EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

4 October 2018
EMA/614680/2018

A Common Data Model for Europe? – Why? Which? How?

Workshop report from a meeting held at the European Medicines Agency, London, United Kingdom,

11-12 December 2017

30 Churchill Place ● Canary Wharf ● London E14 5EU ● United Kingdom
Telephone +44 (0)20 3660 6000 **Facsimil** +44 (0)20 3660 5555
Send a question via our website www.ema.europa.eu/contact

An agency of the European Union

Contents

1. Executive Summary

The unparalleled pace of change in the scientific landscape is driving a paradigm shift in drug development, challenging regulatory agencies to look beyond conventional sources of evidence to support decision making across the entire product life cycle. While the regulatory toolbox is expanding with new initiatives and legal tools to support innovation, regulators must ultimately balance the desire to provide access to a potentially lifesaving or life changing medicine with the need to have sufficient confidence in its long term efficacy and safety. As a consequence we need to consider how we can collect adequate evidence to provide this level of confidence. In the future questions will be broader and will likely require an increased scope, depth and detail of data including digitally collected lifestyle data for example from wearables which offer the opportunity to capture a holistic view of the patient.

Real world data (RWD) has the potential to address many of the current regulatory needs but there are concerns around the reliability and validity of the evidence, especially when conducted across multiple countries and databases. For example in the post authorisation setting regulators are commonly faced with multiple separate observational studies, performed over time in different databases which often deliver variable results. Occasionally if a co-ordinated approach is adopted studies may use a commonly agreed protocol, the aim of which is to standardise the conduct of the studies. However experience shows that even the use of a common protocol can still allow substantial variability in the conduct of the study which can increase the heterogeneity of the results in an unknown way. The need to address these concerns increases the challenge in delivering trustworthy evidence in a timely fashion. An alternative approach to improve both reproducibility and speed is to transform and thus standardise data into a common data model (CDM) which allows the use of common analytics and methods across multiple datasets. This meeting sought to explore whether such an approach would be applicable for heterogeneous European data and if so what would be the key design characteristics which would influence the sufficiency of the data to meet regulatory needs.

In the context of the meeting a CDM was defined ‘as a mechanism by which raw data are standardised to a common structure, format and terminology independently from any particular study in order to allow a combined analysis across several databases/datasets’. The meeting was informed by in depth discussions of two relevant CDMs; the Sentinel CDM developed and funded by the United States Food and Drug Administration (FDA) and the Observational Medical Outcomes Partnership (OMOP) CDM supported by a global community of researchers under the Observational Health Data Sciences and Informatics (OHDSI) collaborative network. Both are distributed data systems encompassing hundreds of millions of person years of observation which structurally re-organise the data but take fundamentally different approaches with regard to the extent of vocabulary mapping. In the Sentinel CDM, the amount of vocabulary mapping is restricted because it is felt that decisions on coding should be made in the context of the specific study question. In contrast, the OHDSI network, which utilises the OMOP CDM, incorporates data from 17 different countries representing more than 82 databases and both restructures the data and standardises its content by mapping the multiple different coding systems or vocabularies used in the source databases to a common vocabulary. A perceived advantage of this is that the pre-specification of mappings removes individual decision making which may be variable. Both systems have developed highly customised, re-usable analytical tools to interrogate the CDM which accelerate studies across the multiple data sources.

Each CDM system represents a trade off on where to set the balance between the efforts invested in data management and data analysis. Where the emphasis is placed is likely to have an impact on the speed of studies. If speed is to be maximised, the diversity of the EU setting would require the standardised vocabularies provided by the OMOP model. However it is important that the model remains sufficiently flexible to answer multiple analytical use cases and scalable to multiple data sources. Furthermore, since not all codes can be incorporated into even the most comprehensive CDM,

it is critical that the source data, including unmapped data, is incorporated and retained within the CDM. Ultimately in order to build trust in studies performed with a CDM with extensive mapping, a careful characterisation is needed to determine whether there is loss of information when EU data is transformed into the CDM and to assess the impact on effect estimates.

Irrespective of the data model, data quality must be understood. While both systems incorporate validation processes, the policing of the process differs; FDA requires the Sentinel network to implement a highly regulated, repeatable, systematic process which is consistent over time and data sources while the OHDSI model depends significantly on its community to challenge and police a validation approach supported by multiple software packages. However the OMOP CDM is used by various networks and some have implemented different governance policies for each use.

Learnings from the National Patient-Centered Clinical Research Network (PCORnet) illustrate that active data curation is constantly required with dynamic, heterogeneous source data and engagement with data partners as a critical part of this process. To reduce variability there should be a single, data holder validated CDM version per database which leverages the expertise of the data holder. Ultimately the CDM must operationalise reliability and robustness by building clear and consistent business rules around transformation of data to support regulatory decisions which have immediate public health impact.

Any system across Europe must address the common challenges of data protection and privacy; moreover government institutions may have specific confidentially requirements beyond those stipulated in law. This was not the main focus of the meeting but clearly is a key requirement for any data governance system. The sensitive and personal nature of healthcare data demands robust data protection and the new requirements of the General Data Protection Regulation will need to be considered, especially in the context of complex real world data, which may in the future include digitally captured data from wearables and smart devices, and the intention to develop cross member state solutions.

Any system must also be built with sustainability at its core. It is clear that a European data platform would require investment which must go beyond the initial investment in the data transformation and encompass an ongoing funding model to enable the continual update and validation of these dynamic datasets. Broad uptake of a CDM would encourage sustainability but is dependent upon defining scientific acceptability and delivering utility from the perspective of all relevant stakeholders. Although complex, we need to start to move towards a situation where decisions around the acceptability of evidence derived from RWD are based on a clear framework. Implementation of a CDM would require a decision tree to consider at each level how utilisation of a CDM may influence the results generated.

There are without doubt multiple challenges in implementing a CDM in Europe and we need to understand where limitations would lie across a range of use cases. No one approach is perfect but a CDM could address many of the limitations of the currently available methods. A set of guiding principles which could underpin a CDM are suggested within this report. However it is important that we do not mix the question of whether a CDM is appropriate or necessary with issues of data quality, concerns over the status of a vocabulary which is dynamic and which can be updated or the possibility that analytical tools to add additional bias. While all these issues are important and must be addressed, they are not necessarily unique to a CDM and it could be argued that they are different discussions and should not be used as a reason for not adopting a CDM.

Regulatory decision making needs timely data that is meaningful for benefit-risk assessment, which supports multiple use cases, is representative of a wide population across Europe, is of sufficient quality and size and is generated through a transparent methodology with robust data governance that meets data protection requirements.

2. Background

The use of healthcare data, generated through the delivery of normal clinical care and encompassed by the term real world data (RWD)[1], is increasingly being proposed as a source of evidence to support drug development and regulatory decision-making. Use of such data is not new; it has been utilised for many years to support decision-making post authorisation where multiple sources of evidence, often of mixed quality, must be combined to reach the best decision possible. In the early 2000's this process was accelerated in the wake of high-profiles safety problems, most notably with Vioxx. However another current driver is the rapid pace of change in the scientific landscape resulting in more products which cannot align with the traditional drug development pathway, challenging regulatory agencies to look beyond conventional sources of evidence to support decision making across the entire product life cycle.

Despite the pressing need to address head on such challenges, there is a significant concern that real world data cannot meet the evidentiary standards required to support regulatory decisions particularly on efficacy and effectiveness. Moreover observational studies across multiple databases are challenging, especially in Europe and can take several years to complete and hence cannot currently meet the need for timely and robust evidence generation. This is particularly pertinent when there is an urgent clinical need e.g. in the event of a serious side effect. Several factors contribute to these problems but they could be appreciably reduced by transformation and thus standardisation of data into a common data model (CDM); this allows the use of common analytics and methods across multiple datasets with the aim of improving both efficiency and reproducibility. Despite use of the OMOP CDM in Asia-Pacific and European/US electronic health record (EHRs), the bulk of experience with CDMs lie in US claims based data[2] [3] and it is unknown whether the use of this approach for heterogeneous European data could result in loss of data integrity and an alteration of the semantics.

On 11 and 12 December 2017 EMA hosted a workshop which sought to answer many questions including how to balance flexibility of question with speed of results; how to validate a CDM and understand if and where information may be lost and/or analytical flexibility following data transformation; how to operationalise a network across Europe to build a sustainable framework through which the expertise of all stakeholders could be incorporated; and finally, how to define the key design choices of a CDM which could influence data sufficiency. It is clear that the approach and method will always be driven by the question and the availability of data sources and therefore even if a common data model is used in Europe, study specific approaches will still be necessary. It is also important to highlight that the current debate is building on a significant platform of previous work investigating approaches to optimise multi-database studies, some of which have utilised a CDM approach and some have utilised different approaches. These include study specific or partial CDMs or common protocol approaches which may include some harmonisation of structure (Ref. 1,2,3,4,5,6)[4] Ultimately data may be combined centrally at an individual patient level or aggregated in tabular form for analysis or, alternatively, fully analysed results may be transmitted and then combined centrally.

3. A Common Data Model – Why?

In the context of the meeting a CDM was defined 'as a mechanism by which raw data are standardised to a common structure, format and terminology independently from any particular study in order to

[1] Real World Data are data relating to patient health status or the delivery of health care routinely collected from a variety of sources other than traditional clinical trials
[2] https://www.sentinelinitiative.org/
[3] https://www.ohdsi.org/
[4] http://www.emif.eu/

allow a combined analysis across several databases/datasets. Standardisation of structure and content allows the use of standardised applications, tools and methods across the data to answer a wide range of questions'. This definition therefore excludes scenarios where only a subset of the data is transformed for a specific study.

RWD has been used for many years to support post authorisation decision making for signal detection and risk management, for life cycle benefit-risk evaluation and to assess the impact of regulatory decision-making. The ultimate aim, with the availability of sufficient timely data, would be the creation of an iterative regulatory system to enable the impact of risk minimisation methods on health outcomes to be actively and continually assessed. The natural extension to these safety orientated applications is initially considered to be in: addressing uncertainties around conditional approvals where there is an unmet medical need; understanding the natural history of disease not only in the context or rare diseases but also when a disease has a long latency and the progression is not understood such as non-alcoholic fatty liver disease; identifying appropriate prognostic markers and endpoints for rare conditions and confirmation of surrogate outcomes with long term clinical outcomes particularly in slowly progressing diseases. From a regulatory perspective, challenges to the use of RWD to address efficacy/effectiveness remain the lack of randomisation, the use of non-contemporaneous controls, biases both known and unknown, whether actionable outcome measures are present in the dataset, recording of adverse events and concerns around data quality and representativeness of the data.

The lack of utilisation of RWD outside of the safety environment is reflected by experience in EMA Scientific Advice. Across 600 scientific advice procedures from July 2016 to June 2017, only 3% of procedures sought advice around the use of RWD. Within this limited dataset the majority of questions in the pre-approval setting focused on the use of RWD to provide historical control data, for which advice was typically that, rather than risk uncertain data quality a small randomised controlled trial (RCT), even if underpowered, was preferred to non-contemporaneous control data. The exception to this advice was for an ultra-rare disease in which an RCT was not possible. Where primary data collection was proposed for post authorisation effectiveness evidence generation, concerns were around capturing adverse events and the validity of evidence generated through US data sources when extrapolated to the European setting. Surprisingly, even though most experience rests in the post authorisation safety setting, here proposals were weaker highlighting the value of early engagement with regulatory authorities for all questions.

One of the challenges facing regulators today is how to enable earlier access to medicines for those patients with limited treatment options when uncertainties around the medicines remain. In these cases, a well-developed post authorisation evidence generation plan coupled with well-defined and robust pharmacovigilance activities needs to be in place to quickly address uncertainties. The EMA Patient Registry Initiative was established in 2015 in part to meet this need; it promotes early engagement among key stakeholders to either facilitate the use of existing patient registries or to support the establishment of new registries if none are available or none are of sufficient quality for regulatory decision making (Ref. 7,8). Early engagement enables the arrangements to be in place prior to product launch, accelerating the collection of evidence in a real world environment.

