database set-up and number of ADR reports in the database. Duplicate searches are generally based on similarities in patient, adverse reaction and medicinal product data. Different search criteria may be suitable for different datasets. For pharmacovigilance systems that do not have to deal with large datasets, a simple table which sorts the reports by age, sex, suspected/interacting medicinal products and adverse events/reactions can be sufficient to detect possible duplicates.

Patient identifiers used for the detection of duplicates are usually patient initials, sex, age and birth date. Reaction details used for the analysis can be the reaction term, usually at MedDRA Preferred Term (PT) level, and reaction dates. Higher-level MedDRA terms might also be used for comparison in order to avoid losing the potential duplicates when the same reaction is coded with a very similar but different PT. With regard to medicinal product data, suspect drug at substance level is mainly used. ADR data, such as country of origin, safety report number and company number details, can also be analysed, mainly in cases where duplicate detection has been performed by an algorithm.

Large organisations often rely on computerised duplicate detection, either in house or as part of a commercial software package, followed by subsequent manual review.[9] The details of these duplicate detection algorithms are generally not published, but they tend to rely on the following premise: if two reports match on certain defined fields, then they are suspected duplicates.

In some instances, more than one algorithm might be applied to detect duplicates (e.g. an industry-specific algorithm using the data field specific to reports received from the industry, general algorithm etc.). Different algorithms use different data fields for comparison.

Adequately chosen number and type of data fields for the duplicate analysis should ensure satisfactory sensitivity and specificity – choosing a small number of general data fields can result in a large list of potential duplicates that have to be manually confirmed, which is time-consuming. On the other hand, by choosing a large set of ADR data fields that need to match, some cases might be missed.

Duplicate detection algorithms should be periodically evaluated and updated as necessary. A good example of algorithm optimization is a new approach applied by Uppsala Monitoring Centre, which uses a probabilistic method to detect duplicate ICSRs. VigiMatch is a duplicate detection algorithm based on the hit-miss model for record matching. The hit-miss model is a likelihood-based approach to identify unexpectedly similar record pairs in large databases. It computes a match score for each pair of records, where matching information is rewarded and mismatching information penalised. This match score reflects the probability that the two records relate to the same underlying entity or, in this setting, that they are duplicates.[8,9] Record pairs with match scores that exceed a certain threshold are flagged as suspected duplicates. This logic is more accurate and assures higher sensitivity and specificity of duplicate detection.

In conjunction with identification of duplicates in the ADR database, efforts can be made to reduce the number of duplicates by other means, e.g. for literature reports, which are generally one of the biggest sources of duplicates in the database. In order to reduce the number of duplicates deriving from the literature, the MHRA publishes the literature report list of scientific literature relating to ADR reports on its webpage. If an MAH identifies a literature article related to an ADR, it should be checked to see if the literature article is on the published literature report list, and if it is, there is no need to send the ADR report to the MHRA. If the ADR report is not already listed, then it should be submitted to the MHRA. The literature report list should be updated on weekly basis.

In addition, the importance of the control of data entry processes should be mentioned: data quality is often compromised by many factors and can affect the duplicate detection process. This includes data entry errors (e.g. wrong patient initials), missing integrity constraints (e.g., allowing entries such as patient age ¼ or 567), and multiple conventions for recording information (e.g. Y.X. instead of X.Y. for patient initials). In independently managed databases, not only the values, but also the structure, semantics, and underlying assumptions about the data may differ as well. For this reason, it is important for ADR reporting forms to have a clean structure and to allow unambiguous data field entry.

2.2 Confirmation of duplicate cases

Upon identification of potential duplicates, a manual confirmation based on the assessors' judgement will always be necessary. A well-documented case, including a case narrative, is a prerequisite to confirm if two cases are duplicates. Following assessment, there are four possible outcomes:

- The case is not a duplicate
- More information is needed
- The case is a duplicate from different sender
- The case is a duplicate from the same sender.[5]

The confirmation process is performed by the assessor. In order to be able to compare two cases, all available information from the individual report should be taken into account. In situations when it cannot be determined if the case is a duplicate and more information is needed, follow-up information should be requested. Follow-up information is more likely to be retrieved when the duplicate check is done during the initial import of the case into the database than at the signal detection stage, when more time might have elapsed.

The outcome of all assessments should be recorded to avoid continually reassessing the same cases when further versions arrive. These recorded outcomes can also be used to refine the duplicate detection methods during future development.

2.3 Management of duplicate cases

Management of confirmed duplicate cases represents the most challenging part of the process.

