Bridging Data Models and Terminologies to Support Adverse Drug Event Reporting Using EHR Data*

G. Declerck; S. Hussain; C. Daniel; M. Yuksel; G. B. Laleci; M. Twagirumukiza; M.-C. Jaulent

1 INSRM UMRS 1142, Paris, France; 2 SRDC Ltd, Ankara, Turkey; 3 AGFA HealthCare, Mortsel, Belgium

Summary

Introduction: This article is part of the Focus Theme of Methods of Information in Medicine on “Managing Interoperability and Complexity in Health Systems”.

Background: SALUS project aims at building an interoperability platform and a dedicated toolkit to enable secondary use of electronic health records (EHR) data for post marketing drug surveillance. An important component of this toolkit is a drug-related adverse events (AE) reporting system designed to facilitate and accelerate the reporting process using automatic prepopulation mechanisms.

Objective: To demonstrate SALUS approach for establishing syntactic and semantic interoperability for AE reporting.

Method: Standard (e.g. HL7 CDA-CCD) and proprietary EHR data models are mapped to the E2B(R2) data model via SALUS Common Information Model. Terminology mapping and terminology reasoning services are designed to ensure the automatic conversion of source EHR terminologies (e.g. ICD-9-CM, ICD-10, LOINC or SNOMED-CT) to the target terminology MedDRA which is expected in AE reporting forms. A validated set of terminology mappings is used to ensure the reliability of the reasoning mechanisms.

Results: The percentage of data elements of a standard E2B report that can be completed automatically has been estimated for two pilot sites. In the best scenario (i.e. the available fields in the EHR have actually been filled), only 36% (pilot site 1) and 38% (pilot site 2) of E2B data elements remain to be filled manually. In addition, most of these data elements shall not be filled in each report.

Conclusion: SALUS platform’s interoperability solutions enable partial automation of the AE reporting process, which could contribute to improve current spontaneous reporting practices and reduce under-reporting, which is currently one major obstacle in the process of acquisition of pharmacovigilance data.

Keywords
Pharmacovigilance, adverse drug event reporting, semantic interoperability, EHR data models, secondary use of EHR

1. Introduction

Under-reporting of drug-related adverse events (AE) is a well known phenomenon that has been extensively studied [1 – 3]. It has been estimated that only around 1 to 5% of AEs are reported to health regulatory authorities through spontaneous reporting systems [4, 5]. This phenomenon results in delayed acquisition of knowledge about adverse effects and thus potentially causes substantial patient casualties and major costs for public health [6, 7]. One major cause of under-reporting identified in the literature is the tedious and time-consuming character of the reporting process, double data entry, together with the lack of awareness of the importance of reporting AEs for patient safety [8 –10]. In this paper, we focus on the hypothesis that secondary use of electronic health records (EHR) data can support the process of AE reporting. Part of the data requested in AE reporting forms – e.g. demographics, lab test results, symptoms or past drug history – are collected in EHR systems and could be automatically reused [11]. Such automatic pre-population mechanism could facilitate and accelerate the reporting process and thus contribute to increase the volume of reported AEs.

Efforts have already been made to support the reporting of AEs with automated detection systems providing alerts to the physician [5 –12] or extracting patient data from EHRs to prepopulate automatically AE reporting forms [11, 12]. One of the recent attempts in solving the under-reporting problem by reusing EHR data is ASTER (Adverse drug events Spontaneous Triggered Event Reporting) proof of concept pilot project [11 –13]. ASTER application enables automatic extraction of data from EHR to prepopulate an AE report and electronic submission to the Food and Drug Administration (FDA). However, the ASTER application has two main limitations: i) the completed form has to be processed manually by an intermediate instance, in charge of putting the form in the proper format expected by the FDA; ii) the
extraction of EHR data is not built to be interoperable with several EHR data models. We think that such limitations have to be overcome for building an AE reporting system able to counteract the aforementioned under-reporting phenomenon: a) the post-processing of the submitted reports must be minimal to avoid overloading the pharmacovigilance authorities; b) patient data extraction must be possible independently of the EHR data model to ensure the scalability and the sustainability of the reporting system.

