

CodEX: when Excel meets SAS to code clinical data

Stéphane Chollet, Keyrus Biopharma, Levallois-Perret, France

Catherine Scoupe, Keyrus Biopharma, Lasne, Belgium

ABSTRACT

Coding is the use of international dictionaries to standardize medical terms and treatments reported in a clinical database. The purpose is to standardize and categorize the terminology of these events, avoiding the country-specific classifications, for analysis purposes.

As a Contract Research Organization (CRO), we work with different sponsors on different database structures, which means that development (programming and validation) is necessary to fit the coding model to each study. Another challenge for coding is the presence of multiple profiles (developer, data manager, coder, study physician), all of whom use different computer communication languages. This means that a common language must be found; ideally one which uses a system known to everyone and which is present on all computers.

Therefore, we sought to develop a coding tool that reconciles the need for standardization and the need for a single communication tool.

INTRODUCTION

The Clinical Data Management (CDM) process encompasses the entry, coding, cleaning, validation and reporting of data gathered during a clinical trial (i.e. during a research study conducted to answer specific questions about drugs, vaccines, biological or medical devices, new therapies, surgical procedures, etc.). Data are collected via Case Report Forms (CRF) that are either paper or electronic and are entered into the Clinical Data Management System (CDMS).

In order to facilitate the statistical analysis, medical events (e.g. adverse events, medical history etc.) and concomitant treatments are commonly coded using standard medical dictionaries (e.g. MedDRA, WHO-DD, etc.) – other elements can be coded depending on the project. As a Contract Research Organization (CRO), we work with different sponsors on different database structures, which presented a problem when standardizing the coding model.

Many different profiles work together during the coding process. Each of these profiles talks a different computer language: data managers talk SAS®, coding officers talk web browser, physicians talk paper listing, etc. In order to ease communication we needed to find a single communication channel for all the profiles involved in the process.

The purpose of this article is to present the solution we developed in order to answer the two challenges we faced when developing a hybrid coding tool.

BUSINESS CONTEXT

The purpose of data management is to ensure completion, accuracy and consistency of study data. To this purpose, before the study starts, the CDM:

- Sets-up a (e)CRF to capture the data needed to answer the protocol questions,
- Sets-up controls to ensure that data are complete, consistent and accurate.

During the study the CDM:

- Ensures data collection (from (e)CRF, randomization, central laboratories results, ...),
- Ensures proper execution of the controls and their documentation until resolution or closure.

Finally at the end of the study, the CDM reports events which occurred during the study and provides data for the statistical analysis.

PhUSE 2015

Coding is fully integrated into the CDM process as shown in *Figure 1*.

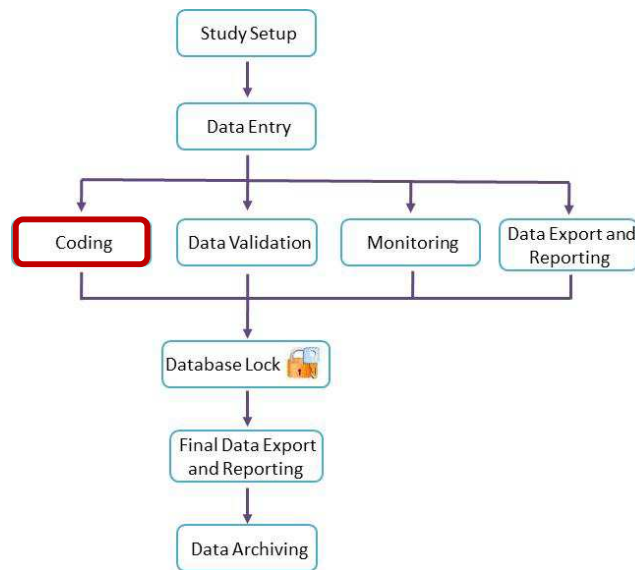


Figure 1: Coding location in the clinical data management process

According to the GCDMP October 2013 Edition, chapter “Medical Coding Dictionary Management & Maintenance”, “The use of medical coding dictionaries for medical terms [...] is valuable from the standpoint of minimizing variability in the way data are reported and analyzed”.

Coding can be summed-up as matching a list of terms present in the clinical database with a predefined list of standard terms. There are several ways to perform coding:

- **Automatic** = match is performed by a program,
- **Hard-coding/manual coding** = match is performed by a coding officer,
- **Hybrid** = a program prepares the coding (automatic coding) and a coding officer provides solutions for terms matched in error (hard-coding) (Cf. *Figure 2*).