Even when medicines are authorised on the basis of larger numbers of patients, the number studied prior to approval remains relatively small, relative to the number who ultimately receive the medicine and the duration of study is limited. Duijnhoven et al (Ref. 9) reported that for new molecular entities approved between 2000 and 2010, the median number of patients studied before approval for medicines containing a new active substance was 1,708 but only 438 for an orphan medicine. Across 84 medicines intended for chronic use, only 68 (82.1%) and 67 (79.8%) met the guideline recommendations for 6-months and 12-month patient exposure respectively. Consequently

uncertainties always remain at the time of approval including the benefit-risk in wider clinical use especially for high risk populations and the identification of rare adverse drug reactions or those with a long latency. Proactive work to continually update our knowledge of a medicine is thus vitally important.

In this setting, regulators must understand the benefit: risk for any medicine across its product life cycle. Often the need for additional data is urgent, for example in the context of a serious safety signal, and needs may change as new information emerges or the population expands on authorisation from the selected population studied in clinical trials. Ultimately there is a regulatory need for timely data that is meaningful and relevant for benefit-risk assessment, which supports multiple use cases, is representative of populations across Europe, is of sufficient quality and generated through a transparent methodology. However the EMA experience is that significant delays are often encountered in the delivery of European multi-database studies conducted via the common protocol approach i.e. where a generic protocol is adapted locally to different dataset structures due partly to the need to make individual decisions on the definitions for exposures, outcomes, confounders and time windows, partly due to different data structures, languages, terminologies and healthcare systems, and partly due to the time required for data governance practices. Moreover the results may be subject to bias due to subtle differences in the interpretation of the study question at different sites which can increase the heterogeneity of the results in an unknown way. Furthermore, the requirement for multiple programmes to be written is a source of delay. While not all these issues will be solved by a CDM, it is envisaged that the *a priori* transformation of data into a common structure, format and terminology independent of any particular study which enables the use of standardised analytics will minimise the need for individual decision making. Of course, such coding would need to be maintained at a sufficiently high detail to enable adequate discrimination of study concepts. Ultimately it is felt this would promote uniformity and consistency in the analysis both over time and across databases and significantly accelerate the conduct of multi-database studies.

3.1. The FDA Sentinel System: meeting regulatory needs

Delivering high quality evidence sufficient to support regulatory decision making for the Food and Drug Administration (FDA) was a central goal of the US Sentinel System and encapsulated in its foundational needs and goals. Sentinel was first launched in 2008[5] on a background where safety data were previously generated primarily from pre-approval randomised controlled trials or post-approval spontaneous reporting systems. Key legislation mandated the creation of an active risk identification and analysis (ARIA) system to improve post-approval safety monitoring. ARIA is a subcomponent of the Sentinel System comprised of predefined analytic tools and data formatted in the Sentinel CDM. The Sentinel System was developed to deliver high quality evidence and meet a legislative requirement requiring FDA to consider the ability of the Sentinel system to address the regulatory need before requiring a market authorisation holder to conduct a postmarking safety study. Thus the legislation balanced the burden of safety surveillance between FDA and manufacturers. Deciding whether Sentinel is “fit for purpose” requires an understanding of data adequacy across multiple use cases, appropriate methods and the definition of a satisfactory level of precision.

Sentinel introduces the concept that a CDM based system involves more than a simple structural organisation of the data. Additional elements are required to create a functional ecosystem which for the Sentinel system translates into a data quality assurance framework and highly customised, reusable analytical tools designed to work with the Sentinel CDM. In order to produce credible trusted evidence sufficient for use in regulatory decision making at FDA, the Sentinel Data Quality Assurance

[5] https://www.fda.gov/safety/fdassentinelinitiative/ucm2007250.htm

system[6] encompasses over 1200 data checks of the core well defined data elements at 4 different levels at each single data refresh across all 17 data partners. In this way the CDM also acts as change buffer and unifier against structural and IT changes at the datasets driven by mergers, acquisitions and routine business needs. Importantly this framework delivers a trusted and curated dataset on which the analytical tools can run. Inevitably this introduces a baseline cost to the system as data is curated and checked irrespective of whether one query is run or dozens but equally means the data is always "ready to go".

While the Sentinel CDM is an enabler of analytic scale and customisation it was designed with flexibility in mind and hence incorporates minimal mapping in order to deliver a system which *'enables investigators to implement the most appropriate design and analysis plans for given drug outcome pair'*. Thus the Sentinel CDM structure allows a diversity of data sources to participate but importantly retains values with known meaning and does not mix data from different sources. For example dispensing and medication administration data are both descriptors of drug exposure but have different meaning which is retained in the CDM. In this way Sentinel supports dozens of finely customized analyses suited to regulatory needs, rather than thousands of standard analyses which do not allow precisely tailored design choices. Currently between 20 and 30 highly tailored analysis are being run through the Sentinel system every quarter. FDA has ultimate version control over the CDM to ensure regulatory needs are always met.

The governance of the Sentinel system is built upon a voluntary "opt-in" programme in which data partners always approve analyses and results before release to the FDA which provides a clear and well-designed governance framework. In additional, Sentinel operates under the public health authority of FDA; all of its activities are considered public health practice and are not subject to rules governing research. This delivers operational speed for urgent requests. The query development process begins with the development of a mini study protocol at FDA which is then translated into the analysis parameters of the predefined analytical programmes. Each analysis is tested in a test database before distribution to all data partners. This allows FDA to ensure that all analyses achieve the regulatory objectives and correctly measure quantity of interest. The design phase takes time and often requires multiple iterations informed by input from across the Sentinel network. By design this process aims to speed computation time but not planning time.

The sufficiency of ARIA for regulatory questions is thus enabled by data quality (data management curation + change buffer and unifier) which delivers valid inferences (data quality + analytical customization) with analytical speed (validity + analytical scale).

Sentinel cannot answer all questions but many of the areas of weakness relate more to the characteristics of the administrative claims databases themselves rather than to the system. For example high on the FDA wish list would be the ability to capture longer term follow-up to overcome the fragmentation of the US health system where patients move healthcare providers every 2-3 years. Information on disease staging and progression is also a key need e.g. tumour staging, Child-Pugh liver disease prognostic scores, in addition to access to insurance formularies to help characterise treatment decisions.

Ultimately if the only requirement of the CDM were *'scaling of the most rigorous design and analysis plans'* bigger would be better. However we must be cognisant of the fact *'that scaling of inappropriate design and analysis methods will lead to results that are precisely wrong'.*

[6] https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model/sentinel-data-quality-assurance-practices

3.2. Validation of a Common Data Model Approach

The process of generating observational evidence is far from a straightforward, pre-defined journey from source data to actionable evidence. There are many points at which variability may arise; for example there are multiple types of source data (variable source data arising from different care settings and different national influences in care delivery), multiple types of evidence to generate and multiple purposes or stakeholders. Hence standardisation of any part of the pathway provides opportunities to reduce the variability in the evidence generation pathway, increasing trust in its ultimate product.

Common data models have taken different approaches some standardising only the data structure (tables, fields and data types) while others are more extensive and include standardisation of data conventions (how the data should be represented), of data vocabularies (terminologies to codify clinical domains), cohorts (how the clinical phenomenon of interest should be represented), covariate construction (definition of the variables used in the statistical analysis), the analysis itself and finally how the results are reported. All of these criteria should be defined within a protocol but in practice the protocol only provides a superficial summary of each of these elements. As a result study design can lead to dramatic variability in results even when examining the same drug outcome pair in the same dataset over the same timeframe (Ref. 10,11,12).

Understanding how much of this process should be standardised within a CDM and the downstream impact on the study results requires identification of the key attributes influencing reliability and the key needs of the system. The fundamental question is whether for the same research question comparable results are found if the analysis is performed by different researchers, in different countries, in different databases or with any combination of these potential sources (or other sources) of variability. Inter-country reliability is particularly important in Europe where the influence of different health policies and guidelines, different procedures and different cultures results in significant heterogeneity, for example the concepts behind outcome measures and the circumstances under which they are measured are unlikely to be consistent. In such cases, transformation to a CDM should not remove or hide heterogeneity but rather by ensuring a consistent interpretation of concepts across countries enables it to be assessed. Validity refers to the performance of a method (study design or analytical method) in identifying true associations (using positive controls) or absence of associations (using negative controls). Measures of validity could be influenced by database characteristics such as coding systems.

The question is whether a CDM and the use of common analytics may increase transparency, reliability and validity. It could be argued that a CDM, especially when it involves mapping source data to a different terminology, introduces another layer of complexity and hence variability. Nevertheless by breaking the link between data and researchers, the CDM divides the journey into data transformation and the analysis plan which enables each part to be assessed and analysed independently. Thus data transformation, if applied consistently combined with the use of common data analytics, increases transparency. A CDM also enables the re-use of the first segment of the journey, the data transformation, for multiple questions assuming the CDM encompasses sufficient elements to be relevant for multiple questions. It also helps in understanding better the underlying reasons when validity is not met across different research questions. However, interpreting the validity of the CDM should not be considered to reside purely in assessing the data transformation but considerations should also encompass validation of the software e.g. of the ETL process (the software implementing the data transformation) and also validation of the clinical and statistical methods. It is thus important to consider how any validation procedure would address not only the transformation of the data (the first part of the journey) but also the analytics procedures.

Development of different approaches may be driven partly by different analytical use cases but also by demands of the different source data which has resulted in different solutions. Each solution results in different trade-offs in data management and analytical complexity. The common protocol approach represents the highest complexity for the analyst and the least data management responsibilities. In contrast a CDM, especially one incorporating a common structure, conventions and vocabularies, represents the greatest responsibility for data management which must ensure that mappings to a common vocabulary are appropriate and at the right level of granularity. There is no right answer; each represents a trade off as to where to balance activities between data management activities and data analyst activities. It must be remembered however that where the emphasis is placed will have an impact on the speed of studies but not necessarily the quality, with the highest complexity for the analyst reducing the speed of delivery.

It is therefore important to define where we are starting from, where we want to finish and the minimal acceptable reliability across the whole evidence generation pathway and set expectations accordingly. Different solutions will be needed for different scenarios.

4. A Common Data Model Which?

If one were to select a CDM for Europe, it would be important to consider whether the domain of regulatory science really has unique requirements which are not applicable to other domains and warrant unique solutions. Being rooted in the belief that unique, tailored solutions are needed for your domain can result in a failure to consider all viable options. On the syntactic level, multiple solutions are possible but models must be dynamic to evolve to changing needs. The debate normally resides around the semantic and pragmatic features of a model and it is thus critical to define what the expectations of each stakeholder are and who will play what role? Ultimately what is required is a model which will deliver reliability and validity and yet deliver transparent and rapid research at an unprecedented scale across Europe.

4.1. The Sentinel System

The Sentinel CDM is used within a distributed database network which currently provides access to 66.9 million members, 14.4 billion pharmacy dispensings and over 13.3 billion unique medical encounters[7]. The network predominantly includes claims data but also incorporates EHR and registry data when available and the model is extendable to any data source. This is enabled by the fact that data are stored at the most granular level available which provides the flexibility to support any type of analysis demanded by the question.

The governance of the Sentinel framework employs a distributed query approach in which as much of the analysis as possible occurs behind the firewalls at the individual data partner sites. Critically Sentinel partners enter into contract with Harvard Pilgrim Health Care Institute who manage all CDM-related interactions on behalf of the FDA; as such FDA primarily interacts with Harvard Pilgrim although Harvard, FDA and the Sentinel data partners commonly collaborate on CDM (and other) activities. The CDM was designed around the specific needs of the FDA with the involvement of the data partners; despite this there is almost a continual need for iteration and development of the model which is currently on version 6. FDA provides all the funding for Sentinel and therefore has complete control of the Sentinel system; nothing is changed within it without the knowledge and consent of the FDA.