Duplicate cases are generally managed through a process of merging two-or-more cases into one 'master case'. This process can consist of one of the following approaches:

- The master case can be based on one of the existing cases, with information from the other subordinate duplicate case added, unless the same, or more-precise, information is already present in the master case, or
- The master case can be created as a new case, combining the information from the subordinate duplicate cases.

In cases where the master case is based on one of the existing cases, it needs to be decided which one of them will become the master case. Usually the one which has been imported first becomes the master case. The other option is to use a qualitative approach and choose the more complete case as the master case. In order to decide which is considered a more complete case, a completeness score might be applied. Completeness scores can be measured manually or applied as an automated algorithm for larger datasets. In addition, sometimes a case is considered the master case by default if certain data fields have been populated (e.g. information on the fatal outcome). In some instances, NCAs can decide to choose a health professional report as a master case rather than a report received from the patient, spontaneous versus literature or study report, etc.

A good example for simplifying case management in case of duplication is the use of a computer program that allow a parallel screen view of an ADR report and its possible duplicate case. The format of the presented reports allows easy creation of a master case by ticking the ADR data fields containing information that needs to be included in the master case.

Regardless of the approach, the master case should always contain all the case reference numbers from all subordinate duplicate cases, so that they can be easily traced. The master case should reflect the most accurate and up-to-date information available to the organisation; however, there are situations where information in the duplicate cases is not the same or is even conflicting: all this should be written down in the case narrative and follow-up should be attempted. After the creation of master case files, one of the reports becomes redundant and can be nullified.

All subordinate duplicate cases and related ICSRs should be retained and there should be adequate cross-referencing between case files and/or pharmacovigilance database entries. If follow-up information is received for any of the duplicated cases, the master case should be updated accordingly.

More detailed guidance on duplicate management can be found in the CHMP Guideline on Detection and Management of Duplicate Individual Cases and Individual Case Safety Reports (ICSRs), and will not be repeated here.

Whatever the specificities of the duplicate detection procedure are, for the purpose of applying consistent standards and practices within the MS, it might be useful to apply work instructions or a standard operating procedure describing the duplicate detection procedure.

3. Conclusions

Duplicates are separate and unlinked records that refer to one and the same case of a suspected ADR. Duplication of cases is an important data quality problem and can pose significant problems for analysing signals arising from pharmacovigilance databases by misleading clinical assessment or distorting statistical screening, both by artificially inflating signals and by masking signals of disproportionate reporting.[6]

The results of the SCOPE questionnaire indicate that over 92% (25/27[c]) of MSs perform duplicate detection. SCOPE results also revealed differences among MSs with regards to duplicate detection processes, ranging from MSs who have more than one duplicate detection system in place (e.g. a computer algorithm and SOPs/written procedures) to MS without any duplicate detection system in place.

Based on the SCOPE results and the importance of the issue, it was considered relevant to summarize the applicable EU guidance documents related to the duplicate detection process, to promote the importance of duplicate detection, to emphasize the key points of the process and to provide an overview of current MSs' experiences in relation to duplicate management, indicating examples of good practice.

Although there are differences between MSs with regard to the duplicate detection, several practices can generally be considered as a good example: automatisation of the process for creating a list of potential duplicates (duplicate identification); moving towards more sophisticated algorithms for duplicate detection; detection of duplicates during the data entry stage rather than during the signal detection process; having written procedures in place for ensuring the consistency of management of duplicates between the assessors.

This document should also serve as a starting point for the development of a learning tool on duplicate detection processes specifically aimed at MSs. This learning tool will allow the use of the experiences from the SCOPE project in a more sustainable way and can also enhance the use of available guidance documents related to duplicate detection processes.

Due to the inherent differences in MSs' reporting systems, ADR databases, number and type of ADRs received, the information presented in this document cannot be applicable for all settings and is intended to provide ideas that can be used to adjust MSs' duplicate detection process.

