University of Louisville *Journal of Respiratory Infections*

Towards Multi-Lingual Pneumonia Research Data Collection Using the Community-Acquired Pneumonia International Cohort Study Database

William A Mattingly[1]*, Kimberley Buckner[1], Senen Pena[1]

Abstract

Background: Although multilingual interfaces are preferred by most users when they have a choice, organizations are often unable to support and troubleshoot problems involving multiple user languages. Software that has been structured with multiple languages and data interlinking considerations early in its development is more likely to be easily maintained. We describe the process of adding multilingual support to the CAPO international Cohort study database using REDCap.
Methods: Using Google Translate API we extend the supported Spanish language version of REDCap to the most recent version used by CAPO, 8.1.4. We then translate the English data dictionary for CAPO to Spanish and link the two projects together using REDCap's hook feature.
Results: The Community Acquired Pneumonia Organization database now supports data collection in Spanish for its international collaborators. REDCap's program hook functionality facilitates both databases staying up to date. When a new case is added to the Spanish project, the case is also added to the English project and vice versa.
Conclusions: We describe the implementation of multilingual functionality in a data repository for community-acquired pneumonia and describe how similar projects could be structured using REDCap as an example software environment.

Introduction

Sharing the results of clinical research continues to play a large role in scientific discovery [1], and more and more funders and publishers are encouraging investigators to adopt sustainable means to share the data used along with their results. For ad-hoc studies the structure of data is not crucial beyond the needs of accurate analysis. However, when data sharing and reproducibility of research are important, time and effort must be invested to structure a dataset to be efficiently combined with other datasets. Efficient means of integrating data from different sources are needed to leverage the full benefits of data sharing because new research questions can be examined using integrated, multi-source data.

Studies that are international in scope can provide large, diverse samples of patient populations, but require investment in translation, and the appropriate ontological structuring of data. In this paper we describe the process of extending the Community Acquired Pneumonia Organization (CAPO) clinical research database to support data collection for multiple languages. After English, Spanish is the most common spoken language of members of the Community Acquired Pneumonia Organization with almost 40% of member sites being in Spanish speaking countries. Starting with Spanish we establish a general multi-language workflow for data entry into the CAPO database, with the eventual goal of supporting all CAPO member languages.

Methods

The study database for CAPO currently resides in a web-hosted instance of the REDCap electronic data capture software. REDCap is a secure web application for building and managing online surveys and databases. Members of CAPO access the REDCap instance remotely from its web URL https://id.research.louisville.edu/capo using a web browser and their assigned user credentials. Demographic and clinical history information can then be entered for new cases in the database.

Table 1 Glossary of terms

Term	Description
API	Application Programmer Interface – a standard language for extending software
CDISC	Clinical Data Interchange Standards Consortium
Hook	Customized software instruction created by end users of the software
ODM	Operational Data Model
Project-level	Pertaining to a feature that only exists in a single project
REDCap	A popular EDC
System-level	Pertaining to a feature that is present throughout an entire system
User credentials	User name and password
Web-hosted instance	An online version of the software

***Correspondence To:** William A Mattingly, PhD
Work Address: 501 E Broadway, Suite 120
Louisville, KY, United States, 40202
Work Email: bill.mattingly@louisville.edu
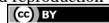
System Language

REDCap was designed to support a multilingual setup, with two levels: system-level and project-level. System-level languages are stored as files in REDCap's web directory. The default is English, but other files can be added from the REDCap community consortium page. If the default language is changed in REDCap's settings, then the entire REDCap instance, including technical configuration pages, will be in the new language. Alternatively, project-level languages can be changed, meaning the home page of REDCap and configuration pages will be in the default language, but a project can be configured to use a language different from the default system-level language. While it is not currently possible for REDCap users to have different language preferences in the same project, we describe how to duplicate and link a project to allow similar functionality.

REDCap language files store each segment of text instructions for every page of the software, one per line in a text document. The REDCap software is consistently updated with new features, and these features are pushed out to REDCap users in new versions of the software. All text accompanying new software features is added to the structured REDCap English language file. REDCap users can then translate the English language file and add the translation to their REDCap language files. All REDCap users are encouraged to share their translation files with other members on the REDCap community page. There are currently nine translations that have been tested and approved by the REDCap consortium, and these are listed in **Table 2**.

The CAPO REDCap database is currently using version 8.1.4 of REDCap. After installing the latest Spanish translation from the REDCap consortium, much of the descriptive text for REDCap will correctly display, but all new features added since version 6.4.1 will still have English text. To complete the translation, all text for those features must be added to REDCap's Spanish language file. We perform the language translation for the remaining phrases with a first pass translation using Google's translation application programmer interface (API) and a second pass with Spanish speaking collaborators.