One of the main technical challenges for building an EHR enabled AE reporting system is establishing interoperability between EHR systems and regulatory bodies collecting AE reports. Current EHR systems use heterogeneous data models to record patient information in hospital data warehouses [14]. On the clinical side, several standard-based information models have been proposed to the needs of sharing or exchanging data, such as CEN/ISO 13606-5 Archetypes/Templates or HL7’s Detailed Clinical Models (DCMs), e.g. HL7 Clinical Document Architecture (CDA) meta-standard and the derived Continuity of Care Document (CCD) and Patient Care Coordination (PCC) templates [15, 16]. These models require an associated, robust data type model such as that defined by ISO 21090, and clinical terminologies such as ICD-10 (International Classification of Diseases, 10th revision) or SNOMED-CT – to the tar-marketing services to transform the source medical terminologies used in EHRs – e.g. ICD-9-CM (International Classification of Diseases, 9th Revision, Clinical Modification), ICD-10, LOINC (Logical Observation Identifiers Names and Codes) or SNOMED-CT – to the targeted terminology MedDRA (Medical Dictionary for Regulatory Activities) for AE reporting.

2. Bridging EHR Data Models and AE Reporting Standards

SALUS project aims to enable the re-use of heterogeneous EHR data for running clinical research studies and AE reporting. This is achieved through the SALUS Semantic Interoperability Layer, in the core of which there is a common information model (CIM) represented as an RDF ontology to act as common denominator. To collect clinical data from the EHR systems or CDWs, the local data models are mapped to SALUS CIM ontology through translation rules. In the following subsections, SALUS CIM ontology and the mediation process are detailed.

2.1 E2B(R2) Data Model

E2B is an international standard developed by the International Conference on Harmonisation (ICH) and promoted by the World Health Organization (WHO), the FDA and the European Medicines Agency (EMA) for electronic reporting of AEs. E2B is used by the Uppsala Collaborating Centre for International Drug Monitoring (UMC), which centralizes a unique database (VigiBase) of spontaneous AE reports submitted all across the world. The E2B specification is both i) a data model in a machine-processable format for providing relevant information when reporting an AE; ii) a protocol describing how the report should be transmitted electronically to regulatory authorities. The E2B data model includes more than 230 data elements – only a few of them are mandatory. The commonly used version of E2B is release 2 (R2), but a third release (R3) is achieved and under deployment. E2B(R2) is used in SALUS.

2.2 SALUS Common Information Model (CIM)

To ensure the possibility of prepopulating E2B(R2) compliant AE reporting forms with patient data from EHR systems using different data models, we have aligned standardized EHR data models to E2B(R2) by converging the overlapping entities of both models to a harmonized data model specific to SALUS: the CIM. This pivot model, represented in an ontological format, enables to rely on unique correspondences to link together the standard EHR data models integrated and interoperable with E2B standard (Figure 1). The CIM Ontology thus acts as a common denominator for the existing standards used in clinical care as well as for proprietary models. Ensuring interoperability...
with proprietary models is important because EHR data remain based on such models in many healthcare facilities, despite the on growing dissemination of standard-based clinical documentation. The CIM Ontology is built through a systematic approach by i) examining the content models, ii) extracting and harmonizing Common Data Elements (CDEs) from these, iii) representing the related terminology systems as ontologies within the SALUS Semantic Resource Set, and iv) linking them with the CDEs in an ontological framework. We have focused on the requirements of a set of selected pilot applications by identifying and formally representing CDEs as an RDF model. In this process, we have analyzed and taken into account the content models from other standards and initiatives as well, to provide a common mediator that can interoperate with well-established state of the art. These include HL7/ASTM CCD and IHE PCC templates, HL7 Health Quality Measures Format (HQMF), HITSP C32/C83 components, Consolidated CDA templates, ICH E2B(R2) and ISO/CEN EN 13606 archetypes. We have also analyzed the Common Data Model (CDM) of the Observational Medical Outcomes Partnership (OMOP), which is used as a target model in three of our pilot applications to carry out statistical analyses on top of the EHR data extracted. The main CDEs resulting from this selection process are: demographics, address, healthcareProvider, dataReporter, diagnosis, adverseEvents (AEs), vitalSigns, labResults, encounterVisit, encounterStay, familyHistory, socialHistory, immunotherapy, medication, medicationOrder, procedure, pregnancy, fulfillmentHistory. After identifying the required CDEs we have modeled them in RDF using existing ontologies and terminologies such as schema.org or SNOMED CT. This strategy has been chosen to avoid creating from scratch entities that are already defined by existing resources and to favor the re-use of our entity models in the healthcare and EHR communities.

