SYNCHROS

WORK-PACKAGE 2
Repository of the European/ International main initiatives that harmonised and integrated cohorts D2.3

SYNergies for Cohorts in Health: integrating the Role of all Stakeholders

Grant Agreement No. 825884

Start Date: 01/01/2019

Duration: 36 months

DOCUMENT INFORMATION

Authors	Albert Sanchez-Niubó
Contributors	Teresa Torres
Reviewer	Ellen Vorstenbosch
Responsible Partner	PSSJD
Dissemination Level	Public
Nature	Report
Keywords	Repository, Harmonisation, Integration, Methods, Methodology, Infrastructures
Due Date	31st of August 2020
Actual Submission Date	31st of August 2020
Version	1.0

TABLE OF CONTENTS

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

EXECUTIVE SUMMARY

The SYNCHROS repository has been created to share all key information about initiatives that harmonised and/or integrated cohort data. The repository contains information obtained from 4 different exercises:

1) The mapping of existing initiatives that harmonised/integrated population, patient and clinical trial cohort studies
2) The mapping of existing cohort studies assessing the impact of new exposures on health
3) The mapping of existing population and patient cohorts using social media and new communication technologies
4) The identification of the main European/ international efforts to harmonise and integrate cohorts

The mappings 1 to 3 have been described elsewhere (Deliverable 1.3, 1.4 and 4.2, respectively), hence here we will focus on the main European / international efforts to harmonise and integrate cohorts that have been identified, and subsequently have been included in the repository.

This report contains a description of the SYNCHROS repository of main European and international initiatives, and explains how data have been harmonised and integrated across different cohorts, either population, patient or clinical trial data.

The document contains the following information:

1. Description of the methodological information on harmonisation and integration collected from the identified initiatives.
2. The state of the repository and some preliminary results of the methodological information found.
3. Limitations and conclusions based on the experience of searching for the methodological information of the different initiatives.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

1. INTRODUCTION

The main objective of this report is to provide an overview of the collected methodological information from main European and international initiatives that have harmonised and integrated population, patient and clinical trial cohort data (or are currently in the process of doing so).

The process of identifying and collecting the initiatives, explained in Deliverable 1.3, was based on three different mappings: 1) the initiatives that integrate population cohorts, 2) the initiatives that integrate patients' and clinical trials cohorts, and 3) the initiatives that integrate existing population cohorts, patient cohorts, and clinical trials assessing the impact of new exposures on health. As the main focus of the first and second mapping (population cohorts, patients and clinical trials) has been on initiatives that have harmonised and integrated cohort data, or plan to do so, we will build on these in this report. The same mappings already included key information/variables to collect the information related to the harmonisation and integration methodology reported in Deliverable 2.1.

The initiatives and, specifically, the methodological information are being incorporated into the SYNCHROS repository (https://repository.synchros.eu/). More details about the technical functioning of the repository have been presented in Deliverable 1.2 (submitted March 2020).

2. METHODOLOGICAL INFORMATION ON HARMONISATION AND INTEGRATION

To collect the information, a template was constructed on a spreadsheet with fields to be filled in. The collection of methodological information was transversal between the initiatives, carried out first by one researcher and then reviewed by two other researchers.

The fields of the methodological information collection that were included can be seen in the following figure 1.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