Table 2 List of supported languages for REDCap and their current supported REDcap software version

Language	REDCap Versions Supported
Chinese	8.0.2
Danish	6.5.10
French	7.6.2
German	5.7.4
Italian	7.6.3
Japanese	7.0.5
Portuguese	4.8.8
Russian	6.4.3
Spanish	6.4.1

Project Language

REDCap supports assigning a translated language file to a specific project. This allows a multilingual setup by having one REDCap project per language for a multilingual database or repository. For the CAPO database, we duplicate the CAPO English REDCap project, set the new project to the Spanish language, and link the two projects so that data entered in the English project will populate in the Spanish project and vice versa. A diagram for this process is shown in **Figure 1**.
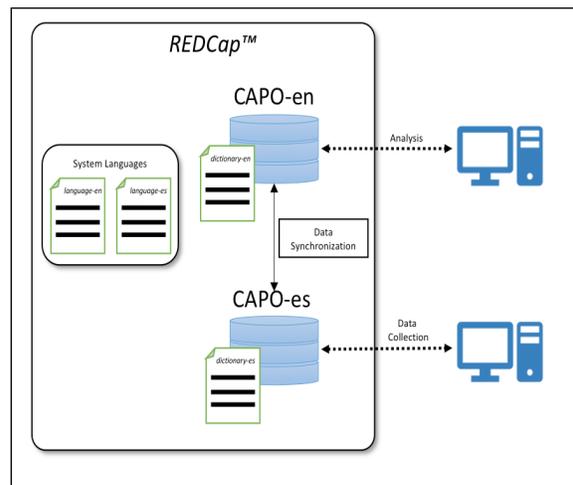
Figure 1 The data collection and analysis pipeline for multilingual projects. A REDCap system with two system supported languages, English and Spanish, is shown. The master project, English in this case, is used to perform analysis. All other projects are used for data collection. Projects are synchronized at the time of data entry.

Data Dictionary Changes

Changing the REDCap system and project languages will not affect any data entered by the user when designing a case report form (CRF) or entering data. Since the CAPO database and its associated variable labels were created in English, these must also be translated for a project with a different language. It is important for variable names to remain identical across projects, as this allows data analysis and reporting to be consistent for the entire database.

Each language needs to have translated versions of the Field Label column, the Choices Labels column and the Field Note column. Each of these columns contains text that is displayed on screen during data collection to assist in the collection of accurate information. **Table 3** shows a comparison of the two languages for a selection of variables from the CAPO data dictionary.

Table 3 Sample of labels and choices for variables in CAPO data dictionary shown in English and Spanish

Variable	Field Label (English)	Choices Labels (English)	Field Note (English)	Field Label (Español)	Choices Labels (Español)	Field Note (Español)
dem_age	Age			Años		
dem_sex	Gender	1, Male | 0, Female		Género	1, Hombre | 0, Mujer	
dem_pregnant	If female, is she pregnant?	1, Yes | 0, No | 2, Puerperal State		Si es mujer, ¿está embarazada?	1, Sí | 0, No | 2, Estado puerperal	
dem_trimester	If pregnant, what trimester?	1, 1st Trimester | 2, 2nd Trimester | 3, 3rd Trimester		Si está embarazada, lo trimestre?	1, 1er trimestre | 2, segundo trimestre | 3, 3er trimestre	
hosp_vent	Did the patient need ventilatory support on day 0?			¿El paciente necesita soporte ventilatorio en el día 0?		
hosp_vaso	Did the patient need vasopressors on day 0?			¿El paciente necesita vasopresores en el día 0?		
prevent_pneumovaxyear	If patient already received the vaccine, approximate year of receipt	4-digit year e.g. (2001)				por ejemplo año de 4 dígitos (2001)

Data Synchronization

As mentioned above, user-specific languages are not yet supported in REDCap, but two separate projects can be linked allowing users preferring English data entry to be assigned to the English project, and users preferring Spanish data entry to be assigned to the linked Spanish project. Keeping data consistent requires a synchronization operation be performed between the two projects. This is supported in CAPO by means of a REDCap hook created by the authors.

It is often desirable to link records in REDCap projects to similar information that may exist in other projects on the same REDCap instance. For example, two studies may be performed on an overlapping patient population. Following the completion of the studies, it may be beneficial to study the overlap or to consider a third study which could involve the same population. In the past this has been managed using REDCap hooks.