[7] https://www.sentinelinitiative.org/sentinel/data/snapshot-database-statistics

The development of the Sentinel CDM was based on a number of guiding principles which drove its ultimate design. One of the fundamental guiding principles underpinning the model is that nothing is mapped that is not necessary and no derived variables or tables are created with the premise that if you can generate the variables from the information in the model, then generate at execution rather than prepopulate the model. The CDM currently contains 12 tables of data elements[8] with record linkage across tables achieved with a unique person identifier. Furthermore distinct sets of data are kept separate to prevent misinterpretation e.g. prescriptions, dispensing and drug administrations but also because the data refreshes at different frequencies within the source systems and this may cause problems under certain scenarios. Importantly Sentinel captures medications distributed in other settings and adjustments, for example indicating a dispensing was cancelled or not picked up, are processed before table creation.

The model was designed to be as intuitive as possible so that users could easily understand the data structure. Due to the extendable design, adding a new value set to the model is relatively straightforward in part because most data is not mapped.

Sentinel processes queries in a distributed manner across the network. Requests are distributed to the data partners, the query is run locally, the analytical data set with all direct identifiers removed is sent to a central secure server and aggregated and analysed centrally by Harvard Pilgrim (figure 1).

In order to accelerate simple studies, Sentinel has built a toolbox that enables queries[9] of increasing complexity and various functionalities building from simply descriptive to comparative studies e.g. propensity score matching and stratification, self-controlled designs to patient profiles and line lists.

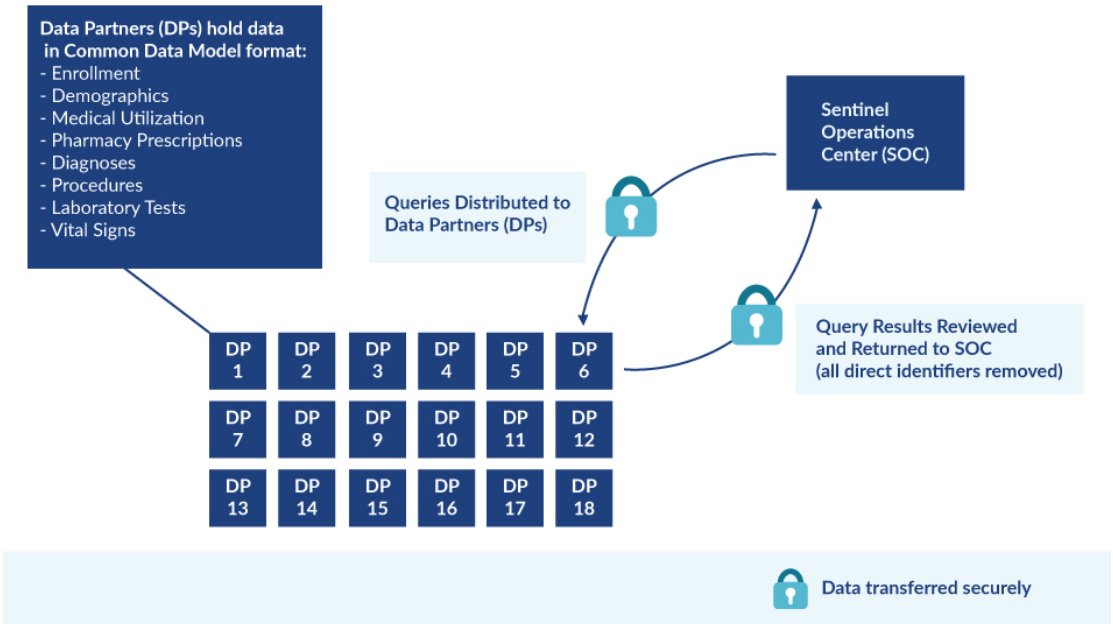

Figure 1: Data flows across the Sentinel Distributed Database[10]

The Sentinel toolbox is now primarily based on the Cohort Identification and Descriptive Analysis (CIDA) tool plus a suite of tools which allows the implementation of any of the available analyses. The analyses which have been done for FDA have been highly customised increasing the complexity of the analytical tools but the data is always used at the most granular level available. Fundamentally the

[8] https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model/sentinel-common-data-model
[9] https://www.sentinelinitiative.org/active-risk-identification-and-analysis-aria
[10] https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model

approach prioritises flexibility to assess individual exposure/outcome relationships allowing customisation for each assessment. For example the granularity of the model allows the FDA to specify the allowable episode gap which is appropriate for each analysis and define medication exposure at the level of the dose, route and formulation. Equally outcome definitions can be defined for each analysis which is considered critical and generally changes for each analysis.

A number of key considerations were emphasised during the meeting: firstly inclusion of a variable does not imply its completeness and moreover the level of completeness may vary by source and over time and affects the usability of the data which cannot be assumed; flexibility of question was a key design choice and the model minimises data mapping to enable this; finally standardisation and model refinements are ongoing actions but any changes are driven by FDA needs.

4.2. The National Patient-Centered Clinical Research Network (PCORnet)

Sentinel predominantly incorporates claims data and the challenge of applying the Sentinel CDM to EHRs was described through a description of the National Patient-Centered Clinical Research Network (PCORnet). PCORnet is a large, highly representative, national patient centred clinical research network[11] encompassing 79 distinct health systems with representation in every US state, with the mission of understanding what works best for patients. The ultimate vision is to enable high quality, efficient large scale clinical research both observational and interventional.

The Sentinel CDM is at the heart of the PCORnet data strategy and, as for Sentinel, PCORnet has a focus on data quality and hence curation. As a result the PCORnet strategy took the Sentinel guiding principles as a starting point and amended where appropriate. One key principle which remained consistent across both data networks was that the CDM reflects values found in the source data; this reduces the burden on the data management team for extensive mappings but also prevent the loss of information and local knowledge as data makes its way through extensive mappings. In establishing the network the need for continual evolution was recognized and that the CDM would change with time and in line with this the PCORnet CDM is currently on Version 4.0 and has added further tables to the Sentinel CDM to meet the needs of the PCORnet data partners[12]. However frequent changes have downstream impact especially for network partners who must update their ETL processes and the analytical tools; a continual deep dialogue with the local data partners is thus essential. Lastly it was recognised that all parts of the CDM would not be populated by all data partners and that the network should be free to add data/domains to their local CDMs.

An iterative data curation cycle is a key part of the PCORnet data strategy and the importance of this process was underscored. Each cycle begins as the data partner refreshes their data, the frequency of which depends upon the relative involvement in ongoing activities, for example involvement in interventional studies/randomised trials often involves the need for monthly data refreshes. As for Sentinel the PCORnet co-ordinating centre provides a data curation package which the centres can download and run, determine where their version deviates from the CDM and following discussions as to whether any deviations are acceptable or not, the refresh is approved. The Coordinating Center is also key in facilitating the interactions between the researchers and the PCORnet partners (Figure 2).

[11] PCORnet consists of 20 Patient-Powered Research Networks (PPRNs), 13 Clinical Data Research Networks (CDRNs), 2 Health Plan Research Networks (HPRNs) and 1 Coordinating Centre. For further information please visit http://www.pcornet.org/

[12] http://pcornet.org/wp-content/uploads/2018/01/PCORnet-Common-Data-Model-v4.0_Specification.pdf

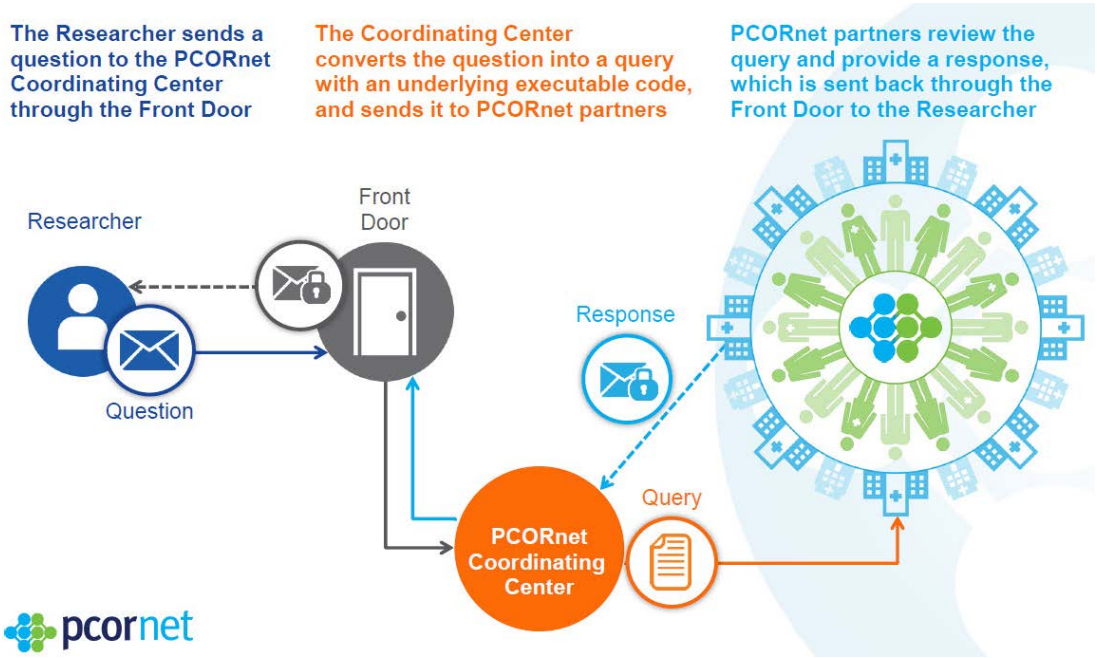

Figure 2: Data flows across the PCORnet Distributed Data network

A number of key learnings have emerged from the PCORnet experience; firstly EHRs are highly heterogeneous and source data are dynamic. Despite the presence of only two main EHR software providers in the US, significant customisation occurs in different institutions such that EHRs are different between different clinics in the same institution. Thus an implementation guidance is required for at a minimum every table within the CDM, if not for every data element which is built up iteratively as questions arise from the data partners. Similarly data partners must provide an annotated data dictionary at each refresh to the data element level describing how each ETL procedure occurred. Active data curation is essential but even 4 years into the cycle missing data still occurs and therefore constant attention is essential. It is the tools and processes which make the CDM usable and useful.

4.3. Observational Medical Outcomes Partnership (OMOP) CDM

An alternative CDM which has been extensively used, predominantly in the US, is the Observational Medical Outcomes Partnership (OMOP) CDM. The network which underpins the OMOP CDM is called OHDSI (Observational Health Data Sciences and Informatics), a community which now incorporates over 200 researchers across 17 different countries representing more than 82 databases and approximate 1.2 billion health records. More recently there has been a growing interest in the OMOP CDM in Europe reflected in the opening of a European chapter of the OHDSI network[13] and the launch of an Innovative Medicines Initiative (IMI) project to fund transformation of a large number of European datasets into the OMOP CDM[14]. As for Sentinel and PCORnet, the OMOP CDM standardises different structures across disparate data sources into common tables which harmonise structure, field datatypes and conventions. This should not result in information loss as it simply structures the data differently. However a key difference between OMOP and Sentinel/PCORnet is that additionally OMOP seeks to standardise the content by mapping the multiple different coding systems or vocabularies

[13] http://www.ohdsi-europe.org/
[14] http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/imi2-2017-12-04.html

used across the source databases to a common vocabulary for example ICD9/10 to SNOMED-CT[15]. The thinking behind this is that if common terms such as cardiovascular disease, bleeding, drugs etc are standardised semantically across databases this would significantly enhance the speed of observational research. The clinical data is stored in the model under standard concept IDs which refer to the standardised vocabulary e.g. SNOMED-CT in the case of conditions. Terms are also assigned a descriptive name e.g. atrial fibrillation and a domain to which the concept belongs. However the model also stores the original source concept IDs in the clinical table and the verbatim source code as found in the source record. This enables analysis using the standardised vocabulary but also allows analysis through the source data, although adopting the latter approach would reduce the speed for multi-database studies. The OMOP CDM is also part of an ecosystem which has common data analytics sitting above the CDM (figure 3). No patient level data is shared in the process of completing a query, only aggregate summary statistics.