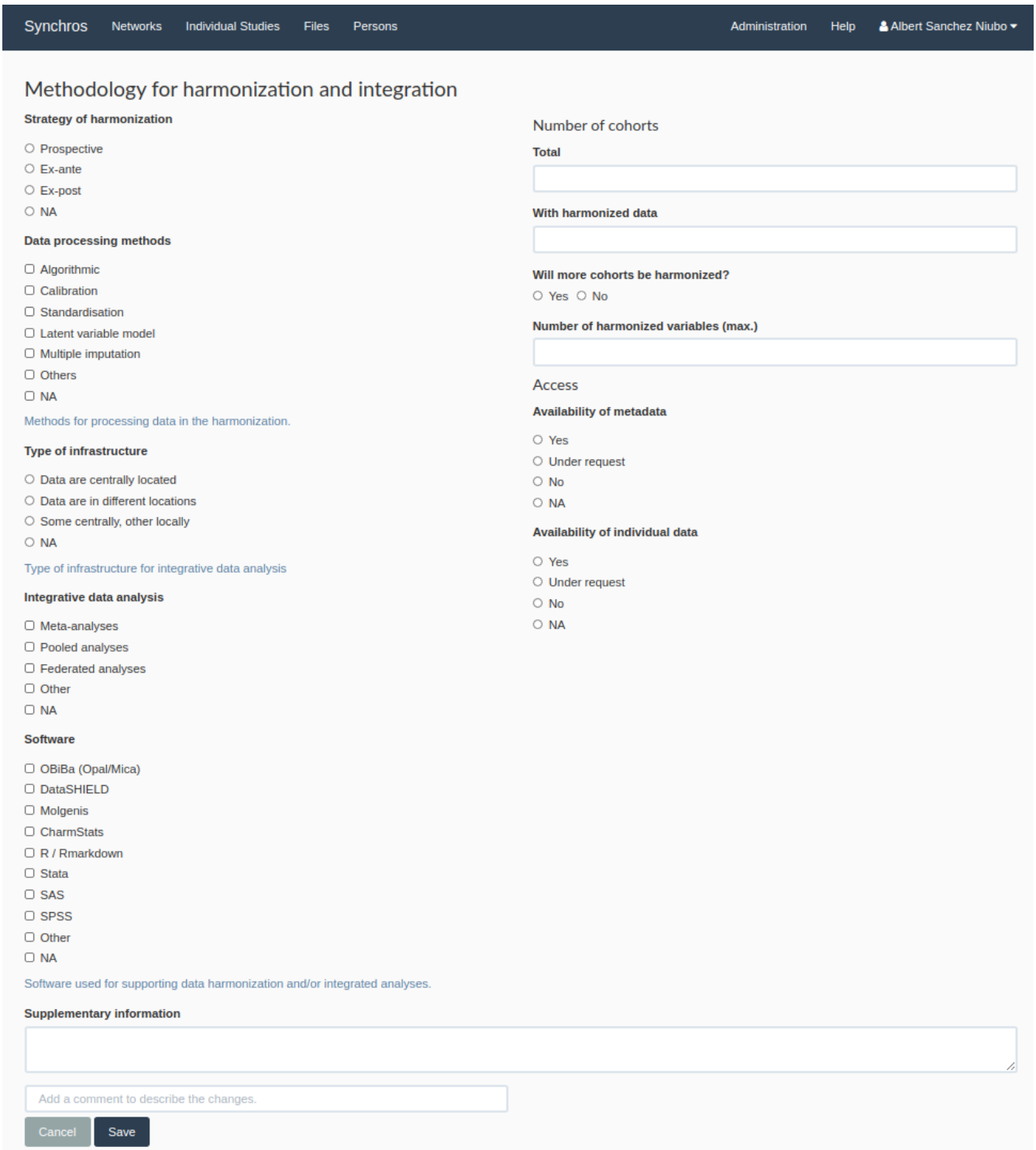

Figure 1. Fields collected for the methodology for harmonization and integration.

The selected fields are relevant to identify, first, to know which strategies and methods are mostly used in the harmonisation process and for which types of cohorts. Secondly, it is important to know how many initiatives

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

and which types of cohorts use infrastructures according to the availability of individual cohort data, i.e. centralised or federated infrastructures. In addition, the type of infrastructure can influence the type of analysis and type of software that are used. Thirdly, comparing the number of cohorts that are initially included in the initiatives and how many of them are successfully harmonised can help to see in which situations this may lead to more difficulties in the harmonisation process. Finally, knowing in which initiatives and type of cohorts they offer access to metadata and/or individual cohort data can tell us about the facilities for data and information sharing between different initiatives and with researchers outside the initiatives.