A REDCap hook is software code saved on the REDCap instance that can perform an action under certain circumstances. These circumstances can include saving a record, opening a data entry form, or completing a survey. After creating a hook, the custom action is performed whenever one of the above conditions occurs. For linking the English and Spanish language CAPO projects, the hook condition is whenever a record is saved. For the English project hook the action performed is to save all record information from the current record to the Spanish project data fields. The Spanish language hook is identical except that it saves information entered to the English project.

Organisms and Medications

Another section of the CAPO database requiring translation is the list of organisms commonly associated with CAP and medications prescribed. These are stored in the CAPO REDCap instance separate from the CAPO project. The names in these lists will generally be the same across languages, but the lists can be extended to support localized names of organisms and medications. In these instances, the numeric codes would remain the same and serve as the link between translations. **Tables 4** and **5** show a selection of medications and organisms in CAPO, respectively, along with associated codes in the RXNORM and SNOMED CT Clinical Terms Ontologies.

Table 4 Sample list of medications in CAPO with associated RxNORM and SNOMED CT codes

CAPO	RxNORM	SNOMED CT
Anidulafungin	anidulafungin (341018)	Anidulafungin (422157006)
Atovaquone	Atovaquone (60212)	Atovaquone (108718008)
Azithromycin	Azithromycin (18631)	Azithromycin (387531004)
Aztreonam	Aztreonam (1272)	Aztreonam (69918003)
Caspofungin	Caspofungin (140108)	Caspofungin (413770001)
Cefaclor	Cefaclor (2176)	Cefaclor (387270009)
Cefadroxil	Cefadroxil (2177)	Cefadroxil (372651006)

**Table 5** Sample list of organisms in CAPO with associated SNOMED CT codes

CAPO	SNOMED CT
Acinetobacter spp.	Acinetobacter (7757008)
Actinomyces spp.	Actinomyces (40560008)
Adenovirus	Adenoviridae (424470006)
Aspergillus spp.	Aspergillus (2429008)
Bacteroides spp.	Bacteroides (57522007)
Chlamydia pneumoniae	Chlamydophila pneumoniae (103514009)
Chlamydia psittaci	Chlamydophila psittaci (14590003)
Citrobacter spp.	Citrobacter (75972000)
Coccidioides immitis	Coccidioides immitis (23439005)

Results

The Community Acquired Pneumonia Organization database now supports data collection in Spanish for its international collaborators, as shown in **Figure 2**. REDCap's program hook functionality facilitates both databases staying up to date. When a new case is added to the Spanish project, the case is also added to the English REDCap CAPO project and vice versa.

Several beneficial changes to the database occurred because of the addition of multilingual functionality. An incorrect antibiotic entry of Defixime in the CAPO medications table was discovered when there were no corresponding entries in either the RXNORM or SNOMED CT databases. We determined this was a duplicate of Cefixime that had been entered as a typo.

Several potentially ambiguous entries for organisms were made more specific by linking them to the standardized organisms listed in SNOMED CT. Pseudomonas pseudomallei was linked to the more standardized Burkholderia pseudomallei (116399000) and Rhinovirus/Enterovirus was linked to Human enterovirus (69239002). All new additions to the CAPO organisms and medications tables are entered in their own REDCap projects and verified with SNOMED CT and RXNORM lookups before being finalized. This prevents the usage of ambiguous or non-standard names.
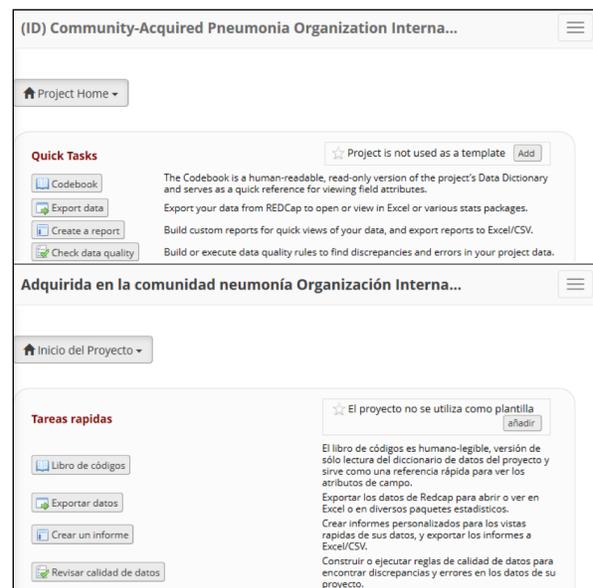
Figure 2 The project level user interface for the database, shown above is the English version with the corresponding interface for the Spanish translation below.