SALUS CIM represents the pivot ontology used for semantic mediation, and is the core of Semantic Interoperability Layer. As a part of Semantic Interoperability Layer, mapping rules are defined to translate EHR data from local EHR models to SALUS CIM as will be explained in the next section.

2.3 Bridging Local EHR Models to SALUS CIM

SALUS Technical Interoperability Layer together with Semantic Interoperability Layer allow accessing disparate EHRs sources to retrieve structured patient medical summaries to be used for the preprocessing of AE reporting forms. Two EHR sources are available through SALUS pilot sites: 1) a regional CDW maintained in Lombardy Region in Italy, which collects and extracts data necessary for administrative and statistical purposes from almost all the public healthcare providers; 2) the AGFA ORBIS installation used as the EHR system at Uniklinikum Dresden (UKD), which is the largest hospital structure with 21 clinics in Saxony, Germany. A non-disruptive approach is used to collect the data in the local models used by both systems and make it available in SALUS CIM based on query mechanisms (see Figure 1 in the Web-based Appendix).

Lombardy Regional Health Infrastructure provides medical summaries of patients in HL7 CDA-CCD/PCC templates. Through SALUS Technical Interoperability Layer, given a patient id, the medical summaries represented in CCD/PCC templates can be collected via the IHE Query for Existing Data (QED) [24] transactions. For the UKD ORBIS system which does not support HL7 CDA interfaces on the other hand, we built a SPARQL query interface to access the medical data sets in the local ORBIS proprietary relational data model. As a result the medical summaries are available in ORBIS Entity Model as the local content entity model. SALUS Semantic Interoperability Layer provides a unified RESTful interface on top of these varying interfaces with different models, so that the end-user tools such as the AE reporting tool access to the comprehensive patient data in a consistent way.

The collected medical summaries from EHRs are represented as instances of local content entity models, and these models are then semantically transformed as SALUS CIM Ontology instances using a set of pre-defined conversion rules. The mapping rules for CDA-CCD based local entity models are detailed in Section 2.4. A simi...
lar approach has been followed to define the mapping rules between ORBIS Entity Model and CIM. The transformed medical summaries, represented in SALUS CIM, enable prepopulation of AE reporting forms in E2B format.

2.4 Bridging CDA-CCD and E2B(R2) Data Models via SALUS CIM

Several initiatives have been made to specify how data should be extracted from EHRs to prepopulate AE reporting forms. The most advanced to date is certainly the IHE Drug Safety Content profile (DSC) [25], an integration profile built as an add-on to the Retrieve Form for Data Capture profile (RFD) [26], aiming at specifying a generic protocol for handling information collected from CCD-based extracts from EHRs to prepopulate AE reporting forms.

DSC focuses on the definition of the RFD retrieve form transaction, and describes the data needed to pre-fill reporting forms using CCD and how to convert it to E2B. However: a) mappings to E2B are only partially defined (some mapping are missed, CCD XPaths are not specified); b) conversion rules needed to map the value sets and value formats of the EHR data elements to the value sets and value formats of the E2B data elements are not specified; c) the direct CCD-to-E2B mapping method used by RFD is poorly scalable: when other EHR information models need to be mapped to E2B to support AE form prepopulation (or, conversely, when the alignment to a new release of E2B is needed), new mapping tables must be built from scratch. SALUS interoperability platform has been designed to overcome these limitations. An important effort was especially devoted to revise and complete DSC mappings and define conversion rules for value sets and value formats conversion to achieve a complete – or at least usable – alignment for E2B form prepopulation.