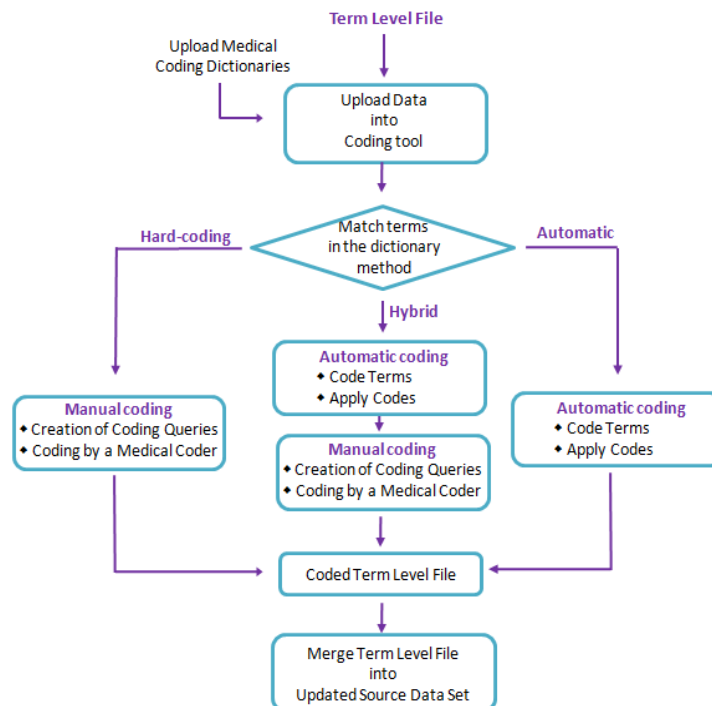


Figure 2: Principle for coding process

PhUSE 2015

The automatic process is frequently combined with electronic data capture (EDC) software and allows real time coding. A list of synonyms is uploaded, allowing consistency between terms to code. However, if updates are required on a frozen database following medical requirement/judgment, the database needs to be unlocked.

The hard-coding process requires trained people to ensure an efficient match. This method provides a medical expertise that the automatic process does not, and for a few terms, avoids developing and maintaining programs that perform automatic coding. But the consistency and accuracy of the coding for a given term is not guaranteed.

The following table lists the basic requirements for each method:

Requirements	Automatic	Hard-coding
Use of a dictionary that answers the project requirements	x	x
Use of a project-specific synonym list	x	
Implementation of an audit trail	x	
People trained in the use of standard dictionaries		x
Storage of dictionaries with versioning	x	

Table 1: Basic requirements for each type of coding

As previously said, we work for several sponsors, which mean that we work with different database structures and EDC systems. Databases can be very extensive, so the hard-coding option is not an efficient standard solution for our company. An integrated automatic solution is also not a good way to proceed because data are not necessarily coming from the same EDC.

This is why we made the choice to work on a hybrid process (described in *Figure 2*). Terms are automatically coded using a program, the standard list of synonyms and the study specific (if existing) list of synonyms. A coding officer reviews and corrects (when necessary) all the coded terms to provide medical expertise. The coding officer also provides solutions for all the terms which were not coded at the automatic coding step.

In order to be consistent during the whole process, the proposed solutions are compiled in a study specific list of synonyms. However, before the specific list of synonyms creation/update and before the medical/sponsor revision, a quality control is automatically done to ensure the consistency and accuracy of the coded terms. In case of inconsistency, a query can be addressed to the correct person (e.g. to the investigator to require splitting or term clarification, or to the coding officer to harmonize coding between two terms with the same verbatim).

Because the hybrid method is a combination of automatic coding and hard-coding, we need to follow both sets of requirements (*Table 1*).

CODEx®: WHEN EXCEL MEETS SAS®

According to the requirements in *Table 1* and the process described in *Figure 2*, the CodEX® tool has been developed to cover:

0. Standard dictionary SAS® data-set creation and maintenance,
1. Data extraction,
2. Automatic coding,
3. Manual coding (i.e. hard-coding),
4. Data management QC and consolidation,
5. Medical or sponsor review,
6. Specific synonyms and data loading if available.

Step 0 is done only once, during dictionary maintenance and can then be externalized from the tool and considered as an entry. Steps 1 to 6 are repeated on an ongoing basing (or at request) while non-coded terms exist.