OMOP has developed a model that accommodates both administrative claims and EHR data from both private and public payers internationally, across care settings and additionally supports registries and longitudinal surveys across a broad range of multiple use cases. In contrast to Sentinel, OMOP contains derived elements e.g. drug period but this does not prevent definition of different periods if required through the tools. The model is extendable and evolving over time driven by the community needs of OHDSI members.

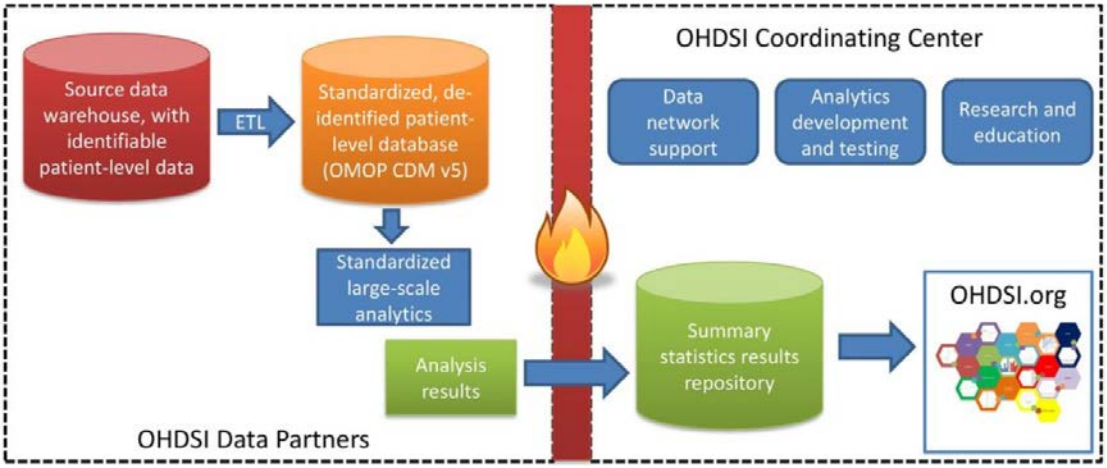

Figure 3: Data flows across the OHDSI network

Given that the extent of vocabulary mapping is a key difference among the model, time was taken in the workshop to describe the basis for this approach. The OMOP vocabulary represents standards compiled from disparate public and private sources largely built upon the platform of the National Library of Medicines Unified Medical Language System (UMLS). It is constantly evolving in line with the continued evolution in the sources themselves and with the OHDSI community input. The variability across data sources is illustrated by the fact that there are currently 78 vocabularies across 32 domains encompassing 5,720,848 concepts and 32,612,650 concept relationships within one large concept table in order to allow the standardisation of queries. While this variability describes the challenge of standardisation, it equally highlights the need. Tools are provided to enable vocabulary

[15] SNOMED CT is the most comprehensive and precise clinical health terminology product in the world, owned and distributed around the world by SNOMED International and has been developed collaboratively to ensure it meets the diverse needs and expectations of clinicians worldwide and is now accepted as a common global language for health terms. SNOMED CT also works to provide explicit links (cross maps) to health-related classifications and coding schemes in use around the world, e.g., ICD-11, ICD-10, ICD-O-3, and Global Medical Device Nomenclature (GMDN).

browsing; for example ATLAS[16] which additionally provides insight into concept relations and the creation and storage of transparent and reproducible and potentially re-usable concept sets. ATLAS also supports the building of complex cohorts composed of multiple components such as conditions, drugs, procedures, measurements, observations and visits and queries can be developed against either the standardised vocabularies or the source concepts. Within these defined cohorts, ATLAS provides other functionalities such as clinical characterisation (descriptive summaries, incidence rate estimation), population level effect estimation (including comparative cohort designs using propensity score matching), and patient level prediction. Critically it is still possible to access the source code and data if wished; this information is retained within the CDM allowing any analysis to be performed on the original data within the CDM.

Mapping to a single vocabulary allows the exploitation of the hierarchy of that vocabulary. For example, there are multiple drug vocabularies across Europe with almost every country having their national code; OMOP maps all of these individual codes onto the RxNorm standard drug dictionary, a granular coding system which includes information about active ingredients, brand names, strength and formulation. However this dictionary is based on the US market and needs extension to include all European medicines which is currently being undertaken by the OHDSI community to make it possible to search across all drugs and databases with a single vocabulary. The Anatomical Therapeutic Chemical (ATC) vocabulary has been commonly used to map local source codes across Europe but with this classification system information on strength and formulation is lost, far from ideal for many regulatory use cases. Recognising this, the regulatory network is implementing the ISO IDMP standards for the identification of medicinal products in regulatory submissions[17] which will allow regulatory questions to be addressed in the same language in which they are phrased.

4.4. Challenges of Implementing a CDM in Europe

Europe is fortunate with its national healthcare systems which provide longitudinal 'cradle to grave' care and in some members states have provided a wealth of data for research. However there is significant heterogeneity across these data sources arising from multiple coding systems, languages, structures, content and governances which complicate the implementation of a CDM across European data. Several approaches have been employed to date to facilitate multi-database studies including common protocol approaches, study specific CDM approaches, disease focused approaches, approaches which incorporate a central pooling of data in a data warehouse and more recently projects have started exploring the use of the OMOP CDM. However, as yet none of these initiatives have delivered a sustainable pan European data platform capable of addressing multiple regulatory use cases and of routinely generating evidence in timely fashion.

It is important that the core principles of pharmacoepidemiology are respected when performing multi-database studies whether using a CDM or not. There are potentially unique biases which could arise with the use of a CDM and which likely differ in magnitude across the Sentinel and OMOP CDMs. Such biases can be broadly divided into biases originating from the source data itself, the data transformation or the analysis of the studies. For example it could be argued that the OMOP CDM, as a result of its multiple mappings, would be more at risk of information bias, due to potential misclassification of outcomes and exposures during the mapping process. This may be exaggerated if there was a different granularity between the source data and the standard CDM vocabulary, if source codes were not available in the CDM or if it was impossible to map from free text fields to the standard vocabularies. If such misclassification is non-differential and independent of exposure, it would lead to

[16] http://ohdsi.org/web/ATLAS
[17] http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000645.jsp&mid=WC0b01ac058078fbe2

a relative risk estimate biased towards the null. A study by Ruigómez et al (Ref. 13) which demonstrated that a broad definition (definite/probable cases) of acute liver injury overestimated cases differently across two European primary care databases illustrates the potential implications of applying the same mapping across different sources of information.

In the context of regulatory decision making exposure misclassification would be of particular concern; while previous publications (Ref. 14, 15), have suggested that some European drugs cannot be completely mapped, the unmapped products appear to be predominantly OTC products. Additionally as discussed in the previous section it is acknowledged that the current standard OMOP vocabulary for drugs RxNorm, is US based and need expansion for the European setting to standardise preservation of strength and formulation. Nevertheless complex exposure definitions will require the flexibility to adapt to specific study questions or databases; such flexibility is achievable with both CDM's but in the OMOP CDM may require the utilisation of the original source data.

A potential loss of or change in information on mapping to a CDM remains a concern especially where such mapping is extensive; in the context of confounding, an incomplete mapping of confounding factors will result in an unknown amount of residual confounding, the impact of which needs to be understood. Conceptually, the impact of confounder misclassification on the accuracy of the results would depend both on the strength of the association between the confounder and the outcome and the confounder and the exposure and also the type of effect to be studied e.g. unpredictable Type B[18] vs predictable Type A[19] adverse drug reactions or intended effects. However confounding misclassification is not unique to a CDM and would be also be a factor when performing multi-database studies or when performing individual participant meta-analyses; it is simply important not to assume that because the data in a CDM is structured similarly, inherent heterogeneity is removed. Statistical approaches such as multilevel multiple imputations to manage systematic missing information or heterogeneity across the databases could also be considered (Ref. 16).

What is critical with multi-database studies, irrespective of whether conducted with a CDM or not, is the transparent reporting of methodology in line with recent guidelines (http://www.record-statement.org) (Ref. 17) in order to be able to replicate and reproduce observational studies. To build trust in studies performed with the CDM, a careful characterisation is needed to determine whether there is loss of information when EU data is transformed into the CDM and assess the impact on effect estimates. Without such information the question will remain as to whether a comprehensively mapped CDM such as the OMOP CDM is the best approach for European data or whether the approach of restricted mapping such as employed by Sentinel which can be enhanced with study specific variables when required is more appropriate. Even with the most comprehensive CDM, the possibility must always be retained to implement study specific solutions if they are needed.

[18] Type B: Idiosyncratic, unpredictable, acute / sub-acute, not related to known mechanism

[19] Type A: reactions are more common and are generally attributable to known pharmacological or toxic effects of the drug.

Sentinel

CDM	Network Ecosystem
• Distributed data network with data queries run locally • US based • Predominantly US claims data, a small percentage linked to EHRs and a recently incorporated hospital EHR system • Source data retained • Built upon the principle of minimal mapping and no derived values • Strict version control • Extendable	• Centralised co-ordinating centre • Access previously restricted to FDA but other stakeholders can work with data partners independently to use data and tools • Data linkage to other settings possible • Limited patient follow up due to switching between insurance providers • Multiple mandatory validation levels and audit steps • Managed quarterly data refreshes • Standard analytics available

PCORnet

CDM	Network Ecosystem
• US based • Predominantly electronic medical records • Modified Sentinel CDM • Follows principle of minimal mapping • Strict version control • Flexibility for individual data partner to add data/domains to local CDMs	• Centralised co-ordinating centre • Distributed data network with data queries run locally • Multiple mandatory validation and audit steps • Refresh rate variable and dependent upon involvement in ongoing studies • Continual evolution of the model requiring detailed implementation guidelines • Standard analytics available

Observational Medical Outcomes Partnership (OMOP) CDM

CDM	Network Ecosystem
• Broad, comprehensive model to incorporate claims data, EHRs and surveys • Substantial mapping of content and concepts to standardise multiple different coding systems. • Source data retained • Extendable	• Distributed data network with data queries run locally • Global community network of users in 17 different countries • Upgrading to new versions of the CDM is not mandatory • Multiple automated validation checks but implementation may vary across databases and networks • Standard data analytics available

Table 1: Summary of the main characteristics of each CDM

4.5. Lessons learnt from concrete case examples

Over recent years a number of European databases, predominantly electronic health record databases, have been transformed into the OMOP CDM which has delivered a number of useful learnings.

The first step of an ETL process is the structural mapping which is the process of importing all of the source data *without any mapping* into the OMOP CDM; the source data is then retained within the CDM. Structural mapping of the UK THIN database into the OMOP CDM resulted in the loss of a number of prescriptions which was found to relate to prescriptions allocated before birth and after death. Similarly work done in 2013 to convert the THIN database revealed a number of issues including, the incomplete coverage of medicines within RxNorm[20] (Ref. 15). These authors produced a

[20] RxNorm has now been extended; the only medicines now not mapped into the OMOP CDM included some herbal medicines, homeopathic products and over the counter vitamin products and products such as nappies and glucose sticks. Similarly of the 109,000 diagnostic codes, only 13 remained unmapped.

heat map to illustrate the efficiency of the process which if implemented routinely might be helpful in enabling the end user to decide the completeness of the conversion and how much missing data is present. The take home message is that the success of the structural mapping process should be confirmed and if, and when, any data is lost the underlying reasons needs to be determined. However loss of information/detail is not necessarily problematic, in fact it may be appropriate, but needs to be understood.

The second step is the vocabulary mapping and again the apparent loss of or change in information that would accompany conducting an analysis using only standardise vocabularies must be explored; in mapping the Dutch IPCI database only about 50% of the drug terms could be mapped to terms in the OMOP CDM but this represented 95% of the prescriptions. Terms which could not be mapped included specials items such as nappies. Again it could be argued that this "loss of information" would have very little downstream impact in terms of the usefulness of the transformed database for observational research and ongoing extensions of the standardized vocabularies may resolve such issues in the future.