The information was first sought at the respective websites of the initiatives. Subsequently, working documents were searched or for some initiatives, even webinars or videos were consulted about parts of the study. Given the difficulty of finding or inferring the information required to complete the fields of methodology, we also searched for information in scientific articles. Therefore, we used keywords from the initiative in order to find clues, especially for the integrated data analysis methods the initiatives had used. In case the information was incomplete or unclear, several attempts were made to contact the initiative's main researcher or contact person in order to contrast and complete the information found.

To further clarify some of the fields presented here, we would like to refer to the repository itself where we explain the definitions of the different harmonisation strategies, the different data processing methods, the types of local or federated infrastructure, types of analysis and some of the most widely used software for the practice of harmonisation and integration (https://repository.synchros.eu/page/methodology). Figure 2 shows the content of this website.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

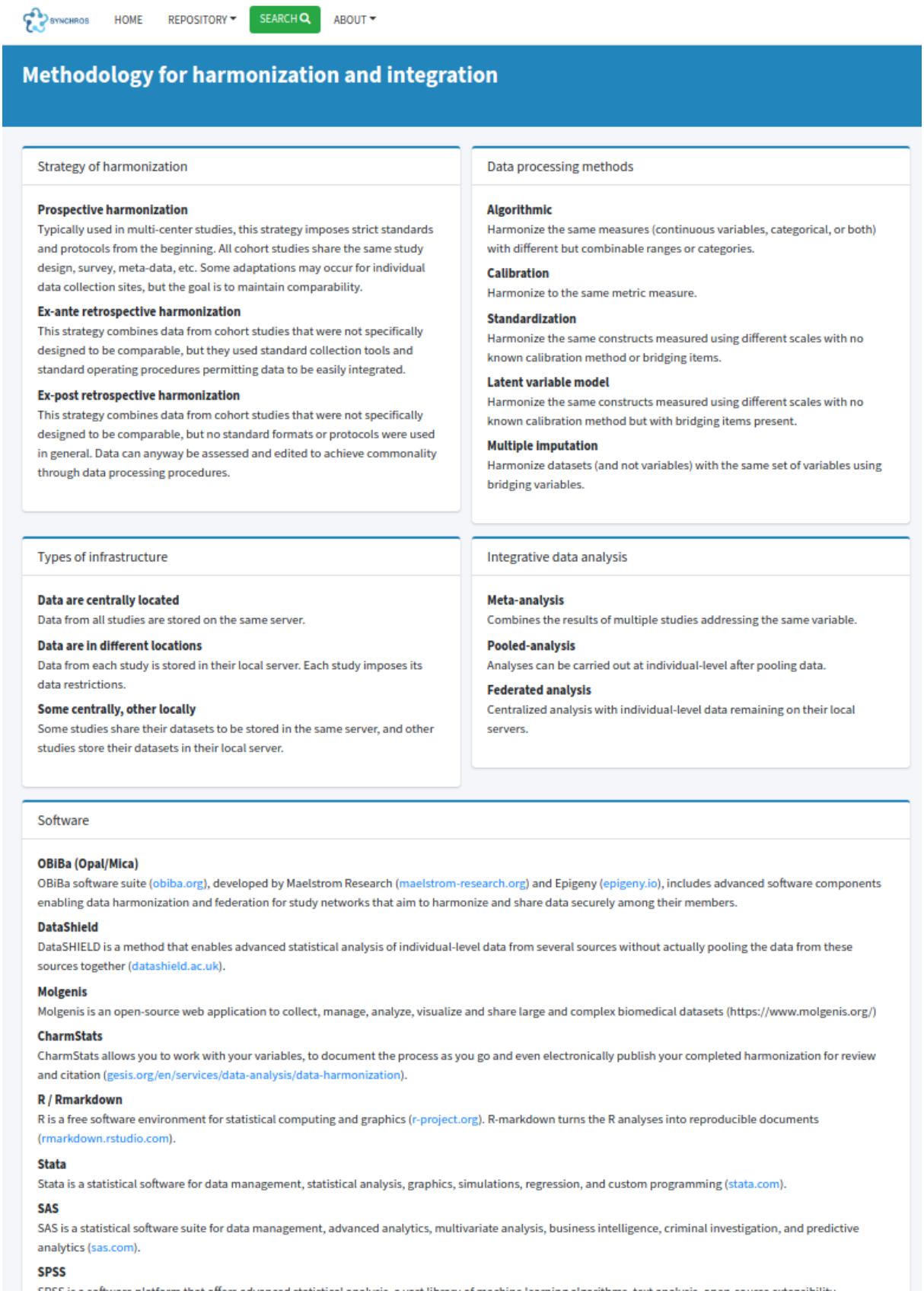

Figure 2. Web page with the methodology for harmonisation and integration

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

In addition, the main bibliography on which the previous terms were based is presented in Figure 3. This can also be consulted online (https://repository.synchros.eu/page/bibliography).