To address the number of sponsors and database structures, step 1 is 'normalized' with a new document to add in the data management plan (DMP). This document specifies how data are sent to the tool and the mapping.

GLOBAL PRESENTATION

CodEX® is cut into several modules with their appropriate tool:

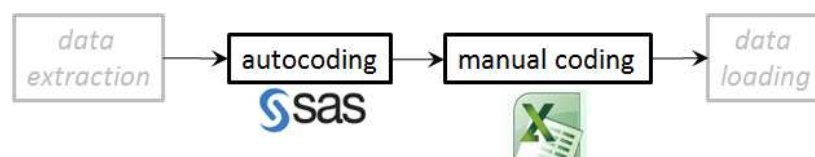


Figure 3: CodEX® uses 2 main softwares for 4 main steps

PhUSE 2015

Coupling SAS® programs (to manage data handling) with MS Excel outputs (accessible on most computers) provides a universal communication format between man and computer on one hand, and between all clinical profiles on the other. The most important point of the development process is simplicity; which would allow us to have a system that is:

- Easy to use and be trained on,
- Easy to develop and maintain,
- Flexible.

Multiple profiles are involved together in the coding process. And as we want it simple, all communications between stakeholders are done using Excel as described below:

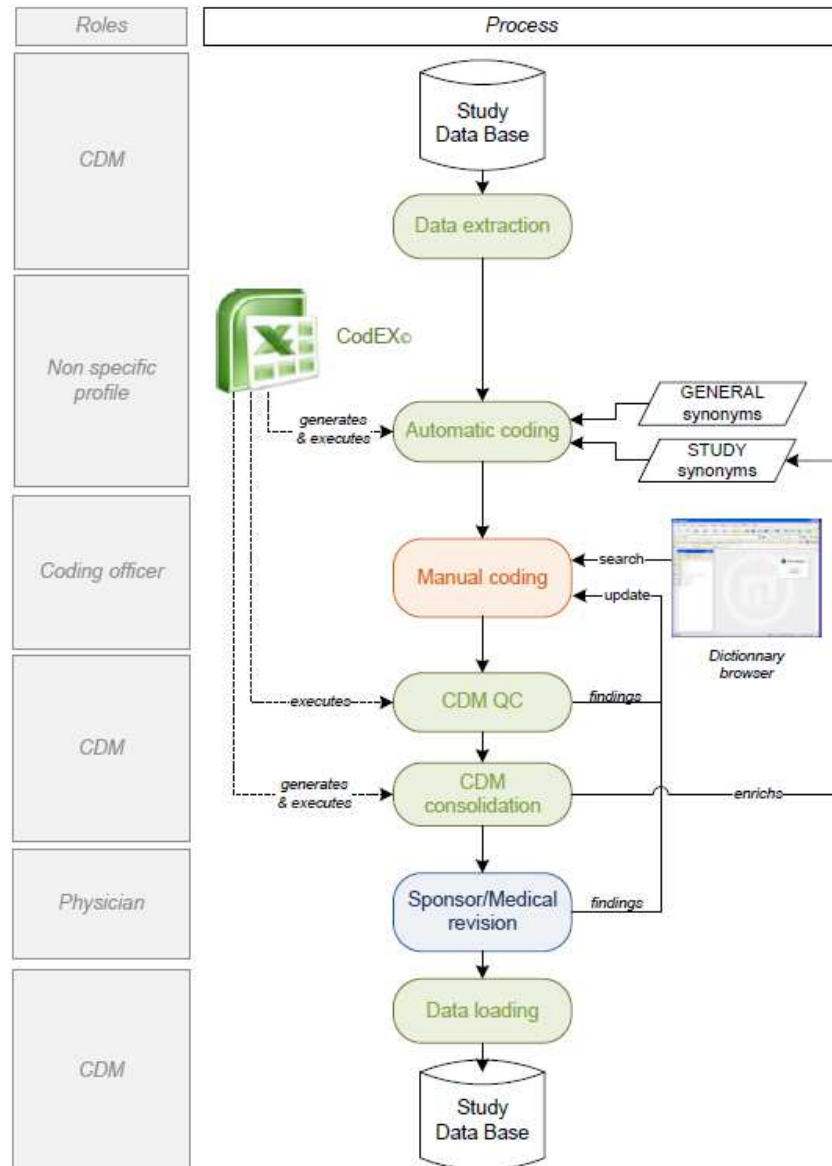
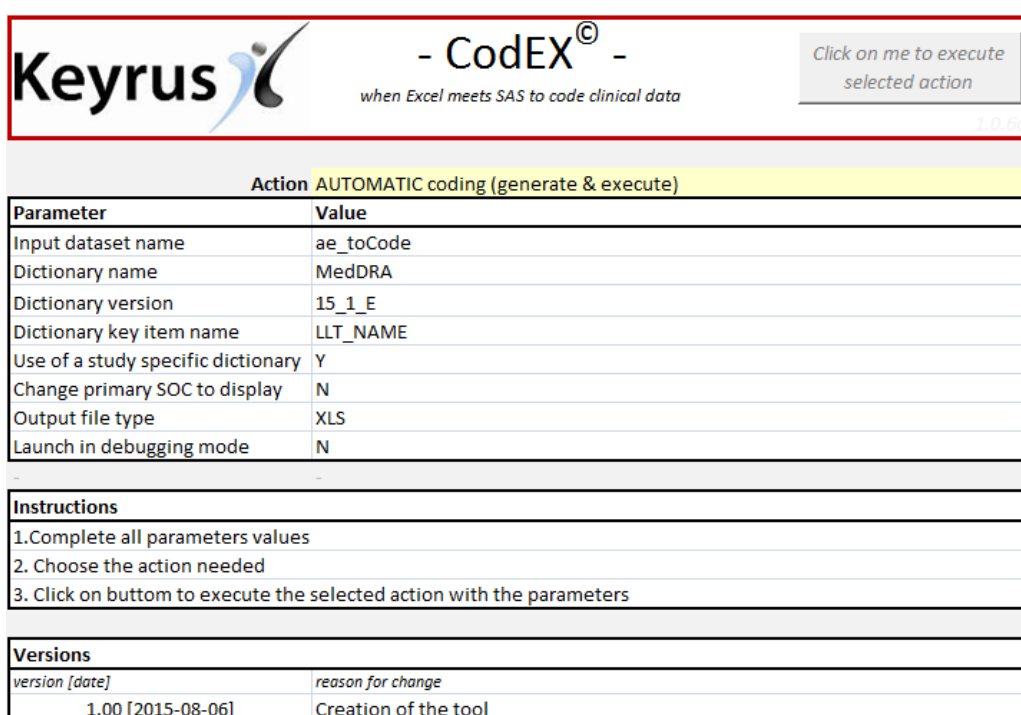