Relevant to multi-stakeholder access to CDM frameworks, is a pilot performed with the Innovation in Medical Evidence Development and Surveillance (IMEDS) programme[21] in collaboration with the Sentinel network. The aims of the pilot were to: to develop and test process and policies around access to Sentinel by non FDA stakeholders; to perform two test queries through the IMEDS/Sentinel distributed data network to evaluate the association between oral contraceptives and venous thromboembolism; and to assess the effectiveness of a label change for proton pump inhibitors. The work demonstrated that such policies and procedures could be developed with support from all relevant stakeholders including the data partners but highlighted the importance of clarity around such procedures when different entities access data networks.

Overall work to date has highlighted a number of potential complications performed on a platform of datasets which themselves change over time, both structurally and in content with additional fields: lack of version control of the CDM where the same database can be converted into different CDMs or different versions of the same CDM or the same version of the CDM by different groups exist; and different analytical tools and different versions of the same tools producing disparate results against the same CDM. Not only do such issues add to the difficulty of resolving complexities between disparate results across pharmacoepidemiological studies but they also reduce the credibility of the field and reduce trust in the results.

In an effort to understand potential sources of variability between CDMs, the Humana dataset was transformed into both the OMOP CDM and the Mini-Sentinel CDM and a single safety query was run through both models using the relevant analytical programmes (Ref. 18). A deliberate decision was taken to use the relevant purpose built analytical tool set so as to test the whole CDM ecosystem as most users would experience it rather than try and standardise tools across the CDMs. The study focused on 6 drug-outcome pairs assessed though 2 different analytical methods (high dimensional propensity score based procedure and self-controlled case control study). Either no or minimal information loss was observed during the transformation of the databases into each CDM but in some circumstances differences (sometimes great) were seen between the cohort creations and the relative risk scores. Differences were fundamentally driven by lack of a priori transparency around different implementations which were clear on more detailed analysis. However a system delivering a routine rapid analysis capability for timely evidence generation could not include such further analysis. Acknowledging there were a number of limitations associated with the study, it however illustrates nicely that the availability of a CDM does not in itself replace the need for local understanding and

[21] http://reaganudall.org/innovation-medical-evidence-development-and-surveillance

expertise and there is a risk that queries run "blindly" and rapidly across multiple unknown databases may produce inaccurate results. This is particularly important as it will reduce trust in CDM outputs for stakeholders who do not have expertise in the particular CDM. Nevertheless if it is to deliver rapid cycle analyses, network CDM outputs need to be readily understandable and replicable.

The studies and pilots listed above in addition to work done through the IMI funded project EMIF which has mapped 10 databases to the OMOP CDM, have resulted in a number of lessons learnt which include:

- the importance of a multidisciplinary team for the data transformation to blend the local expertise on how data is recorded and stored in the local coding systems, with expertise on the OMOP CDM and the analytical tools;
- strong project management;
- sufficient resource investment for the vocabulary mappings;
- the need for version control of CDM;
- transparency around the whole CDM ecosystem including accompanying tools and versioning over time; and
- the need for training of stakeholders in the OMOP CDM and OHDSI tools.

5. Validation of a CDM – what is needed for regulatory decision making?

'loss of fidelity begins with the movement of data from the doctor's brain to the medical record'

Clem McDonald, MD
Director, Lister Hill Center for Biomedical Informatics
National Library of Medicine, USA

While being cognisant of the truth in the above statement, a key concern from the regulatory perspective is how can evidence derived from such data be validated so that the associations arising from it are as near the truth as possible. This concern would be present even if the data were being used in its raw form but it could be argued that transformation of the data to a CDM adds another layer of complexity.

5.1. Validation through the Sentinel network

Validation can be defined as the 'action of checking or proving the accuracy of something' but the specific objectives of the analytical procedure need to be clearly understood as this will determine the characteristics which need to be evaluated. When considering validation of a CDM it must be understood that there are fundamental differences between study specific validation approaches such as would occur in a common protocol study and network data validation approaches such as through the Sentinel network. In common protocol studies data validation occurs as needed at the time of the study and is completely focused on understanding the accuracy of the data required to answer the specific question which in some cases, may require accessing medical records to check whether values have been entered appropriately. Hence in this case the burden is on the study team and the associated cost is included within the cost of a study. In contrast within a network utilising a CDM a validation of the data transformation must be done a priori such that the all the data is "ready on demand" for all questions; validation checks are therefore needed to confirm that the data has been

restructured appropriately and that, where applicable, any mappings to a common dictionary accurately reflects the source data. Implicit within this is a need to understand what is an acceptable level of accuracy. Such validation does not involve checking the accuracy of the source data itself, merely that the data has been incorporated accurately within the CDM. In this case the burden is on the network team and the associated cost for 1 study is the same for 1000. Moreover because the question is unknown the validation must be repeatable, systematic and consistent over time. Such a system avoids the costs and delays of individual projects utilising the CDM devoting significant resources to validation of the data transformation but equally the process validates data that may never be used. Moreover it does not remove the need for data validation for individual studies.

Sentinel employs over 1200 checks which each site must meet in order to be validated. The purpose of the data quality assurance (QA) processes is to assess whether the data contained within Sentinel meets reasonable standards for consistency and quality of data transformation, including reviewing data integrity across data tables as well as characterising data trends and patterns. All data partners are sent a comprehensive QA package which checks the data partners ETL in waiting, touching on every row in every table and delivers back a compliance check at two levels: Level 1 which assesses completeness and content of each variable in each table; level 2 which assesses cross variables and cross table integrity. Thus if a variable such as 'admission' is used frequently it will have multiple checks. At another level there is a Judgement Call Check which assesses both trends and consistency over time and on another level, logic, plausibility and convergence. As the underlying sources are dynamic in both content and structure, at each refresh there is a full overwrite of the data rather than adding onto the new data. Validation must therefore occur at each refresh to provide confidence that any changes are real and not a result of structural changes. Even several years down the line problems still emerge from data partners within Sentinel. For example in recent data deliveries from the 5 largest sites, 24 checks were reported in QA that required follow up; importantly 22 of these were in the judgement call category requiring expertise and knowledge illustrating that not all the process can be automated. This is enabled by a structure where a trusted third party is checking and comparing the quality checks and where checks are not done in isolation.

Critical to the Sentinel system is the creation of a knowledge management structure to record and track all issues and in particular how they were resolved for future reference. The system also records all codes being used across the network, investigates the uptake of new codes and looks for instances of incomplete data capture over time and across data partners, information which subsequently informs on the creation of algorithms for new studies. The data quality checks must change as the model evolves.

Finally it is important to note that while ETL data curation QA checks are done a priori to yield an approved dataset for each refresh, every individual Sentinel query also includes a set of data characterisation checks focused on the specific question. For example when coding system changes e.g. ICD9 to ICD10 and queries need to be run across the ICD9/10 divide, it is important to validate whether the ICD9 and ICD10 based definitions generate comparable cohorts i.e. whether the ICD9 to ICD10 mappings perform as expected. In Europe one may imagine that it would be important to check whether a definition applied similarly across two or more different countries generated comparable findings.

It should be remembered that the Sentinel network represents a unique environment in that Sentinel is a funded contract which requires data partners to follow specific QA processes. In the event of problems or discrepancies in the Sentinel situation, resolution is linked to payment whereas a network of the willing relies on the data holder to voluntarily resolve issues.

5.2. Validation through the OHDSI network

Data validation within OHDSI is approached from the perspective of the whole evidence generation pathway rather than just the data itself and incorporates 4 components that frame the validation process: data validation, software validation, clinical validation and methods validation. However there are important distinctions between the OHDSI and Sentinel approaches in that OHDSI enables validation across the open community network by the provision of tools, workgroups and collaborations platform but does not enforce specific validation processes across the OHDSI community. Hence while the expectation is that groups doing collaborative research would enforce the processes in order to ensure consistency, this process is voluntary. Nevertheless, this does not preclude individual data networks implementing mandatory validation checks within their governance frameworks; this has already been seen through a number of initiatives including the clinical data research network, PEDSnet[22] (Ref. 19).

The validation process seeks to answer a number of questions:

- How do we ensure preservation of source data into a CDM?
- How do we ensure ETL conventions are followed?
- How do we ensure vocabulary mappings are correct?
- How do we detect inconsistencies in the underlying data?

The set of checks on data format and structure occurs naturally because the OMOP CDM exists in a structure with constraints across tables and therefore, if the data is not entered in the right way, the transformation is automatically rejected as part of the ETL process. In addition the OHDSI community provides specific tools for the ETL process (figure 4):

- WHITE RABBIT; profiles the source data and highlights patterns in the source values e.g. variability, frequency etc but is not specific to any particular format
- RABBIT IN A HAT: provides a consistent mechanism for documenting the ETL process for each dataset to the CDM which can be shared among the community.
- USAGI: supports vocabulary mapping which considers all of the concept codes in the original source data and determines the percentage that are not mapped. Hence this exposes the quality and density of fully mapped versus partially mapped versus unmapped content.
- ACHILLES: a data characterisation, quality and CDM conformation package which performs checks on every domain and every concept within the transformed database. OHDSI recommends running ACHILLES following ETL and all subsequent data refreshes.
- ACHILLES HEEL: summarises the data checks in a standardised way reporting on demographics, data inconsistencies, data drop off, data outside observation periods and makes checks for certain types of data e.g. missing data. This encompasses a number of the judgement calls included within Sentinel's level 3 checks. Achilles Heel also provides conformance checks and provides visualisations of identified anomalies; additionally issues may also emerge from the community but this does not provide a systematic check.

[22] https://www.fbo.gov/index?s=opportunity&mode=form&id=02025cd0fe5ae8f7594baeca61f58545&tab=core&_cview=1

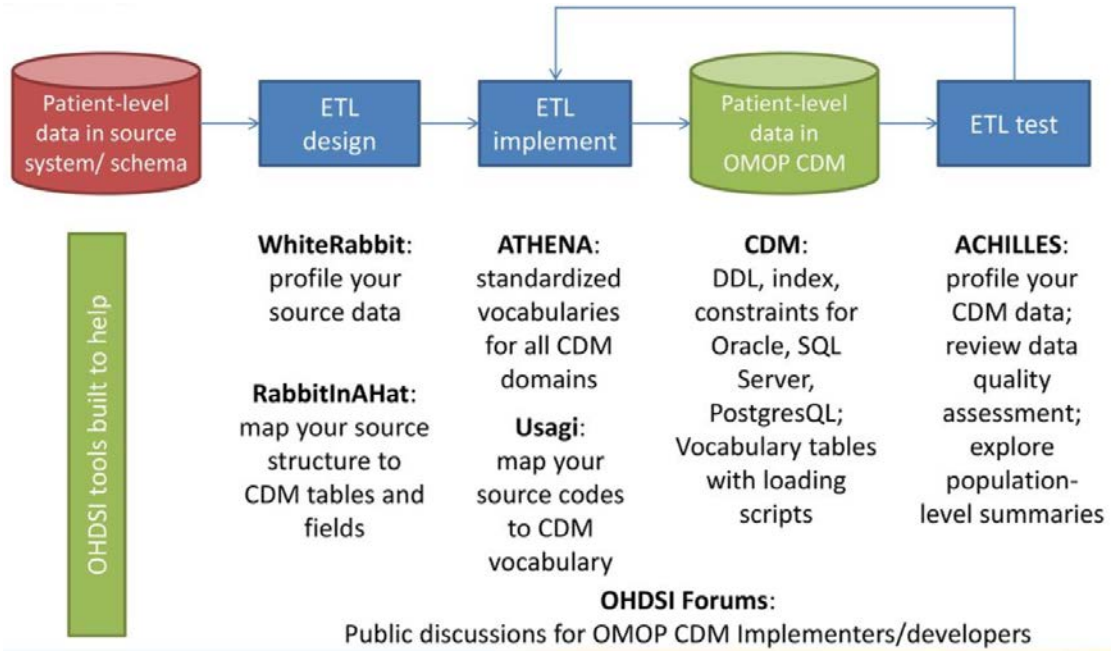

Figure 4: Analytical tools for the ETL process

In addition to these tools there is a working group called THEMIS which is responsible for defining specific conventions or business rules within the ETL e.g. how to handle multiple birth dates, conflicting genders, events after death etc. Importantly, as mentioned earlier any group of sites can enforce a common set of validations/business rules to ensure consistent interpretation of the CDM within their collaborative group.