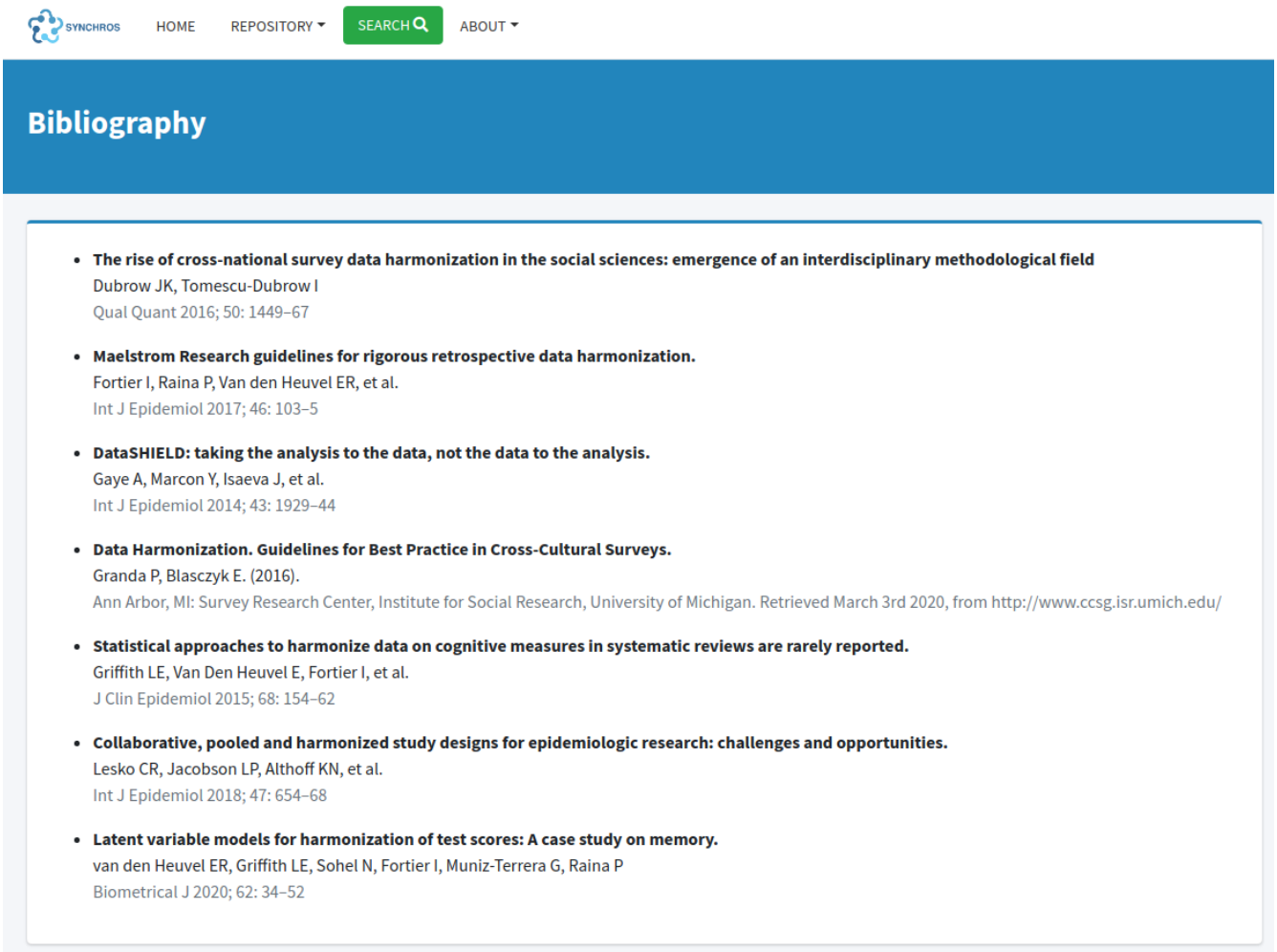

Figure 3. Web page with the bibliography of the methodology for harmonisation and integration.

3. PRELIMINARY RESULTS

The number of initiatives identified by the first and second mapping (initiatives with cohorts of populations, patients or clinical trials) was 121. However, during the review of the methodology we found that several initiatives were rather repositories not yet providing information about the harmonisation or integration of data; one was discarded because it had started before 2000 (exclusion criteria); and two were removed because they were repeated with different names and mixed with different types of cohorts and, thus overlapping. This current report, therefore is based on 114 initiatives for which information on harmonisation and integration methodology has been collected.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

Here it should be noted, that to date only 28 initiatives have received a confirmation from the researchers of the initiatives and could thus be made publicly available in the SYNCHROS repository. All the others have been entered into the web application but are awaiting the researchers' response so that they can be published correctly and with their consent.

Below are some preliminary results of the 114 initiatives, which may change as the researchers who coordinate the initiatives reply. In total, we identified:

- 47 initiatives with only population cohorts,
- 32 initiatives with only patient cohorts,
- 16 initiatives with only clinical trials,
- 8 initiatives with population and patient cohorts,
- 8 initiatives with patient cohorts and clinical trials, and
- 3 initiatives with all three types of cohorts.

3.1 Harmonisation strategies

We can differentiate three harmonisation strategies: prospective, ex-ante retrospective and ex-post retrospective. Of the total number of initiatives, approximately 80% applied ex-post retrospective strategies, with half of the population cohorts, 30% of patient cohorts and 20% of clinical trials. Of the remainder, 8 are considered to have followed ex-ante strategies and at least 12 have used prospective strategies. There are 4 initiatives that need to further clarification with the researchers in order to determine the followed strategy.