Discussion

Besides the organizational challenges in managing multi-site studies, differences in languages pose a problem to the consistency of collected data. Case report forms that have been designed to collect accurate information, must be carefully translated so as not to lose any validity. A study found that the translation of parenting programs to help Hispanic mothers was successful when translated from English to Spanish, but cultural adaptation also played a large role, which is more complex and time consuming than simple translation. (2) A study on the interpreting practices in multilingual healthcare found that the availability of professional interpreters was very limited and needed to be planned far in advance. The ability of an individual to interpret language correctly is still highly related to their socio-economic status and related cultural factors. (3) One study showed differing attitudes towards multilingual patient engagement at the organizational and patient level. Patients are generally enthusiastic whereas organizational and institutional members are concerned about cost and maintenance of complex systems [4].

The translation of focused data collection instruments may be more successful across cultural divides within a language. The Stroke Impact Scale (SIS) was translated to Portuguese and showed high consistency across different interviewers. [5]. Surveys and data collection concepts which map to visual concepts are also easier to translate, and in some cases do not require translation to be effective [6]. Systems that use pictograms and visual icons can successfully bridge language barriers and sometimes achieve a language-neutral solution for health applications [7]. Systems which provide online multi-lingual data collection and patient care are becoming more common.

In addition to effort in translating data collection forms, multilingual research requires adjustment in the structure of data. The CDISC Operation Data Model (ODM) was implemented as a way for physician researchers to upload, comment, rate and download medical data models [8]. The system includes forms from clinical and research systems and supports multi-lingual data files. A later examination of the ODM model has found that it is being used as a metadata standard for many use cases including Electronic Data Capture and Electronic Medical Records, data collection, data tabulation and analysis, and data archival [9]. Systems needing more features than the ODM model can be developed keeping many of the fundamental metadata design principles of multi-lingual and flexible communication in mind. Previous efforts to this end were successful in developing a shared research environment for radiotherapy clinical data mining [10]. The Pediatric Oncology Network Database (POND4KIDS) was designed to be a multilingual hematology/oncology database for use in countries with limited resources [11]. The database recently reported over 1000 users in 66 different countries [12].

A recent study suggests automated translation services, like Google Translate, may begin to be efficient tools for medical translation, as they were able to extract consistent translations and data from Latin-based language manuscripts [13]. Google Translate was also used to provide an effective translation of the Gene Ontology from English to German [14]. The Global Public Health Intelligence Network (GPHIN) was designed to process data from news sources across the globe and uses automated translation across nine languages to provide aggregate information regarding disease outbreaks [15]. Translation of a nutrition survey in Switzerland from German to French and Italian, used a software tool called GloboDiet to design multilingual surveys [16].

Medical ontologies provide ways to organize both the conceptualization and sharing of data among different researchers. The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is a widely used ontology [17] and can be used as a structure to link multilingual medical terms. The Wikimedia Foundation maintains a large set of multilingual data in the form of articles across multiple Wikipedia projects. Garcia et al, previously used the interlanguage links of Wikipedia to apply concept mapping in their efforts to develop a classifier for multilingual biomedical documents [18].

Although people will prefer to enter data or interact with electronic systems in their preferred language, the cost associated with supporting more than one language will continue to be a barrier for adoption of multilingual systems. Systems like REDCap, which are available free of charge to non-profit researchers, struggle to keep up with maintaining multiple languages and depend on volunteer submissions to stay up to date.

One clear advantage to multilingual capability is the emphasis placed on structure. When a new feature is proposed for REDCap time and thought is invested in understanding the different effects a new feature will have. Is there already a similar feature that performs the same function? Where will the button for the feature be placed? Will there be more than one way to access it? Will it conflict with any other current features? This same approach to structure was taken with REDCap's language feature list and systems like CDISC ODM and WikiData. This places emphasis on how concepts will link together and interact, which is considered by some to be the most difficult aspect of translation.

It is important to invest time early in the design of data repositories and data warehouses to ensure they will support future features like data interlinking and multilingualism.

Limitations

Currently the synchronization between databases is triggered when data collection is done directly into REDCap fields and not when data is imported using REDCap's data import tool or the REDCap API. This is a limitation of REDCap's hook feature. Future synchronization schemes will involve a centralized data project or data warehouse that periodically pulls records from language specific projects to catch all data contribution methods.

Conclusion

We describe the implementation of multilingual functionality in a data repository for community-acquired pneumonia and describe how similar projects could be structured using REDCap as an example software environment. We believe there is a strong future in cross-language collaborative studies with tools like REDCap and Google translate that would extend the reach of multi-site clinical and observational trials.