Validation of the vocabulary mapping would be a key concern for a European framework given the number of different coding systems in use across Europe. To date the majority of OMOP vocabulary content is developed outside of the OHDSI community and used without modification; when manual mappings are developed there is a second level of review by independent coding manager, quality and change evaluation scripts are run prior to each release and there is OHDSI community review. The OHDSI community have explored the influence of using the standard terminology versus the source code (Ref. 20).

Across 27 drug-outcome pairs when applying a standardised analytic method across two databases coded in ICD-9-CM, SNOMED-CT, and MedDRA, the relative risks were essentially identical.

In the OMOP CDM, software validation is considered equally important as an integral part of the CDM ecosystem. As such the OHDSI community has created a number of method packages with all updates are transparently recorded within ATLAS[23]. However community input is again key to identifying issues.

Clinical validation can be described as the extent to which the analysis conducted matches the clinical question or intention. To achieve this, the cohorts must be explicit, well described and communicated to allow collaborators to review prior to cohort generation and subsequently must be able to be validated in a standardised and consistent fashion. Within OHDSI, communication is achieved via an intuitive user interface tool within ATLAS which allows the development of extremely complex cohort definitions and moreover produces a readable visual document which allows validation of the cohort. Moreover once generated, OHDSI provides tools for statistical and clinical review in a standardised

[23] https://docs.google.com/document/d/165L3XSwEkxC6eHYF0wCPtjQ4QrvUbqd_dauFgEq9Zj4/edit

fashion; however this again depends upon a voluntary community based review. In order to expedite validation across multiple sites it is possible to annotate a cohort with a custom set of questions which can then be shared with collaborators. This enables automatic calculation of cohort performance measures to generate cross site performance metrics.

Although outside of the scope of the meeting, the final area of validation discussed was the methods validation which describes, for example, a number of methods diagnostic checks, whether the planned statistical method is valid given the data and empirical calibration for example are there systematic errors that require calibration of the results? It also performs the process of empirical calibration which uses negative controls to look at the likelihood of systematic errors in order to calibrate p values. This novel process is a standard feature of the OHDSI methods pack to ensure the results are valid and calibrated appropriately to the individual datasets. Recent results have demonstrated that many negative drug outcome pairs actually return significant P values across multiple datasets (Ref.21).

The fundamental principle underlying the OHDSI approach to validation is that every aspect of the validation approach is improved dramatically by transparency, openness to challenge and engagement of a community who are committed to improving the research process.

5.3. The CNODES network – Lessons learnt from a Sentinel CDM pilot

The CNODES network is a national network of linked provincial administrative health data which includes 7 Canadian sites but also leverages data from the UK Clinical Patient Research Datalink (CPRD) and this in total provides data on over 100 million people. However the complexity of the current CNODES process requires a project team to be formed for each study in addition to the development and implementation of an individual protocols and a statistical analysis plan. As such in a recent study investigating the relationship between high doses statins and the onset of diabetes, the timeline for the study from identification of the query to presentation of the results was almost 2 years (Ref. 22).

While it is clear that the individualised approach provides advantages around analytical flexibility, allows leverage of data holders expertise and methodological freedom, this comes at the expense of speed which is particularly marked at some sites and significant data variability across the sites in terms of exposure and outcome definitions.

The Sentinel pilot was launched in April 2017; at the time of the meeting table conversions were almost complete which had only required some minor amendments. For example data fields not present in the Canadian data need to be completed for the query tools and field digit lengths e.g. ICD codes require standardisation. The choice of the Sentinel CDM was driven partly by existing strong relationships among academics and regulators and the similarity of Canadian data to Sentinel data in terms of the raw data tables. However the existing robust data quality assurance framework and tools designed to meet regulatory needs and the support provided by the Sentinel team was a large factor in the choice. The principle of maintaining data granularity, as fine as possible for as long as possible, the minimal mapping of the data which removes the need for a common vocabulary and the ability to extend the model to other data sources were all additional important factors. CNODES intends to run validation studies utilising the both the CDM and the CNODES standard tools but also with independent validation by the Canadian Institute for Health Information.

In summary CNODES believes the CDM will facilitate rapid responses to simple queries from Health Canada; while significant upfront investment and time is required to establish the model, ultimately it should significantly improve and accelerate data extraction and analysis. Moreover the advantages of implementing the same CDM as FDA will enable cross jurisdiction collaborations and is particularly valuable given the longer average follow up in Canadian data compared with US data. However the

CDM will not eliminate the need for CNODES standard tools and full epidemiological studies for the most complex questions.

5.4. Acceptability of RWD for Regulatory decision making

Evidentiary standards required for regulatory decisions have to date been largely based on the rules around the design and implementation of clinical trials but these do not translate well into observational studies. However we should not be starting from the premise that observational data can answer the same questions as clinical trials but rather ask what questions observational data can answer and indeed for what questions may such data be preferable to clinical trial data. Necessary standards will vary depending on the context under which the question is asked but there is a need to be transparent around these standards to provide not only clear criteria for industry but also to provide clear assessment criteria for regulators to drive consistency and equity around decision making. Critically the public need to understand and trust the process.

Under perfect conditions the power of the clinical trial lies in several areas: firstly it allows the pre-specification of the clinical variable to be measured in terms of the measurement criteria, the timing of the measurement relative to the treatment allocation and the interpretation of that measurement. In the observational setting the variable must be inferred from a range of unsystematically recorded observations where the timing is not controlled relative to the prescribing decision and where there is variation across healthcare professionals in their understanding, in the use of the preferred code and the propensity to record.

Secondly, in most interventional studies the study group is usually highly selected to achieve an unbiased estimate of efficacy; however it is the selectivity and lack of representativeness of an RCT population which calls into question the generalisability of the results of RCTs.

Thirdly most RCT study designs allow for the option of a placebo control and the process of randomisation removes selection bias. This therefore provides for a very controlled environment although the knowledge that both clinical staff and patients are involved in a clinical trial may result in a modification of behaviour in unknown ways. The most significant disadvantage of observational studies is that treatment is given selectively according to perceived patient need which results in significant selection bias and it is not always clear that the prescribed treatment has been taken and for how long.

Fourthly data validation in an RCT is optimal with standardised forms, trial management procedures to check timing and completeness of data, major errors and omissions are queried immediately, monitoring is mandated and external inspections may be implemented. For observational data, some data collection systems may provide mechanisms for checks against medical records but this will be very variable and only related to a subset of the data. Statistical checks may be run within observational studies but remedial measures are usually crude and because studies are not specified at the time of data collection, concurrent validation cannot be focused.

Lastly success criteria and analysis are pre specified for RCTs and much can be done to control for multiplicity. While in observational studies best practice would also require a formal protocol requiring pre specification without prior looks at the data, the difficulty from a regulatory perspective is confirming that this was done and adhered to. The inability to control the data collection adds complexity. Decisions based on results always require post hoc assessment of credible bias but it is not possible to determine the success of any adjustment.

As a result of these limitations, apparently large effect sizes and small P values arising from observational studies do not alone ensure trust. Even with a large effect size many aspects of data quality and study design need to be assessed.

The discussion above introduces the concept that principles borrowed from the clinical trial world could be repurposed to help build trust in observational results. As a useful starting point, principles such as transparent pre-specification of selected data sources on open source metrics, availability of coding details including hierarchical systems and the pre-definition of concepts, possibly pre-validated data systems repeated at regular intervals and automated record of analysis (including versions of CDM and dataset used) would promote transparency and facilitate replication of the analytical procedure. From a regulatory perspective, a carefully designed CDM may provide an environment that not only limits some of the potential sources of bias associated with observational studies but by facilitating replication in a timely fashion allows the key attributes of replicability and generalisability to be addressed.

The cost of establishing distributed data systems and in particular the cost of data transformation is significant. The total annual cost of Sentinel including all of the infrastructure, governance, querying and data holder support is approximately $15 million. The OHDSI network estimated that transformation of a database requires dedicated time of 4-6 months while CNODES mentioned annual costs ran at around $3.5 million Canadian which mostly represented costs for the studies themselves and did not reimburse academic or clinical time.

Robust governance of any network will be critical to its success. Consent processes differ not only across European member states but also across data partners within member states and can be complex and time consuming. However a distributed data network is designed so as to involve the movement of the minimal amount of data and the tools generate a consistent data structure and output which not only allows rapid confirmation of the data by the data holders but builds familiarity of the processes within ethics committees when their views must be sought. Queries should be phrased such that the information is at the highest level of aggregation consistent with the scientific requirement in order to reduce the risk of patient re-identification. A common understanding is needed across datasets on the rules for data access.

The mechanism by which studies would be initiated across a European network requires discussion. In the CNODES network, a two-step process allows centres to opt out of a study due to lack of interest but this option is rarely exercised due mainly to the interaction with CNODES principal investigators at each of the sites, a distinct advantage of a pre-formed network of engaged investigators; however opt out does occur on the basis of feasibility. Within the OHDSI network, projects are initiated by individual investigators who develop a protocol within their own datasets, iterate and check through one or two additional sites before posting it on GitHub[24] for consideration by other partners. Within Sentinel to date, all studies are initiated by FDA but are developed and managed by Harvard Pilgrim in conjunction with both FDA and the data partners. More recently the IMEDS programme provides access to the Sentinel network and tools for other parties and additionally allows the ability for chart review to address specific issues. In order to address a regulatory question of which an open posting of the question and/or a voluntary response may be inappropriate, a closed network of contracted sites able to answer the question could be developed from within the entire OHDSI network.

[24] All OHDSI services are hosted on Github platform which also serves as a repository for developing study packages (https://github.com/OHDSI/)

6. A CDM in Europe: potential solutions

Four separate breakout sessions sought to address in more detail potential solutions for some of the challenges of implementing a CDM in Europe, many of which had been articulated in earlier sessions.

- *What are the specific European barriers and challenges in applying a CDM?*
- *How you operationalise a CDM in Europe?*
- *What principles should underpin validation of a CDM in Europe?*
- *What were the key design choices of a CDM which would influence the range of regulatory questions that could be addressed?*

This section records the key findings from those sessions.

The first question addressed in a breakout session *'What are the specific European Barriers and Challenges in applying a CDM?'* is a key consideration given the heterogeneity of the European data landscape. First and foremost it was emphasised again that the CDM, including its structure, governance and processes must be sustainable and that to achieve this it must be underpinned by a sustainable funding source. Most initiatives to date have been funded by short term funding mechanisms (5 years or less), an approach which to date has delivered no lasting data platforms and limited immediate outputs to support regulatory decision making (Ref. 23). Linked to this point, the fact that FDA has a legislative requirement requiring them to consider the sufficiency of Sentinel before requiring a postmarking study has driven the development of a sustainable system; hence it was argued that the long term success of a European network would be dependent upon a single legal requirement imposed upon relevant EU bodies which must include the EMA who are likely to the key recipient of much of the data. Governance of the network was also highlighted as a key challenge given the heterogeneity in access mechanisms and national legislation across Europe; the importance of involving data partners and key stakeholders in the development of the governance cannot be over emphasised. Any network should build harmonised consent processes which fully comply with European data protection legislation but meet the need for timely access to data. However we are not starting from zero and any network should build on established codes of conduct, such as those agreed by ENCePP[25] and ADVANCE (Ref. 24), to minimise duplication of effort and build structures which facilitate close interaction with the data partners to not only leverage their expertise on the data but also to build trust in the network. Trust must be built at many levels including in the data itself, in the transformation of the data into the CDM, in the associated audit and quality assurance processes, in the analytic processes which sit on top of the CDM and in the governance process. Transparency and documentation of processes and ultimately publication of all results whether positive or not will build confidence.