Figure 4: Coding process using CodEX®

We wanted to find a user-friendly interface, and as such, chose to limit the number of parameters to 10 or less, in order to avoid difficulties with conceptualization and management. Because we had chosen Excel for communication between the profiles, we need a programming language on which Excel documents can be created, updated, and so on. As we also need a 'universal' tool, we need a programming system that is already installed and ready to use on most desktop computers. We therefore chose to develop our system with visual basic (VBA) macros.

PhUSE 2015



Keyrus - CodEX[©] -
when Excel meets SAS to code clinical data

Click on me to execute selected action

1.0.00

Action AUTOMATIC coding (generate & execute)

Parameter	Value
Input dataset name	ae_toCode
Dictionary name	MedDRA
Dictionary version	15_1_E
Dictionary key item name	LLT_NAME
Use of a study specific dictionary	Y
Change primary SOC to display	N
Output file type	XLS
Launch in debugging mode	N

Instructions

1. Complete all parameters values
2. Choose the action needed
3. Click on button to execute the selected action with the parameters

Versions

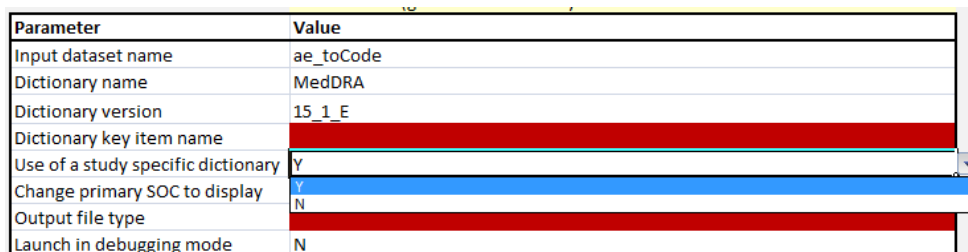
version [date]	reason for change
1.00 [2015-08-06]	Creation of the tool

Figure 5: CodEX[©] interface

Figure 5 shows the interface developed with simplicity in mind. The system is composed of:

- An action part (in yellow),
- A customizable list of parameters,
- An instruction part,
- A versioning part.