[c] total number of respondents

5. References

1. Directive 2010/84/EU of the European Parliament and of the Council
2. Regulation (EU) No 1235/2010 of the European Parliament and of the Council
3. Commission Implementing Regulation (EU) No 520/2012
4. Guideline on good pharmacovigilance practices (GVP) Module VI – Management and reporting of adverse reactions to medicinal products (Rev 1) EMA/873138/2011 Rev 1*8 September 2014
5. CHMP Guideline on Detection and Management of Duplicate Individual Cases and Individual Case Safety Reports (ICSRs), Doc. Ref. EMA/13432/2009, London, 21 June 2012
6. ENCePP Guide on Methodological Standards in Pharmacoepidemiology, EMA/95098/2010 Rev.3
7. Norén GN, Bate A, Orre R. A hit-miss model for duplicate detection in the WHO drug safety database. In – KDD '05: Proceedings of the 11th ACM SIGKDD conference on Knowledge Discovery and Datamining, 2005, 459-468
8. Norén, G. N., Orre, R., Bate, A., Edwards, I. R. Duplicate detection in adverse drug reaction surveillance. Data Mining and Knowledge Discovery. 2007. Published online
9. Tregunno PM, Fink DB, Fernandez-Fernandez C, Lázaro-Bengoa E, Norén GN. Performance of Probabilistic Method to Detect Duplicate Individual Case Safety Reports. Drug Safety, 2014, 37(4):249-258
10. EU Individual Case Safety Report (ICSR) Implementation Guide, EMA/51938/2014, 4 December 2014.

Annexes: Case studies of duplicate detection process management

Annex 1. The United Kingdom

The UK has a long tradition of ADR collection through the Yellow Card Scheme (YCS). Yellow Card reports are submitted to the Medicines and Healthcare products Regulatory Agency (MHRA) from all types of healthcare professional, Marketing Authorisation Holders and members of the public. The cases are entered into a custom built national ADR database holding spontaneous ADR reports dating back to 1963. In 2015, approximately 39,000 ADR reports were received by the MHRA.

Reports can be received on paper, electronically through a Yellow Card website, directly from healthcare clinical systems and via the telephone. As soon as the report is received it is entered into the ADR database within strict deadlines by a multidisciplinary team of life science graduates and pharmacists. The reports go through four stages of processing: the first step, 'input verification', confirms the case is valid and assigns a priority (based upon the seriousness of the case). The second step, 'data capture', is where all the detail of the case is entered into the database. The third step is a quality assurance step to confirm the coding of the case is correct, and the fourth is an assessment step to determine the completeness of the information and whether a request for further information from the reporter is required.

The duplicate detection process begins with identification of possible duplicates by comparing the newly received ADRs against the data already in the database. The identification of potential duplicates is done automatically by using an algorithm that analyses ADR data elements, such as suspect drug, patient, reporter details and MAH reference numbers. The query is run on a weekly basis and potential duplicates are outputted into a report, as well as being identified through an alert on the cases in the ADR database.

The report with cases identified as potential duplicates by the algorithm are shared with assessors in the form of an Excel spreadsheet. The spreadsheet documents the identified potential duplicates, along with some key case details (ADR number, suspect drug, patient sex/initials/age/weight, reporter details, company details, sender report numbers and the report date). The assessors are sent this spreadsheet by team managers and each assessor must review the cases they have been allocated and confirm which are true duplicates that require merging.

When a case has been identified as a potential duplicate, an alert is automatically added to the case and a hyperlink to the case suspected to be a duplicate is added to the 'Duplicates tab' within the electronic case folder, indicating that the case is a possible duplicate.

Should a duplicate case be identified outside of the above mentioned procedure (e.g. during the signal detection process or as a result of an enquiry from an MAH) it is possible to manually populate the 'Duplicates tab' in the case folder, at any stage of the workflow.

Every possible duplicate has to be assessed and a decision recorded, regardless of whether it is confirmed as a duplicate or not. An internal guidance document is available to help assessors decide on whether two cases are duplicates and should be merged. If an assessor is still unsure they must seek advice from a team manager. The technical process of merging duplicate cases on the ADR database is outlined in a Standard Operating Procedure.

When a potential duplicate pair have been checked and found not to be duplicates, the assessor will remove the potential duplicate from the duplicates tab of the case folder. This then prevents the pair being considered as potential duplicates again by the duplicates algorithm.

If the cases are confirmed to be duplicates, a master case is selected prior to beginning the merge process. When deciding which report will be the master case, it is typically the case with the most or latest information already populated. Alternatively, if a report has been received from multiple sources, then the case that is received directly from a healthcare professional or member of the public is chosen as the master in preference to a case received indirectly through an MAH.

To reduce the amount of case processing required for merged duplicate cases, the assessor carrying out the merging of the cases can decide what information to keep from both case folders by using a comparison tool within the ADR database. The comparison tool allows parallel viewing of two potential duplicate cases on the screen, displaying all data fields which contain differing information between the two cases. The assessor can then select the most appropriate information to populate the master case.