It must be noted that not every E2B data element can or has to be prepopulated using EHR data. We categorize E2B data elements into following main types (see also Table 3 in Section 5): i) automatically prepopulated data elements using EHR data (assuming that the required field values are present in EHR for the considered patient); ii) automatically deduced data elements from the value of other E2B data elements, e.g. “Age at time of onset of reaction/event” can be calculated based on “Date of birth” and “Date of start of reaction/event”; “Duration of reaction/event” can potentially be calculated on the basis of “Start date” and “End date”, etc; iii) pre-defined and default data elements that usually do not change from one AE report to another, e.g. “Primary source”, “Sender”, “Receiver”; iv) manually filled data elements, as the related information is simply not available in the patient EHR or cannot be extracted in an automated manner, e.g. the seriousness of the AE being reported or the characterization of the role played by the drug in the suspected AE. Alignment to HL7 CDA-CCD templates was only defined for the first type of E2B data elements.

Table 1 in the Web-based Appendix describes an excerpt of E2B(R2) data elements, which can be prepopulated from the EHR data, and the mappings between these elements and SALUS CDEs. CDA-CCD XPath expressions used to extract patient data from hospital EHRs are highlighted, prior to its representation in the CIM ontology as RDF class instances. The IHE PCC templates are also used. Those mappings are necessary to ensure prepopulation of E2B(R2) compliant AE forms with the patient data extracted from the source EHRs.

3. Using SALUS Mediation Platform to Prepopulate AE Reporting Forms

When a new AE must be reported, the extracted patient summary based on SALUS CIM is queried using SPARQL (see example in Figure 2), and the queried data is used to prepopulate the AE reporting form, presented to the healthcare professional for further validation and completion.

Due to format and value set differences between source EHR data and targeted E2B, several conversion mechanisms, depending on the data types of the data elements, are applied before prepopulating
the AE reporting forms (Table 1 and Table 2).

As shown in Table 2, even for quite simple value sets, only a partial alignment can be achieved, and the conversion from one data model to the other implies a loss of information. For instance, the “undifferentiated” value of the “Administrative Gender” E2B element cannot be equated to the empty value of the “Patient sex” E2B element, because an empty value generally means that no information is available regarding the patient sex, which is very different from the statement that the patient gender is undifferentiated (as in the hermaphrodites case). Similarly, value set of the “Continuing” E2B element (which is used to specify if a medical episode – disease, surgical procedure, etc. – is still active) is less precise than the value set used in CDA-CCD to indicate the status of clinical observations (“Problem Status Observation” or “Procedure status” sections).

Partial mappings however provide a basis for prepopulation and the health care professional can still adjust and validate the prepopulated fields’ values in the E2B form.

### 4. Medical Terminology Reasoning and Conversion

EHR systems use different terminologies to describe patient data. Since most medical data, e.g. description of the AE, must be described with MedDRA in E2B, terminology reasoning is necessary to ensure prepopulation. In order to handle terminology reasoning within the SALUS Interoperability Framework, first of all we represent the necessary terminology systems as ontologies inside the SALUS Semantic Resource Set. For this, we prefer the well-established Simple Knowledge Organization System (SKOS) vocabulary [27]. When available, we retrieve the terminology systems (e.g. ICD-10, SNOMED-CT, MedDRA, ICD-9-CM) from BioPortal and further fine-tune them to stick to the SKOS vocabulary. For some others such as the German Modification of ICD-10 (ICD-10-GM), which is used in one of our pilot sites, we create their semantic representations ourselves. We create a “skos:ConceptScheme” for each terminology system, and a “skos:Concept” for each code. We represent the original hierarchical relationships within a terminology system with “skos:broader” property.

The second important step for achieving terminology reasoning is utilizing mappings across terminology systems. We benefit from several reliable mapping resources such as the OntoADR ontology [28] of the PROTECT project for one-to-one mappings between MedDRA and SNOMED-CT; OMOP Vocabulary for exact or broad matches from ICD-9-CM to SNOMED CT and from ICD-10-CM to SNOMED-CT; one-to-one mappings between ICD-10-GM and ICD-10 based on identical codes; and some further manually reviewed BioPortal mappings. We represent these mappings through the SKOS vocabulary, e.g. using the “skos:exactMatch” or “skos:broadMatch” properties.