On a more technical level, the CDM must operationalise reliability and robustness by building clear and consistent business rules around transformation of data. The presence of multiple languages and coding terminologies demands a unique European solution which should as much as possible leverage existing European terminologies including ISO SPOR. Since not all codes may be incorporated into a CDM, it is critical that all source data, including unmapped data, is incorporated and retained within the CDM. To aid in a fuller capture of the richness of the EHR data, the CDM should ultimately aim to incorporate free text into the model.

[25] http://www.encepp.eu/code_of_conduct/

All the successful CDMs described in the meeting incorporated an iterative approach into their development, building scalable systems which learnt from the issues and problems which arose over time. The situation will not be different in Europe. One area of scalability is in the scope of use, another relates to geographical cover and types of datasets. In terms of the scope of use cases, we need to build an understanding of the extent of uncertainty in results arising from the CDM starting first with core use cases such as drug safety and drug utilisation and then extending towards more methodologically challenging cases such as drug effectiveness and relative effectiveness studies. In this regard we must also understand and incorporate the needs of all stakeholders including HTA bodies. However it must also be acknowledged that significant variability can arise from common protocol studies as a result of differences in interpretation and implementation of the same protocol across multiple sites.

The second group discussed more specifically on '*How you operationalise a CDM in Europe'*. The issue of funding was raised again highlighting the absence of sustainable funding and challenging the decision and policy makers to suggest how such a system could be funded in Europe including establishing the basic infrastructure, maintaining and running the system and finally performing the studies. Not only this, it must be remembered that initial investment is required for the transformation of the data into the CDM; a recent IMI call in Europe[26] will provide funding for the transformation of a large number of datasets in the OMOP CDM model mechanisms but how the continual maintenance of the CDM including routine data updates would be sustained is unclear. Additionally the role of industry in the infrastructure needs to be considered; in neither Sentinel nor CNODES are industry actively involved in the infrastructure and if a different model is envisaged for Europe with a partnership with industry through IMI, any implications of this need to be carefully considered. The development of the IMEDS programme by the FDA, which provides access for industry to Sentinel, could provide a useful model for Europe.

An important aspect of any successful network is securing the active engagement of the data holders which requires development of a clear value proposition for the database holders. Database partners need to be re-assured that participation in a CDM does not equate to a loss of control of either the data or their role in participating in studies, that their data would be used appropriately and additionally need to agree with and accept the principles of a CDM. To ensure consistency there should be a single version of the CDM per database which is controlled and verified by the data partner with tracking of any supplementary conversions. The Sentinel example emphatically emphasised the importance of the data partners and a key part of its success is the leadership shown in engaging with data holders and securing their trust by developing a system which defines a clear role for the data partners as scientific partners not just as data providers. Additional enabling factors were the wish to improve public health in line with the mission of many of the data partners and clarity around the work load, the processes and security around the data to ensure institutional review board clearance for public health.

Two alternative governance structures were presented at the workshop which showed what may be possible in Europe ranging from the tightly organised and controlled structure of Sentinel through to the open community of OHDSI. Where on this continuum might a European CDM sit? What would drive the repetitive cycles of learning, refinement and expansion within a network? How would the needs of different stakeholders be prioritised and how can this be achieved in Europe?

A common feature across both systems is the incorporation of the element of transparency with open protocols, open reports and open tools and this is a key condition for delivering scientific independence and trust in the system. When considering purely the regulatory use case, what would be the

[26] http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/imi2-2017-12-04.html

organisational and decision making structure and would the open community structure of OHDSI be appropriate in this context? One possibility is that regulatory questions may be answered through the development of a specific network which contracted a subset of the OHDSI partners who agreed to adhere to a number of business and procedural rules. This is in line with current industry practice where, despite holding much data in house in the OMOP CDM, the majority of regulatory required studies are outsourced to provide independence and increase the likelihood of acceptability from a regulatory perspective. This key point was picked up in discussions where it was emphasised that if regulators could articulate where studies performed in the OMOP CDM may be acceptable, reluctance to utilise the CDM within industry would reduce considerably. While regulatory acceptability will always vary depending on the specific use case, some general principles may be defined.

The third group addressed *'What principles should underpin validation of a CDM in Europe?'* with a focus on data validation as opposed to software validation. As a first level validation process it was proposed that all variables should be checked but only with regard to presence or not, which mirrors the compliance checks used by Sentinel. Over time a level of prioritisation could be developed as undoubtedly validation of some variables would need to be performed more critically. However some data will always be more challenging to harmonise and thus validate and laboratory values were given as one key example. Cognisant of the need to understand variability between the measurement of datasets, rates or proportions may sometimes be preferred to absolute counts. Whichever validation mechanisms are used, the openness, transparency and contestability of the specifications will be an important part of any CDM. It should also be acknowledged that this would be a major step forward as such specifications are not currently a feature of epidemiological research. Moreover it would allow pre-defined checks to be done upfront, creating a 'state of readiness' such that individual studies could be run in a more efficient and timely manner.

Considering the empirical calibration approach used as part of the methods validation by OHDSI, the development of a reference set of known effects in order to assay the sensitivity and specificity of the system, could provide a powerful calibration tool. However recent studies (Ref. 25,26) have highlighted the challenges around this and it may be that for example there are possibly very few drug-event pairs that could be agreed on. Additionally if such event pairs could be developed they would likely reflect the validity of the dataset itself rather than the CDM.

What was not discussed in any depth was whether there would be a need in a European CDM framework for a Co-ordinating Centre, as created by Harvard Pilgrim for the Sentinel network. The Sentinel Co-ordinating Centre (SCC) develops analyses based on FDA safety questions, co-ordinates with the data partners, reviews and aggregates results and sends the summary results to FDA. However the SCC is also responsible for validating the transformation of datasets into the CDM and the subsequent maintenance. If there were no equivalent European Centre overseeing a European network, how would the validation checks be policed? Could this responsibility be assigned to a data partner(s) on a rotating basis? Would the community eyes approach of the OHDSI be sufficiently robust from a regulatory perspective?

The final group addressed '*What were the key design choices of a CDM which would influence the range of regulatory questions that could be addressed?'* The scope of the CDM will be influenced by the end users of the system and by the datasets that are included, neither of which are currently defined. However whatever system is adopted, it should be extendable. In this regard an important consideration is not only current needs but also future anticipated needs and what we would like to capture, an obvious key example being patient reported outcomes. Could the CDM design if sufficiently innovative, enable and be a driver for this wider data capture?

As nicely illustrated by the comparison of the Sentinel and OMOP CDMs, the choice of a CDM may require a compromise between speed and versatility. Both CDMs restructure the data to a standardised

format to allow the utilisation of common analytics across the data partners. They differ however in the extent of terminology mapping. The minimal mapping of the Sentinel CDM requires the question to be specified within each coding terminology which, if there were multiple terminologies, would come at the expense of speed but does not affect the discrimination of concepts as specified in the source data. In contrast the OMOP CDM maps all the clinical terminologies to SNOMED which in the context of multiple European terminologies significantly accelerates studies. However if SNOMED was not able to discriminate the concepts required for the study as compared to the source dataset terminology, this may alter the extracted data. Importantly in both systems the source data are available for customised studies when required and the analytical tools which sit above the CDM should allow the introduction of analytical flexibility if required. Importantly in itself the need to digress from the standard choices provided by the mapping provides valuable information and a requirement of the researcher to justify this choice would drive more transparency in methodological choices.

Lastly if different CDMs are ultimately adopted by international regulatory authorities they should ideally be as interoperable as possible in order to allow interrogation of global databases in the event of a rare reaction, exposure or disease. Where convergence cannot be achieved, initiatives such as the Biomedical Research Integrated Domain Group (BRIDG) model[27] which has the objective of developing data interchange standards and technology solutions will help to enable semantic interoperability across datasets. Nevertheless while it could be argued that creating multiple parallel approaches may defeat the objective of global harmonisation, it could also be argued that different approaches may be complementary in that they will have different strengths and weaknesses. In this context in the event that different CDMs delivered disconcordant results to the same question, it would be important to understand how each CDM influences the interpretability of results. Certainly one would need to be reassured that any disagreement did not originate from the CDM itself but was a genuine reflection of the source data.

7. Discussion

In the face of a rapidly changing scientific landscape which is driving a paradigm shift in drug development, the regulatory environment must keep pace. An increasing number of innovative medicines face challenges to align with a traditional drug development pathway which can then result in uncertainties in the data package at authorisation. For example medicines have already been authorised in cases of unmet medical need, with only Phase I data[28] or with uncontrolled Phase II data[29] which result in increased post approval obligations and often the need for long term data capture to address any uncertainties on safety and effectiveness. While the regulatory toolbox is expanding with support initiatives such as PRIME and legal tools which allow flexibility such as accelerated assessment, compassionate use and conditional authorisations, regulators must still ultimately balance the desire to provide access to a potentially lifesaving or life changing medicine with the need to have sufficient confidence in its long term efficacy and safety. In the future questions will be broader and will likely require an increased scope of data and length of collection, extending from primary care into secondary care, and to digitally collected data for example from wearables. To better support decision making there is a regulatory need for timely data that is meaningful and relevant for benefit-risk assessment, which supports multiple use cases, is representative of a wide population across Europe, is of sufficient quality and is generated through a transparent methodology with robust data governance.

[27] https://bridgmodel.nci.nih.gov/
[28] http://www.ema.europa.eu/ema/index.jsp?curl=/pages/medicines/human/medicines/003854/human_med_001985.jsp&mid=WC0b01ac058001d124
[29] http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/human/medicines/002801/human_med_002016.jsp&mid=WC0b01ac058001d124

Real world data has the potential to address many of these needs but delivering evidence in a timely fashion is challenged by concerns around the reliability and validity of that evidence especially when derived across multiple data sources and countries. A CDM offers an opportunity to address some of these issues; by providing a common structure and in some cases common semantics together with a transparent description of how these elements were derived from the source data, it allows standardisation of some parts of the evidence generation pathway. Principles such as transparent pre-specification of selected data sources based on open source metrics, availability of coding details, the pre-definition of concepts and automated record of analysis (including versions of CDM and dataset used) would promote transparency and facilitate replication of the analytical procedure. A validation system applied consistently across data sources, which enabled their characterisation, would deliver an understanding of whether that data source was suitable to answer the question at hand. Thus from a regulatory perspective, a CDM may provide an environment that not only limits some of the potential sources of bias associated with observational studies but by facilitating replication in a timely fashion allows the key attributes of reliability and validity to be addressed.

The meeting was informed by in depth discussions of two CDMs; the Sentinel CDM and the OMOP CDM which both restructure the source data but which take different approaches in mapping of data to common terminologies. In developing the Sentinel CDM a conscious decision was made to restrict the amount of vocabulary mappings to enable the maximum flexibility in protocol design. However this is on the background that coding within American claims databases upon which Sentinel relies, is generally restricted to a limited set of coding terminologies which in itself delivers a degree of harmonisation. In the light of the diversity of the EU setting and number of terminologies standardised vocabularies as provided by the OMOP model are considered an essential component if timeliness is an important characteristic. However, the acceleration in delivery of studies is predicated on the basis that every mapping from source terminology to SNOMED is trusted and does not need validation at a study specific level.

Irrespective of the fundamental approach, any CDM must operationalise reliability and validity by building clear and consistent business rules around transformation of data. The validation approach of the Sentinel network delivers a highly regulated, repeatable, systematic and consistent process over time and data sources. In contrast the OHDSI model depends significantly on the ‘eyes of the community’ to challenge and police its largely automated validation approach. In a regulatory context, a dependence on a voluntary process is unlikely to deliver sufficient reassurance across very disparate data sources and it is envisaged that the imposition of a mandatory and more formalised process would be required. As such a subset of sites could agree to enforce a common set of validations and business rules to ensure consistent interpretation of the CDM within their collaborative group.

In order to build trust in studies performed with a CDM with extensive mapping, a careful characterisation is needed to determine whether firstly whether there is loss of information when EU data is transformed into the CDM and secondly to assess any impact on effect estimates. Without such reassurance the question will remain as to whether a more fully mapped CDM such as the OMOP CDM is the best approach for European data or whether the approach of minimal mapping such as employed by Sentinel which can be enhanced with study specific variables when required is more appropriate and achievable. Irrespective of the model a key take home message is that the possibility must always be retained to implement study specific solutions via the original source data if needed; this is already possible in both models.