When cases are merged, these appear in the 'Duplicate ADRs' section of the duplicates tab in both cases for audit purposes. The master case gets an updated version created in the database. All original documentation from both cases is stored in the master case folder and the case must progress through the workflow as usual. The duplicate case is nullified on the database.

Literature reports are generally one of the biggest sources of duplicates in the database and place a significant burden on both the NCA and MAHs. In order to reduce the number of duplicates deriving from the literature the MHRA has published on its webpage a report listing all scientific literature relating to ADR reports received: https://www.gov.uk/guidance/send-and-receive-information-on-adverse-drug-reactions-adrs#literature-report-list. This list is updated on a weekly basis.

When an MAH identifies a literature article containing an ADR, which qualifies for expedited reporting, they should check to see if the literature report is on the Excel spreadsheet published by the MHRA; if it is there then there is no need to send it to the MHRA. If the ADR is not already listed, it should be submitted to MHRA via the usual reporting processes. The MHRA is currently determining whether the literature list on its website is still a useful tool for MAHs given the introduction of centralised literature monitoring by the EMA in September 2015.

In 2014, the MHRA was leading a work package as part of the IMI Protect project, investigating innovative approaches to duplicate detection in spontaneous data. This involved testing a novel approach developed by the Uppsala Monitoring Centre using probabilistic record matching[d]. This work indicated significant benefits for MHRA duplicate detection, which would bring efficiencies for handling of duplicates within the Pharmacovigilance Information Unit, particularly in light of the increasing numbers of ADR reports received by the MHRA in recent years. It is anticipated the MHRA will transition to this new model for its duplicate detection algorithm in the near future.

[d] http://link.springer.com/article/10.1007/s40264-014-0146-y

Annex 2. The Netherlands Pharmacovigilance Centre (Lareb)

The Netherlands Pharmacovigilance Centre (Lareb) has been receiving ADR reports since 1986. Initially, only reports from healthcare professionals were received, but in 2003 consumer reporting was introduced. Additionally, reports from Marketing Authorisation Holders (MAHs) concerning ADRs from the Netherlands are imported into the Lareb database from the Eudravigilance database. Reports from ADRs can be submitted via the Lareb website, by post or through a recently developed app for tablets and smartphones, resulting in approximately 24,000 ADR reports in 2015.

At Lareb, separate duplicate detection algorithms were developed for vaccine and non-vaccine reports, since they have distinct characteristics in terms of relevant identifiers for duplicate detection.

For non-vaccine reports, the automated duplicate detection algorithm is based on age/date of birth, sex, received date, suspect drug (based on generic product code from a Dutch thesaurus) and ADR (based on MedDRA High Level Term (HLT)). Since the majority of vaccine reports concern individuals of similar age, and the reported types of ADRs are rather limited, the duplicate detection algorithm for these reports is based on age/date of birth, sex, suspect drug and drug start date (for vaccines this is the vaccine administration date). Duplicate detection for MAH reports is based on the non-vaccine algorithm, since Lareb rarely receives vaccine reports from this source.

In general, the duplicate detection process consists of two steps: first, each new ADR report is screened by a fully automated duplicate detection algorithm during the assessment procedure of individual case reports. Then, when a possible duplicate report is detected by the algorithm, the assessor gets an alert and manually evaluates it if it concerns a true duplicate.

Although most of the duplicate reports are detected during the case-by-case assessment procedure, detection during the screening procedure may also occur. If this is the case, assessors can manually process these reports as duplicates.

Once a set of reports has been designated as duplicates, the responsible assessor appoints one of them as the master case report. The choice for the master report is based on the quality of the reports and is at the discretion of the assessor. Lareb has chosen not to merge two reports into one new master report, since this would lead to interpretation issues in the case of contradictory information in the separate reports. In addition, follow-up information that needs to be added to the original reports can be received after the duplicate detection process is complete.

Annex 3. The Czech Republic

The Czech Republic introduced ADR collection in 1970, but the current Czech database contains data from 2004 with roughly 2200 ADR reports per year.

The ADR reports received by SUKL from various sources are entered into the national database. The duplicate detection process begins with identification of possible duplicates among the newly received ADRs against the data already in the database.