The median number of cohorts in the ex-ante and ex-post retrospective strategies is 12 cohorts; if we consider those that successfully harmonised cohort data, the median is 8-9. For the prospective strategy, the median is reduced to 7 in total but all were successfully harmonised.

3.2 Harmonisation strategies

In general, no practical information is available on how the different initiatives processed the data for harmonisation. At most, some can be inferred if they show the harmonisation algorithms for conversion between the study-specific variables and the harmonised end variable. In total, we have only been able to collect information from 6 initiatives. The rest have either not offered us any information, have not harmonised data yet, or we are still waiting for their response.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

The most common data processing methods in these 6 initiatives are algorithms and standardisation. As mentioned above, efforts are been made to obtain this information from the coordinators of the respective initiatives.

3.3 Type of infrastructure

We can differentiate between two main types of infrastructures for the *sharing* of individual data from the cohorts within the initiative: 1) the individual data is centralised in one institution or server, or 2) the individual cohort datasets reside in different institutions (federated), mostly on the server of origin. We found the same percentage of initiatives in each type of infrastructure. However, we found 18 initiatives for which it is not clear how the individual cohort data is located; we're currently awaiting the researchers' clarifications.

It should be noted that the percentage of initiatives with population cohorts with a centralised infrastructure is higher than those with the individual cohort data residing in different locations (57% vs 35%), and this is reversed in initiatives with patient cohorts (40% vs 54%), and varies little in initiatives with clinical trials cohorts (23% vs 27%).

3.4 Type of analysis for data integration

We differentiate between three types of analysis: meta-analysis, pooled-analysis and federated-analysis. More than half of the initiatives have performed pooled-analysis whereas the rest have performed meta-analysis or federated-analysis. The numbers are imprecise because there are quite some initiatives that are unclear about the type of analysis or are still in the process of harmonisation. There seems to be no clear distinction between types of analysis according to the type of cohorts.