Based on these relationships, we apply a series of terminology reasoning rules to calculate the transitive closures of hierarchical relationships and the mappings, e.g. for an ICD-10-GM code, a corresponding MedDRA term can be inferred via terminology reasoning, while no direct link between these two codes were originally asserted. We compute such specific mappings in advance to semantically enrich patient data represented in SALUS CIM (e.g. reaction codes, active ingredient codes, problem codes) with the codes from the terminology systems preferred by the requester, i.e. MedDRA in the case of the AE reporting tool. For example, assuming that the source code is "Anaphylactic shock due to serum, not elsewhere classified", coming from ICD-9-CM terminology, three MedDRA terms are proposed as candidate mappings to the user: "Anaphylactic shock", "Anaphylactic reaction" and "Anaphylactic transfusion reaction" (Figure 3).
5. Results

To evaluate the performances of the AE reporting form prepopulation process, we have estimated for both pilot sites the percentage of data elements of a standard E2B report that can be completed automatically (see Section 2.4).

We can see in Table 3 that in the best scenario (i.e. the available fields in the EHR have actually been filled), only 36% (for UKD) and 38.1% (for Lombardy) of E2B data elements remain to be filled manually (the data elements that can be filled once for all for subsequent reports are excluded). In addition, most of these data elements shall not be filled in each report. For instance, some data elements must be filled only in case of death of the patient. Others are used only when the submitted report warrant being evaluated together with other cases known from the reporter, or when additional document potentially relevant to evaluate the reported case are available. Conversely, this is of course the decision of the reporter to use or not use in the AE report the data available in the EHR, or to modify, discard or complete the data that has been prepopulated. The tool provides support to facilitate the completion of the form; it does not substitute to the reporter’s expertise.

6. Discussion

SALUS interoperability solutions provide an indisputable advantage to the supported AE reporting tool compared to existing solutions using EHR-enabled prepopulation, such as the ASTER system: i) the possibility of prepopulating E2B compliant XML reports ensures a minimal workload of the pharmacovigilance authorities when processing the submitted data: the report being ensured in the expected format; and ii) due to SALUS platform’s scalability, AE reporting tool can be plugged onto new EHR systems with minimal adaptations: the only requirement is to map the new EHR data model to SALUS CIM; additional mappings between the new EHR data model and E2B do not have to be defined. Similarly, when the E2B data model evolves (a third release has recently been published), the alignment to the multiple EHR data models plugged onto SALUS platform do not have to be amended one by one; only the mapping from E2B to SALUS CIM need to be updated. In addition, the SALUS interoperability solutions are generic: the aim is to make the EHR data available in a common model. In this paper we exemplified how this common model can be used to pre-fill AE reports; however, other pilot applications can be built easily to consume medical summaries in common model, all needs to be done is writing new SPARQL queries.

Table 3

<table>
<thead>
<tr>
<th>Completed automatically</th>
<th>Prepopulated from EHR</th>
<th>Based on calculation mechanisms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UKD pilot site</td>
<td>Lombardy pilot site</td>
</tr>
<tr>
<td>Prepopulated automatically</td>
<td>14.2%</td>
<td>24</td>
</tr>
<tr>
<td>Completed manually</td>
<td>24</td>
<td>48</td>
</tr>
<tr>
<td>Once for all</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>For each report</td>
<td>36.0%</td>
<td>75</td>
</tr>
</tbody>
</table>
We expect that the partial automation of the AE reporting process enabled by SALUS solutions will contribute to reduce AE under-reporting, which is currently one major obstacle in the process of acquisition of pharmacovigilance data. Projects such as ASTER have already demonstrated that prepopulating AE reporting forms using EHR data improves substantially spontaneous reporting practices [13]. However, whether SALUS solutions can really increase the volume and improve the quality of AE data can only be demonstrated based on its real deployment in hospital settings. SALUS project is still on-going. A first prototype of the AE reporting tool fully integrated into SALUS interoperability platform has been produced. The prototype has been deployed on the two pilot sites with simulated EHR data using their respective data models and databases. But the evaluation will be performed in a later phase of the project, including: 1) a testing phase checking the robustness of the pre-population and conversion mechanisms with real (although de-identified) EHR data; 2) a functional and non-functional software quality evaluation phase; and 3) an end user evaluation phase. The methodology to be used in the latter case remains to be decided in more details. End users satisfaction questionnaires might be relevant. But a more robust evaluation method would consist in a longitudinal study enabling to quantify the added value of the tool in terms of number of submitted reports and data quality. Measure of the mean time needed to complete and submit a report with and without the tool could also be compared to determine if the tool actually facilitates AE form filling and minimizes the time required by the reporting process.