The identification of potential duplicates is done automatically by using an algorithm after entering the reported data to the electronic report form. The algorithm analyses ADR data elements such as patient and adverse reaction details. Initials, sex, age and birth date are used as patient identifiers. The reaction detail used for the analysis is the adverse reaction start date. This algorithm, implemented in 2006 at the time of electronic database development, has not proved to be very appropriate because the given list of potential duplicates is usually too extensive and inaccurate. More data fields to the duplicate detection algorithm are going to be added in a new database, which is under development now, to increase the precision of the search. Data elements planned to be added are: suspect drug name (or, better, an active substance level of the suspect drug or their ATC – as the suspect drug brand name is often not known), SOC MedDRA term of the reported ADRs (again to cover the possibility that the observed ADR is named differently by various reporters) and suspect drug administration start date. It is hope that this change will significantly improve the algorithm and the resulting candidate list will be more helpful in the search for duplicates.

Upon the identification of potential duplicates, cases need to be manually confirmed by the PHV data manager (sometimes in cooperation with PHV assessor). When a case is confirmed as a duplicate it is merged with the initial case in order to create a master case. Usually, the case which was received earlier or has more complete data – i.e. a better case description – becomes the master case. Additional data from the duplicate case is then transferred to the original case, together with duplicate case identifiers in 'Report duplicate' fields (A.1.11.1 and A.1.11.2 in R2 format) of the report. The duplicate case is given the status of 'duplicate case' in the ADR database, with the master case number written down. Both master and duplicate cases are visible in the database and in the data analysis tool, but, using the appropriate search filter, only master cases (or duplicates) can be tracked. Duplicate cases are only nullified in the EV database once both master and duplicate have been transmitted to the Eudravigilance database.

The last part of duplicate detection process is the assessment of the newly received ADR reports by PHV assessors on a weekly basis. The assessors work with the ADR report list from the given time period (the last week) and they assess the ADR reports according to ATCs allocated to them for assessment. Through detailed, case-by-case assessment, any remaining potential duplicates which have not been recognised by automatic processes are identified. The process of duplicate detection management is described in a Standard Operating Procedure.

Annex 4. Uppsala Monitoring Centre (UMC)

Uppsala Monitoring Centre (UMC) is the World Health Organization's Collaborating Centre for International Drug Monitoring. Responsibility for managing the WHO Programme for International Drug Monitoring has been held by the UMC since 1978. Today, more than 120 member countries have been included in the WHO Programme for International Drug Monitoring.

One of the main tasks of the UMC is to collect, assess and communicate information from member countries about the benefits, harms and risks of drugs and other substances used in medicine to improve patient therapy and public health worldwide. The WHO global individual case safety report database is called VigiBase and is maintained and developed by the UMC. The UMC develops and provides several tools for use by organisations involved in drug safety, tools for searching in the database, and a program for case report management called VigiFlow. In order to be able to adequately analyse information deriving from a large number of spontaneous reports, UMC developed a duplicate detection algorithm (vigiMatchTM) for individual case safety reports and implemented it for routine use in VigiBase. Duplicate detection method details and performance are described in articles published by Norén et al., Duplicate detection in adverse drug reaction surveillance, and Tregunno et. al., Performance of Probabilistic Method to Detect Duplicate Individual Case Safety Reports.

Unlike rule-based methods, UMC's probabilistic approach allows for effective matching in the presence of errors and a variable amount of information in reports. While based on a sophisticated mathematical model, the algorithm operates in intuitive ways. Matches are rewarded based on how common the matching events are, and mismatches are penalized based on how common mismatches are for that record field among manually identified duplicates. For example, a match on gender receives a more modest reward than a match on an adverse drug reaction, since two randomly selected reports are more likely to list the same gender than to list the same adverse drug reaction. Similarly, a drug mismatch receives a greater penalty than an ADR mismatch, since mismatches on listed drugs are more rare than mismatches on listed ADRs, which may be expected on account of the more clear-cut division between different medical diagnoses. In its routine signal detection, UMC excludes from first-pass screening all suspected duplicates identified by vigiMatch. However, in subsequent assessment, all reports are available for review. At the same time, the national centres are requested to review suspected duplicates from their country and identify any confirmed duplicates that can be deactivated (or merged with an existing record).

Upon the identification of potential duplicates, cases need to be managed in order to carry out the de-duplication. The most complete ICSR becomes a 'Preferred case'. Due to the large amount of ICSRs, the selection of preferred cases is performed automatically based on the vigiGrade completeness score.

At the moment, there is an ongoing 'pilot project' to evaluate how to communicate and how to make the output from UMCs duplicate detection algorithm, vigiMatch, as useful as possible for all countries that share their PV data with the WHO Programme for International Drug Monitoring. The aim is that each member country should eventually have the option of getting their vigiMatch result and using it for the optimization of their internal processes.