EudraCT Safety Data Input Software Tool

2. Abstract

A suite of software tools have been provided, across the three main statistical programming languages, to provide an efficient means to convert safety data collected from clinical trials, from a one-row-per-event format into the XML format of aggregated frequency statistics required by the EudraCT web portal for upload. Thus a time-consuming step that could require hundreds of hours of work to enter the data manually has been reduced to a few lines of code.

3. Introduction

European Union Drug Regulating Authorities Clinical Trials Database (EudraCT) is run by the European Medicines Agency (1). All studies that are officially registered clinical trials must enter the results of the final analysis into the website thence to be made public. This is a laudable aim to prevent publication biases affecting meta-analyses, and to avoid repetition of negative studies incurring unnecessary expenditure of resources and risks to patients. However, the detailed entering of safety information into the portal is an onerous task if it is to be done by hand. An alternative to manual data entry is the facility to upload an Extensible Markup Language (XML) file.

A prototype of a tool that creates the required XML files was developed internally within Cambridge Clinical Trials Unit after the introduction in 2015 of the regulatory requirement to upload the results of Clinical Trials of Investigational Medicinal Products (CTIMPs) to the EudraCT portal. The funding call for "Efficient / innovative delivery of NIHR research – 2018" provided a timely opportunity to share this tool with the wider UK-wide clinical trials unit (CTU) network and beyond. This project commenced following a successful application in October 2018 for a period of one year.

4. Methods

The project breaks down into several main phases as described below

4.1. Planning

An online survey sampling from registered CTUs was run using Survey Monkey (2) to: establish existing practices across the CTU network; gauge the level of interest in the proposed tool; understand how the tool would need to interact with the users.

Following the survey a formal specification document was written to reflect the requirements ascertained from the survey, and which would be used to compare against at the stage of User Acceptance Testing (UAT)

4.2. Building

The version of the tool aimed at users of the R statistical software package (3) was built as the primary objective so as to reach a high standard of quality control with rigorous unit testing and manuals documenting the tool within the time scales of the project. However preliminary versions of the tool for the other main statistics package in use, SAS and Stata (4), were also created.

We wished to acknowledge that widely varying practices and pre-existing tools already exist across the CTU network, that produce the core content (as opposed to format) of the frequency counts that were required by EudraCT. The key challenge lies in converting this content into the very specific format required by EudraCT to allow a one-step upload of an XML file. Hence the tool is modularised into 3 parts:

- Production of statistical frequency counts
- Saving into a common standard of a semi-readable xml data file
- Converting into EudraCT format

The first of these is relatively straightforward and would be considered a basic task that any competent statistician could accomplish. Nonetheless, a tool is provided to standardise this step, but also which could be supplanted by the user's own code.

The conversion of an arbitrary data set into a simple, semi-readable XML file is a generic task that each of the statistics packages can achieve easily with their own inbuilt tools. However the precise format of the outputted XML file varies slightly between the three statistical packages. So a common format was given in the specification requiring a degree of subsequent reformatting.

The language XSLT (eXtensible Stylesheet Language Transformations) (5) is ideal for converting XML documents between different formats, and is widely adopted in the wider programming community. A single script is used to achieve the transitions to the third part above, which gives an accessible means to be amended should the need arise in future. Also it is used to achieve convergence to the common standard needed for step 2 from the simple XML file output from the statistical packages.

XSD (XML Schema Definition) is a technology that allows a set of rules regarding the desired format of an XML file to be specified in a file, called a *schema* (6). A schema is then used to check instances of XML files against the rules. A specific schema is provided by EudraCT to define appropriate rules for the output files (7), and we wrote a new schema to check the output from part 2 above. Checks are made internally within the tools against these two schemas.

4.3. Testing

Unit-testing was performed by an independent programmer with a formal set of tests written in advance based on the specification document. Tests were performed and compared to the pre-specified desired output.

Testing using SAS and Stata is planned for future work.

An R package was created to bundle together the tools, documentation, and tests for convenient sharing and installation. An R package must meet an extensive list of criteria for internal consistency and documentation.

5. Results and Conclusions

5.1. Survey

The survey revealed several strands regarding current practices and desiderata for future tools.

There was a need and demand to enter safety data more efficiently across CTUs, with a lack of formal procedures current in place. Data were entered into EudraCT by hand most commonly. Stata and then SAS are the statistical packages used primarily the most, with R being the commonest use in a secondary role. MACRO was the leading choice of database system, although less than half of the total, with the remainder being split over a large variety of systems. Medical Dictionary for Regulatory Activities (MedDRA) is the majority choice for coding safety data.