Several challenges needing to be addressed for successful implementation of SALUS AE reporting tool have also emerged during the design phase: i) SALUS is a proof of concept project, and a key assumption of the technologies being designed is that structured data is available. This could be considered a strong limitation since this is clearly not the case when considering current state of the art in hospitals. Tests conducted in the pilot sites showed that although some very important data elements such as diagnoses and medication names are available in a structured manner, a substantial amount of data in the EHR is still available only as free text [29]. Structuration of data in EHRs is a well known problem and several solutions are currently examined to mitigate it, such as tools to facilitate the coding of data (e.g. terminology browsers or auto-completion mechanisms) or Natural Language Processing technologies to enable automatic extraction of structured data from free text [30-32]. These solutions are in principle usable in the future to complement SALUS platform. SALUS strategy is also prospective: the assumption is made that providing powerful tools that can only work with structured data will have a motivational impact on healthcare institutes and work practices, boosting structured data input in EHRs.

ii) Sometimes, mappings between E2B data elements and EHR data models can only be partially achieved: some E2B sections are simply not present in EHR data models (e.g. “Seriousness of the AE” or “Recurrence of AE on re-administration” have no corresponding section in CDA-CCD templates) or value sets only partially overlap. We also need to ensure the alignment between terminologies, as E2B requires the use of MedDRA for medical data whereas EHR systems often resort on LOINC, ICD-10 or SNOMED-CT. This is not a straightforward operation, as the terminologies often have various levels of granularity, then mappings can only be approximate. Secondly, terminologies are evolving, therefore mappings need to be regularly updated. In our approach, we address this problem by reusing current available mapping resources (see Section 4). The most important issue at this phase is to utilize reliable mappings. Mappings from PROTECT [28] and OMOP projects are created manually by terminology experts, and whenever a mapping from BioPortal or UMLS metathesaurus is to be introduced into the SALUS Semantic Resource Set, it is checked by the medical experts of the SALUS project. The main limitation of such approach is of course its cost in terms of human resources. Verifying the quality of mappings requires specific expert knowledge and is time-consuming.

The precedent issue is essential because it implies that AE reporting process cannot reach full automation: the physician needs to validate and/or complete pre-populated data, or make a selection among several terminology mapping candidates, before submitting the AE reporting form to the regulatory authorities. Medical judgment must also be exercised to select the relevant patient history items from the EHR. A big challenge for the conception of such tool is consequently to find equilibrium between blind automation and manual expertise based completion of data.

iii) Last but not least, access to hospital CDWs storing EHR data also poses some ethical-legal difficulties that are country-specific. Patient data must generally be de-identified before being accessed and cannot leave the Clinical Care Zone, i.e. the zone where identified data is maintained and accessed locally. In order to ensure ethical-legal regulations, we de-identify and pseudonymize AE reporting forms before being sent to regulatory authorities.

7. Conclusion

Secondary use of EHRs in supporting AE reporting process demands establishing syntactic and semantic interoperability between varying EHR data models and AE reporting standards. We presented a knowledge mediation approach for establishing interoperability between i) EHR data models – such as HL7 CDA-CCD/ PCC templates and AGFA ORBIS data model, and ii) AE reporting standards – such as E2B(R2), by converging and mediating both modalities using SALUS CIM and CDEs. In order to deal with medical terminological differences, we utilize terminology reasoning and mapping services provided by the SALUS platform. Future work includes further development, and deployment and evaluation of SALUS AE reporting tool in the healthcare settings: i)
prepopulating AE reporting forms by extracting data available in the patient EHR, ii) providing assistance to clinicians in further completing and validating the prepopulated forms, and (iii) transmitting the completed and validated forms to regulatory authorities. A prototype is already designed and discussed in [13].

Acknowledgments

The authors express their gratitude to Yves Parès, Romain Ng and Adrien Ugon for their technical contribution to the conception of SALUS AE reporting tool. The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007–2013) under grant agreement no ICT-287800, SALUS Project [18].

References