Acknowledgements

References

1. Kiley R, Peatfield T, Hansen J, Reddington F. Data Sharing from Clinical Trials - A Research Funder's Perspective. N Engl J Med. 2017 Nov;377(20):1990–2.

2. Beasley LO, Silovsky JF, Espeleta HC, Robinson LR, Hartwig SA, Morris A, et al. A qualitative study of cultural congruency of Legacy for Children™ for Spanish-speaking mothers. Child Youth Serv Rev. 2017 Aug 1;79:299–308.

3. Hadziabdic E, Lundin C, Hjelm K. Boundaries and conditions of interpretation in multilingual and multicultural elderly healthcare. BMC Health Serv Res. 2015 Oct;15(1):458.

4. Ochoa A 3rd, Kitayama K, Uijtdehaage S, Vermillion M, Eaton M, Carpio F, et al. Patient and provider perspectives on the potential value and use of a bilingual online patient portal in a Spanish-speaking safety-net population. J Am Med Inform Assoc. 2017 Nov;24(6):1160–4.

5. Brandão AD, Teixeira NB, Brandão MC, Vidotto MC, Jardim JR, Gazzotti MR. Translation and cultural adaptation of the stroke impact scale 2.0 (SIS): a quality-of-life scale for stroke. Sao Paulo Med J. 2018 Mar;136(2):144–9.

6. Lim L, Ng TP, Ong AP, Tan MP, Cenina AR, Gao Q, et al. A novel language-neutral Visual Cognitive Assessment Test (VCAT): validation in four Southeast Asian countries. Alzheimers Res Ther. 2018 Jan;10(1):6.

7. Wołk K, Wołk A, Glinkowski W. A Cross-Lingual Mobile Medical Communication System Prototype for Foreigners and Subjects with Speech, Hearing, and Mental Disabilities Based on Pictograms. Comput Math Methods Med. 2017;2017(2):1-9. doi:10.1155/2017/4306416.

8. Breil B, Kenneweg J, Fritz F, Bruland P, Doods D, Trinczek B, et al. Multilingual Medical Data Models in ODM Format: A Novel Form-based Approach to Semantic Interoperability between Routine Healthcare and Clinical Research. Appl Clin Inform. 2012 Jul;3(3):276–89.

9. Hume S, Aerts J, Sarnikar S, Huser V. Current applications and future directions for the CDISC Operational Data Model standard: A methodological review. J Biomed Inform. 2016 Apr 1;60:352–62.

10. Roelofs E, Dekker A, Meldolesi E, van Stiphout RG, Valentini V, Lambin P. International data-sharing for radiotherapy research: an open-source based infrastructure for multicentric clinical data mining. Radiother Oncol. 2014 Feb;110(2):370–4.

11. Antillon FA, Ribeiro RC. POND4Kids: a web-based pediatric cancer database for hospital-based cancer registration and clinical collaboration. International perspectives in health informatics. 2011 Feb 9;164:227.

12. Quintana Y, Patel AN, Arreola M, Antillon FG, Ribeiro RC, Howard SC. POND4Kids: a global web-based database for pediatric hematology and oncology outcome evaluation and collaboration. Stud Health Technol Inform. 2013;183:251–6.

13. Balk EM, Chung M, Hadar N, Patel K, Winifred WY, Trikalinos TA, Chang LK. Accuracy of data extraction of non-English language trials with Google Translate.

14. Hailu ND, Cohen KB, Hunter LE. Ontology translation: A case study on translating the Gene Ontology from English to German. InInternational Conference on Applications of Natural Language to Data Bases/Information Systems 2014 Jun 18 (pp. 33-38). Springer, Cham. https://doi.org/10.1007/978-3-319-07983-7_4.

15. Dion M, AbdelMalik P, Mawudeku A. Big Data: Big Data and the Global Public Health Intelligence Network (GPHIN). Canada Communicable Disease Report. 2015 Sep 3;41(9):209.

16. Chatelan A, Marques-Vidal P, Bucher S, Siegenthaler S, Metzger N, Zuberbuehler C, Camenzind-Frey E, Renggli A, Bochud M, Beer-Borst S. Lessons learnt about conducting a multilingual nutrition survey in Switzerland: Results from menuCH pilot survey. Int. J. Vitam. Nutr. Res. 2017.

17. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. J Am Med Inform Assoc. 2014 Feb;21 e1:e11–9.

18. Antonio Mouriño García M, Pérez Rodríguez R, Anido Rifón L. Leveraging Wikipedia knowledge to classify multilingual biomedical documents. Artif Intell Med. 2018 Jun;88:37–57.