There was an even split between the preferences in terms of the user-interface and the subsequent level of UAT to be placed as the users' responsibility. One choice was to provide code directly and give more flexibility and responsibility to the users; the other was for a closed-box tool that could be accessed on a web site, with less flexibility and responsibility.

These results led to the modification of the specification to ensure that access to the tool would be provided to users of SAS and Stata, as well as R, which was the language used to develop the initial prototype.

We considered that developing code would be a necessary first step that would precede any web-portal delivery method for the tool. Also there would be issues around providing back-up systems to a web page which would be in conflict with the need to respect General Data Protection Regulation (GPDR) and data confidentiality regarding any uploaded clinical trial data. A statistician was commonly involved in uploading data to EudraCT hence the audience for the tool would be literate in programming. The decision was taken to develop modifiable code rather than a closed-box web-portal.

5.2. Tool

A website provides the results of the project: https://eudract-tool.medschl.cam.ac.uk/ . This contains explanation, links to external material, and downloads of the tools and documentation.

The core of the code and files that constitute the project are mirrored on a GitHub repository: https://github.com/shug0131/eudraCT/ . This provides a structured framework that is openly accessible world-wide for programmers to inspect, copy, and feedback potential changes. Should there need to be modification in future then this allows collaborators to easily access the existing code and commence the required work.

The R package has been released on the CRAN network: https://CRAN.R-project.org/package=eudract . This provides an extra degree of external quality control, and enables the tool to be installed very easily within R with a single command: install.packages("eudract") . This is issued under a GPL2 licence (6), which, in brief, allows anyone to freely use, copy, or modify the software, under the condition that these terms are preserved.

Hence the overall objectives of the project have been met: to provide a tool that avoids the need to expend hundreds of hours of data entry into the EudraCT portal for an individual study, by instead using a freely available and rigorously tested software tool.

6. Dissemination

The results of the initial survey were presented at a Webinar hosted by the UKCRC Registered CTU Network on 22nd February 2019.

The website, GitHub repository, and CRAN package all combine to provide a multi-facetted means for users to access the tool and its documentation.

The biannual meeting of lead statisticians from NIHR-registered CTUs has been used as a forum to update the wider community of the project's status. Several external CTUs have used the software and since gotten in touch to offer helpful feedback which has be incorporated into the tool.

A poster and presentation at the 5th International Clinical Trials Methodology Conference in Brighton was used to disseminate the tool to an audience beyond the NIHR CTU network (7).

A workshop will be held to provide a training session for interested users within the next 6 months, and incorporate a workshop and seminars on reporting of adverse events in clinical trials.

A journal article will be produced to describe the tool over the next year.

7. Acknowledgements

8. References

1. **European Union Drug Regulating Authorities.** *EudraCT Public Website.* [Online] [Cited: 20 11 2019.] https://eudract.ema.europa.eu/.

2. **Survey Monkey.** [Online] [Cited: 20 11 2019.] https://www.surveymonkey.co.uk/.

3. **R Foundation for Statistical Computing.** *R: A Language and Environment for Statistical Computing.* Vienna, Austria : s.n., 2013.

4. **StataCorp.** *Stata Statistical Sofware: Release 13.* s.l. : College Station, Tx: StataCorp LP, 2013.

5. **W3C.** XSL Transformations (XSLT) Version 1.0. [Online] 1999. [Cited: 20 11 2019.] https://www.w3.org/TR/1999/REC-xslt-19991116.

6. —. XML Schema Part 1: Structures Second Edition. [Online] 2004. [Cited: 21 11 2019.] https://www.w3.org/TR/xmlschema-1/.

7. **EudraCT.** EudraCT results schema. [Online] [Cited: 21 11 2019.] https://eudract.ema.europa.eu/docs/technical/schemas/clinicaltrial/results/adverseEvents.xsd .

8. **Free Software Foundation.** GNU General Public Licence Version 2. [Online] 1991. https://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html.

9. Meeting abstracts from the 5th International Clinical Trials Methodology Conference. **ICTMC 2019.** 579, 2019, Trials, Vol. 20.

9. Appendices
- Survey Slides
- Specification document.

10. Conflict of Interest declaration
No conflicts of interest.