Because most of the program's errors/bugs are linked to misuse during production, a lot of these parameters are not able to be edited and are available only in a drop-down menu. This avoids the user entering data which is invalid.



Parameter	Value
Input dataset name	ae_toCode
Dictionary name	MedDRA
Dictionary version	15_1_E
Dictionary key item name	
Use of a study specific dictionary	Y
Change primary SOC to display	N
Output file type	
Launch in debugging mode	N

Figure 6: Conditional formatting shows unanswered parameters, drop-down lists avoid errors in data entry (Y for YES and N for NO)

Conditional formatting is another Excel functionality that is used to highlight errors before going through the process (shown in Figure 6 and Figure 7).

HOW TO USE THE TOOL

CodEX[©] is 'hosted' in an Excel file, so we can easily 'install' the tool in a study-specific folder by copying the template to the folder. The system detects the folder in which the document is stored, and captures all the study external network addresses (such as SAS[®] executables, dictionary data sets, user identification etc.). To avoid any misuse, these parameters belong to a hidden password-protected sheet which only the business owner is able to unlock for update.

PhUSE 2015

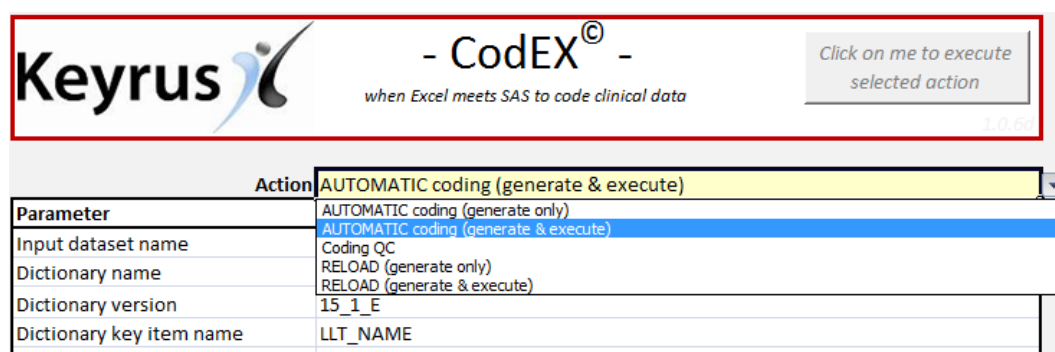


Figure 7: Actions modes

As shown in Figure 3, the tool is mainly composed of 4 modules, each of them using distinct software:

- Data extraction (not predefined, dependent on the project),
- Automatic coding (SAS®),
- Manual coding followed by data management QC (Excel),
- List of synonym updates (SAS®) and data loading (not predefined, dependent on the project).

The CodEX® tool is universal; therefore it needs to be universally connected with the 'external world'. This means that data entry has to be normalized and key items used for synonym matching have to be specifically named and formatted, which leads to the CDM performing a short mapping from their database (this has to be documented in the coding section of the DMP).

Mapping only captures the information useful for coding (e.g. patient country for medication coding), and blinds the data if the manual coding is being outsourced and allows distinct terms to be kept (this is very interesting because the medical officer only has to code each term once with no more risk of duplicate codes for a given term). Mapping also allows a selection of country, center, patient, and visit-related choices for coding.

Once the data to code are present in a dataset in the CodEX® folder location, then the automatic coding can be performed. The system generates and executes the SAS® batch with the parameters provided by the user, and then the run history is generated (see Figure 8). The run history documents each run along with information such as user id, date, time, parameters and number of errors. The run history is available for reading and is password protected to avoid modification.

Because Excel is used to build the SAS® batch and uses the windows system command to launch it, some SAS® errors can occur, which need to be solved by the system owner. This is why we chose to split the SAS® actions in two: one called 'generate only', the other called 'generate & execute'. With the first one, the system owner generates the SAS® batch and is able to launch the batch using SAS® in debug mode to solve the issues. With the second one, SAS® user training is not mandatory since the system generates and launches the batch itself. But in this case debugging is limited to the SAS® log reading.