3.5 Software used for harmonisation or in integration

56% of the initiatives did not report what type of software they used for the harmonisation and integration methodology. Of those we did know, more than the half used regular statistical software, mainly SAS and STATA; 28% used specific software such as OBiBa applications from Maelstrom, Datashield or Molgenis; and 14% reported using other software. This "other" software was not reported by the initiative researchers and as a consortium we have the intention to learn more about it, through contacting the researchers directly.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

3.6 Availability of harmonised data

Neither was it easy to see whether the metadata (37%) or individual data (66%) could be requested by researchers outside the initiative or that these were considered private (and thus not shared). For those that we did find this information, 86% offered the metadata, of which 35% were under petition. Furthermore, only 45% (17 initiatives) also offered the individual harmonised data, 13 of them under petition.

4. LIMITATIONS

The results presented here are preliminary because there are still 85 initiatives to be reviewed by the researchers who coordinate the initiatives. On the other hand, some of the 28 initiatives left blank fields, some of them because they had not yet started the harmonisation process, Therefore, the statistics shown in the previous sections should be interpreted with caution.

Although our effort in Deliverable 2.1 has been to define and classify the different strategies and methods of the harmonisation and integration process, all these terms may not yet be sufficiently known and clear to the researchers of the initiatives we have contacted, as the practice of the harmonisation and integration process may be more complex in some cases. In particular, it is not always clear in the retrospective strategy whether it is ex-ante or ex-post as sometimes both can be used according to certain types of subsets of data between cohorts. For example, socio-economic data may be constructed more arbitrarily (ex-post) while data on psychological or clinical measures may be well standardised in advance (ex-ante). This complicates the determination of an initiative as either ex-ante or ex-post, as sometimes both are being used by an initiative.

The form we created for the collection of methodological information has already had several variations as the different initiatives were reviewed. Therefore, it is not exempt that there may be future variations, more fields or more categories as we keep updating the repository with new initiatives and/or due to new technological developments new insights might be generated.

5. CONCLUSIONS

Knowledge about harmonisation and integration methodology has evolved greatly in recent years due to the efforts of researchers who sought ways to integrate data from different cohorts as validly as possible and, consequently, thanks to research groups such as Maelstrom Research and other institutions that have established practical harmonisation guides to facilitate in these efforts.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884

However, during the process of reviewing and collecting information we have found that in general, the researchers leading the initiatives do not follow the recommendations well enough to make the harmonisation process transparent and rigorous enough to assess its validity and facilitate reproducibility in future initiatives. It has been very difficult to find clear and concise information on how the process of harmonisation and integration has been carried out. Most probably, this is due to a lack of knowledge about the importance of documenting this information. On the other hand, we also understand that efforts are naturally directed more towards more profitable objectives of the initiative itself such as results of the research project than towards explaining the harmonisation process in detail.

We'd like to conclude this report by stressing that is highly important that future initiatives aimed at harmonising and integrating data from different cohorts follow systematic guidelines depending on the type of methodological scenario and the availability of data shared by the different cohorts. The process should be transparent, rigorous and well documented in order to justify its validity and to be able to be of help for other initiatives.

The SYNCHROS repository offers researchers the opportunity to search initiatives and cohort studies that meet certain criteria of interest. With regard to the methodologies, researchers will find out with a simple search which initiatives use specific harmonisation strategies, data processing methods, types of infrastructure, software they use and whether they can access the data. Since the initiatives contain links to the initiatives' websites and contact information, researchers can address them directly.

The availability of this repository in the research community provides a useful tool for disseminating existing initiatives and their key information, and as such promoting good practices in harmonisation and integration. In addition, contact between researchers from different initiatives can promote new synergies and share harmonised data or at least their practices into new initiatives to facilitate and leverage resources.

The SYNCHROS project (www.synchros.eu) has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under Grant Agreement No. 825884