Importantly the CDM cannot be considered in isolation, there is a need for a trusted CDM infrastructure which leverages the expertise of the data partner and allows for ongoing development in iterative cycles. To reduce variability there should be a single, data partner validated CDM version per database cut. Embedding transparency into the platform is a key condition for delivering scientific independence

and trust and lastly any system must be built with sustainability at its core. It is clear that a European data platform will require investment which must go beyond the initial investment in the data transformation and encompass ongoing funding to enable the continual update and validation of these dynamic datasets. This is particularly relevant for recent projects such as the European Health Data and Evidence Network (EHDEN)[30] where initial funding is focused on data transformation and mechanisms to ensure the long term sustainability must be found.

Key to the uptake of a CDM is defining scientific acceptability and enabling benefits from the perspective of the relevant stakeholders. However many of the factors which impact on acceptability are not unique to data generated through a CDM but to the use of observational data in general. It is challenging to generalise and be definitive on where an observational study would be acceptable to support regulatory decisions as use cases and questions change across the product life cycle as does the audience receiving the data. Nevertheless defining acceptability is a critical step as any model must be based on a reasonable expectation of future utilisation and benefits to stakeholders. As such we need to start to move towards a situation where decisions are based on a clear framework or decision tree. Within such a framework there should be general principles which need to be considered independent of any question posed which include the regulatory setting in which the data will be used, the feasibility of capturing other data and indeed the delay which may be imposed by generating the other data, the unmet need and the methodology for controlling confounding and bias. In terms of study specific questions, we need to ask whether we can contextualise the resultant effect size in terms of the uncertainties which will exist around it. This is done routinely in the regulatory setting but formalisation of the process may be helpful. Considerations would include what is the level of understanding of disease progression and characteristics, what is already known about the benefit-risk in terms of understanding whether an effect size would be discernable, what precise biomarkers of disease are likely to be recorded in the data set, what is the ability to record exposure dose, duration and adherence and finally whether there is an actionable endpoint on which to base a regulatory decision. Implementation of a CDM would add an additional step to this process as a decision tree would need to consider whether the benefits of using this approach outweigh the perceived risks; would it add additional bias or bring additional confidence?

Any system across Europe must address the common challenges of privacy and data governance; this was not a focus of discussions which rather concentrated on the technical requirements for a CDM but clearly is a key feature of the governance system especially with implementation of new data protection legislation in Europe (General Data Protection Regulation (GDPR)). The sensitive and personal nature of healthcare data demands robust data protection and the new requirements of GDPR will need to be considered, especially in the context of complex real world data including digitally captured data from wearables and smart devices and the intention to develop cross member state solutions. Nevertheless distributed data networks where the analysis is brought to the data with no need to share patient identificable data could provide the most likely mechanism to meet these requirements.

Despite rich European datasets which benefit from national healthcare systems which provide access for all and which often contain lifelong data on individuals, Europe currently has no pan European data network and is lagging behind other regions in delivering answers for healthcare related regulatory questions. This is not necessarily a result of a lack of investment; recent assessments (Ref. 23) have estimated that initiatives linked to RWE have benefited from over 734 million Euros of public funding. However while this has generated a lot of tools and learnings the immediate utilisation of their outputs to support regulatory decision making to date is limited.

30 https://www.imi.europa.eu/sites/default/files/archive/uploads/documents/Future_Topics/IndicativeTopic_EHDN.pdf

If Europe implements a different CDM to other regulatory authorities such as FDA and Health Canada, efforts should be invested to maintain as much alignment as possible. It should however be noted that FDA are also exploring utilisation of the OMOP CDM via the BEST initiative[31]. However heterogeneity of approach is not necessarily to be discouraged as it provides for different solutions and these maybe leveraged to add value and knowledge.

There are definitely multiple challenges with regard to implementing a CDM in Europe which must be flexible enough to answer multiple analytical use cases, scalable to multiple data sources and realistically feasible. Transparency of approach would support the development of an understanding of where limitations in the evidence derived from such a system would lie across multiple regulatory use cases. However it is important that care is taken not to mix questions of whether to move to a CDM is appropriate or necessary with issues of data quality, concerns over the status of a vocabulary which is dynamic and which can be updated, the ability of analytical tools to add additional bias or governance of the data network. While all these issues are important and must be addressed, it could be argued there they are different discussions and should not be used as a reason for not adopting a CDM.

8. Conclusions

Current thinking is that a hybrid system for generation of evidence will always be required to meet the needs of regulatory decision making. No one system can answer the multitude of questions across the product life cycle and the need for complex epidemiological studies delivered across multiple databases via the common protocol approach will always remain as will specific studies in a single data source. However, currently the regulatory experience is that multi-database studies even when performed with a common protocol are usually slow and can still allow substantial variability in the conduct of the study which can increase the heterogeneity of the results in an unknown way. In addition delays are imposed by the lack of a common governance structure across multiple national databases. Implementation of a CDM has the potential to significantly accelerate studies by delivering a system 'primed and ready for use' and improve reproducibility by standardising parts of the evidence generation pathway. Investment would be required which must go beyond the initial investment in the data transformation and encompass ongoing funding to enable the continual update and validation of these dynamic datasets. To prevent fragmented solutions and duplication of effort and deliver a sustainable solution, we must strive for agreement as a community as to the best method to meet varying needs.

9. Guiding Principles

From the presentations and discussion at the workshop, a number of fundamental guiding principles emerged which are detailed below.

Structure

- A common data model can be defined as a mechanism by which the raw data are standardised to a common structure, format and terminology independently from any particular study in order to allow a combined analysis across several databases/datasets.
- A common data model should not be considered independently of its ecosystem which incorporates standardised applications, tools and methods and a governance structure.
- The ability to access source data should be retained.

[31] https://www.fbo.gov/index?s=opportunity&mode=form&id=58ae31bd5d84cb5bf93bdc5891635418&tab=core&_cview=1

- The CDM should be the simplest that achieves security, validity and data sufficiency.
- The CDM should be intuitive and easy to understand
- The CDM should enable rapid answers to urgent questions when required, be efficient and feasible

Operation / Governance

- The CDM governance model must respect data privacy obligations across all data partners and regions
- The CDM should be built with sustainability as a priority.
- Development of the CDM should maximally utilise data partners expertise. The CDM must be agreed by and accepted by the participating data partners.
- There must be version control.
- The CDM should be dynamic, extendable and learn from experience.
- The value package should be clear to data partners.

Quality of Evidence Generation

- The CDM must operationalise reliability and validity by building clear and consistent business rules around transformation of data across multiple databases. Where divergence is unavoidable this should be recorded.
- The focus should be on data characterisation to understand if the data is fit for purpose.
- The CDM should be transparent on how data is defined, how it is measured and incorporate and document its corresponding validation.
- The CDM should allow transparency and reproducibility of data, tools, study design to facilitate credible and robust evidence across multiple datasets.

Utility

- The CDM should provide a common set of baseline concepts which should enable flexibility when required and meets the needs of potential users.
- All the concepts that are commonly used in safety and effectiveness studies should be mapped to the CDM to maximise regulatory utility.
- The CDM should address recognised use cases for which an established need is present.

10. References

1. Coloma, P. M. et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiol. Drug Saf. 20, 1–11 (2011).
2. Trifirò, G. et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? J. Intern. Med. 275, 551–561 (2014).
3. Bollaerts, K., De Smedt, T., Donegan, K., Titievsky, L. & Bauchau, V. Benefit-Risk Monitoring of Vaccines Using an Interactive Dashboard: A Methodological Proposal from the ADVANCE Project. Drug Saf. 41, 775–786 (2018).
4. Eurosurveillance editorial team. ECDC in collaboration with the VAESCO consortium to develop a complementary tool for vaccine safety monitoring in Europe. Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull. 14, (2009).
5. Mor, A. et al. Antibiotic use varies substantially among adults: a cross-national study from five European Countries in the ARITMO project. Infection 43, 453–472 (2015).
6. Myocardial infarction and individual nonsteroidal anti-inflammatory drugs meta-analysis of observational studies - Varas-Lorenzo - 2013 - Pharmacoepidemiology and Drug Safety - Wiley Online Library. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/pds.3437. (Accessed: 31st July 2018)
7. Bouvy, J. C. et al. Registries in European post-marketing surveillance: a retrospective analysis of centrally approved products, 2005-2013. Pharmacoepidemiol. Drug Saf. 26, 1442–1450 (2017).
8. Jonker, C. J., Kwa, M. S. G., van den Berg, H. M., Hoes, A. W. & Mol, P. G. M. Drug Registries and Approval of Drugs: Promises, Placebo, or a Real Success? Clin. Ther. 40, 768–773 (2018).
9. Duijnhoven, R. G. et al. Number of Patients Studied Prior to Approval of New Medicines: A Database Analysis. PLOS Med. 10, e1001407 (2013).
10. Cardwell, C. R., Abnet, C. C., Cantwell, M. M. & Murray, L. J. Exposure to oral bisphosphonates and risk of esophageal cancer. JAMA 304, 657–663 (2010).
11. Green, J. et al. Oral bisphosphonates and risk of cancer of oesophagus, stomach, and colorectum: case-control analysis within a UK primary care cohort. BMJ 341, c4444 (2010).
12. Dixon, W. G. & Solomon, D. H. Bisphosphonates and esophageal cancer--a pathway through the confusion. Nat. Rev. Rheumatol. 7, 369–372 (2011).
13. Ruigómez, A. et al. Ascertainment of acute liver injury in two European primary care databases. Eur. J. Clin. Pharmacol. 70, 1227–1235 (2014).
14. Matcho, A., Ryan, P., Fife, D. & Reich, C. Fidelity Assessment of a Clinical Practice Research Datalink Conversion to the OMOP Common Data Model. *Drug Saf.* **37,** 945–959 (2014)
15. Zhou, X. et al. An evaluation of the THIN database in the OMOP Common Data Model for active drug safety surveillance. Drug Saf. 36, 119–134 (2013).
16. Jolani, S., Debray, T. P. A., Koffijberg, H., van Buuren, S. & Moons, K. G. M. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. Stat. Med. 34, 1841–1863 (2015).
17. Wang, S. V. et al. Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1.0. Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res. 20, 1009–1022 (2017).
18. Xu, Y. et al. A Comparative Assessment of Observational Medical Outcomes Partnership and Mini-Sentinel Common Data Models and Analytics: Implications for Active Drug Safety Surveillance. Drug Saf. 38, 749–765 (2015).
19. Khare, R. et al. A longitudinal analysis of data quality in a large pediatric data research network. J. Am. Med. Inform. Assoc. JAMIA 24, 1072–1079 (2017).
20. Reich, C., Ryan, P. B., Stang, P. E. & Rocca, M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. J. Biomed. Inform. 45, 689–696 (2012).
21. Madigan, D., Ryan, P. B. & Schuemie, M. Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies. Ther. Adv. Drug Saf. 4, 53–62 (2013).
22. Dormuth, C. R. et al. Higher potency statins and the risk of new diabetes: multicentre, observational study of administrative databases. BMJ 348, g3244 (2014).
23. Plueschke, K., McGettigan, P., Pacurariu, A., Kurz, X. & Cave, A. EU-funded initiatives for real world evidence: descriptive analysis of their characteristics and relevance for regulatory decision-making. BMJ Open 8, e021864 (2018).
24. Kurz, X. et al. The ADVANCE Code of Conduct for collaborative vaccine studies. Vaccine 35, 1844–1855 (2017).

25. Hauben, M., Aronson, J. K. & Ferner, R. E. Evidence of Misclassification of Drug-Event Associations Classified as Gold Standard 'Negative Controls' by the Observational Medical Outcomes Partnership (OMOP). Drug Saf. 39, 421–432 (2016).
26. Slattery, J. Measuring Signal Detection Performance: Can We Trust Negative Controls and Do We Need Them? Drug Saf. 39, 371–373 (2016).