UserID	Date	Time	Action	Input dataset name	Dictionary name	Dictionary version	Dictionary key item name	Use of a study specific dictionary	Change primary SOC to display	Output file type	Launch in debugging mode	Number of Error
stephane.chollet	27/02/2015	21:06:07	PRE-CODING (generate & execute)	ae_to_code_20feb2015	MedDRA	15_1_E	LLT_NAME	Y	Y	CSV	Y	5
stephane.chollet	27/02/2015	21:07:34	PRE-CODING (generate & execute)	ae_toCode	MedDRA	15_1_E	LLT_NAME	Y	Y	CSV	Y	1
stephane.chollet	27/02/2015	21:12:05	PRE-CODING (generate & execute)	ae_toCode	MedDRA	15_1_E	LLT_NAME	Y	Y	XLS	Y	0
stephane.chollet	27/02/2015	21:14:27	PRE-CODING (generate & execute)	ae_toCode	MedDRA	15_1_E	LLT_NAME	Y	N	XLS	Y	0
stephane.chollet	27/02/2015	21:35:27	PRE-CODING (generate & execute)	ae_toCode	MedDRA	15_1_E	LLT_NAME	Y	N	XLS	N	3
stephane.chollet	27/02/2015	21:36:13	PRE-CODING (generate & execute)	ae_toCode	MedDRA	15_1_E	LLT_NAME	Y	N	XLS	Y	3
stephane.chollet	27/02/2015	21:41:31	PRE-CODING (generate & execute)	ae_toCode	MedDRA	15_1_E	LLT_NAME	Y	N	XLS	Y	0

Figure 8: CodEX® run history

The SAS® program is designed to generate an Excel file with the coding information (Cf. Figure 9). The CDM prepares the file for the coding officer's review by completing the comments section (e.g. with answers to previous queries). Manual coding can be done in the Excel listing.

Once the files are completed, the CDM performs a QC to ensure that:

- No forbidden characters have been used,
- Each distinct term has the same code.

If there are any findings, the Excel file is sent back to the coding officer to be updated.

At this stage, the terms are coded and can be medically reviewed by the study physician or by the sponsor. The file can also be uploaded to the database using reverse mapping and can be used to implement the project-specific list of synonyms.

PhUSE 2015

CONCLUSION FOR STANDARDIZATION

Working on several studies with several structures results in a large amount of repeated work for each project (writing specifications, programming and validation). The purpose of standardization is to perform this job only once for all projects.

We also chose to focus on the standard parts of the coding process, externalizing specific tasks such as input management and SAS® dataset thesaurus conversion management.

Because the automatic coding step makes a match between a term from the clinical database and the dictionary key item from the list of synonyms, with all additional information (on the way this term is computed), the project specificities can be externalized.

Export of data from the clinical database into CodEX® is documented at the study level by the CDM. Data export can be reversed at output, and additional information can be provided to help coding (e.g. country for medications).

The purpose is to limit the energy used to set-up a study specific coding solution. Without CodEX® all SAS® programs need to be specifically updated and validated. With CodEX® development is only focused on data export validation.

CONCLUSION FOR A SINGLE COMMUNICATION TOOL

Excel is a very common software. Everyone has already worked with an Excel sheet (managing, sorting, filtering data etc.) and people can use their personal set-up, thereby customizing the display. Excel is therefore a very good candidate for communication between the profiles.

CodEX® also provides an Excel file which has to be consolidated by the CDM in order to provide information from the study database (additional information such as queries already posted, query answers etc.). By using cell color it is possible to provide clear guidance for which information should be completed:

Coding status	Coding comments	Change primary SOC code to:	Term to be coded	LLT code	LLT name	PT code	PT name	HLT code	HLT name	HLGT code	HLGT name	SOC code (primary)	SOC name (primary)	SOC code(s) (secondary)	SOC name(s) (secondary)
Manual coding required			ABDOMINAL PAIN CRAMP												
Automatic coding (MedDRA 17.1 E)			ACUTE ABDOMEN	10000647	Acute abdomen	10000647	Acute abdomen	10027678	Gastrointestinal signs and symptoms NEC	10018012	Gastrointestinal signs and symptoms	10017947	Gastrointestinal disorders		
Manual coding required			FISSURES HANDS												
Manual coding required			THROMBOEMBOLIC EVENT (DVT)												
Manual coding required			LOST OF APPETIT												
Automatic coding (MedDRA 17.1 E)			NERVOUS	10029197	Nervous	10029216	Nervousness	10002889	Anxiety symptoms	10002851	Anxiety disorders and symptoms	10037175	Psychiatric disorders		
Manual coding required			OBSTRUCTION OF SMALL INTESTINE												
Manual coding required			SENSORY POLYNEUROPATHY												
Manual coding required			SHIVER												
Automatic coding (MedDRA 17.1 E)			UROSEPSIS	10048709	Urosepsis	10048709	Urosepsis	10040054	Sepsis, bacteraemia, viraemia and fungaemia	10021879	Infections - pathogen unspecified	10021881	Infections and infestations	10038359	Renal and urinary disorders
Manual coding required			SKIN RASH (TORSO + ARMS)												
Automatic coding (MedDRA 17.1 E)			NEUROSIS	10029333	Neurosis	10029333	Neurosis	10068751	Anxiety disorders NEC	10002851	Anxiety disorders and symptoms	10037175	Psychiatric disorders		

Figure 9: example of CodEX® output

Before data loading, cleaning is performed to ensure that file structure has not been altered (even if file is password protected) and that forbidden characters have not been used (like carriage return). This allows successful transfer to the SAS® dataset and avoids SAS® misreading. All consistency checks are documented and validated in the tool.

At the end, after QC, a final review can be performed by the study physician in order to validate changes and request updates in case of medical inconsistency. As for all steps, a new file name is attributed to the file, meaning that all changes can be tracked easily.

CONCLUSION

CodEX® was conceptualized by our data managers in response to two key challenges encountered by CRO data management teams when dealing with coded clinical data:

- Standardization of SAS® programs despite the number of different data management systems and database structures,
- A common computer communication language that can be understood by any clinical data expert involved in the coding process.

By the use of SAS® and Excel combination, CodEX® is able to face these challenges.

PhUSE 2015

Now that CodEX[®] is routinely used in our organization, we are convinced that it can also be used as a communication tool between CROs and Sponsors, especially when they are sharing data cleaning activities. For example, if a biotech company wants to keep responsibility for medical monitoring including the coding, CodEX[®] is the ideal solution for managing the split responsibility in the data coding process. By giving the sponsor study physician access to CodEX[®], the manual coding or the coding review can be kept as a sponsor activity. The only pre-requisite for this use of CodEX[®] is that the sponsor must have licences for the coding dictionaries, MedDRA and WHO-DD.

In its current version, CodEX[®] is developed to code with MedDRA and WHO-DD. However, it could also be linked with any type of dictionary, either official (NCI Common Terminology Criteria for Adverse Events, ICD-10 etc.) or sponsor custom-made (allergen codes, solicited AE codes etc.).

Clinical data management, including data coding remains a critical phase in clinical research. CodEX[®] will contribute to the ultimate goal of fast-tracking the drug development process and will ensure the generation of high-quality data for accurate drug evaluation.

REFERENCES

MedDRA = Medical Dictionary for Regulatory Activities
WHO-DD = World Health Organization Drug Dictionary

GCDM October 2013

Ref FDA : US Food and Drug Administration. Reviewer Guidance: Conducting a Clinical Safety Review of a New Product Application and Preparing a Report on the Review. Washington, DC: US Government Printing Office; 2005. Available at: <http://www.fda.gov/CDER/GUIDANCE/3580fml.pdf>. Accessed December 3, 2008.

Ref EMA: Guideline on good pharmacovigilance practices (GVP) Module VI – Management and reporting of adverse reactions to medicinal products
http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/09/WC500172402.pdf

Ref MedDRA: http://www.meddra.org/sites/default/files/guidance/file/intguide_18_0_english.pdf

ACKNOWLEDGMENTS

Dominique Begon
Kirsten Dumaz
Michel Brouillard

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Stéphane Chollet
Company: Keyrus Biopharma
Address: 18-20, rue Clément Bayard
City / Postcode: F-92300 Levallois-Perret - France
Email: stephane.chollet@keyrus.com
Web: <http://www.keyrusbiopharma.com/>

Author Name: Catherine Scoupe
Company: Keyrus Biopharma
Address: Chaussée de Louvain, 88
City / Postcode: B-1380 Lasne - Belgium
Email: cathy.scoupe@keyrus.be
Web: <http://www.keyrusbiopharma.com/>

Brand and product names are trademarks of their respective companies.