sciendo

RURAL SUSTAINABILITY RESEARCH
47(342), 2022
ISSN – 2256-0939



Received: 1 June 2022 Revised: 29 June 2022 Accepted: 04 July 2022

Research Support Model for Improving the Effectiveness of Medical Study Data Collection

Signe Balina[1,2], Edgars Salna[2], Ilona Kojalo[3], Eliza Avotina[3]
[1] University of Latvia, 19 Raina Blvd., Riga, Latvia
[2] Datorzinibu centrs Ltd, 41 Lacplesa street, Riga, Latvia
[3] Institute of Clinical and Preventive Medicine, University of Latvia, 19 Raina Blvd., Riga, Latvia

Abstract. The paper describes the research support model for improving the effectiveness of the medical research data collection process and data quality. Every research project involves a data collection phase, during which different organisation, legal and technology factors are involved, including various procedures (questionnaire design, annotation, database design, data entry, data validation, discrepancy management, medical coding and data mining). The key task of clinical data management is to obtain high-quality data, which can be achieved by minimising data input errors and timely identifying missing data. This process is often time-consuming and takes up a significant part of the research project budget in both veterinary and human medicine. The aim of this study is to elaborate the research support model for the creation of a data collection automation software tool, which will allow one to ensure better data quality, shorten the time for data collection and minimise human work volume and respective human resource expenses, making research projects more effective in terms of their timing and budget. Research work included analysis of the current situation, its shortcomings, typical research project budget distribution and existing automated electronic data collection tools (EDC). Research was carried out in partnership with the Institute of Clinical and Preventive Medicine of the University of Latvia.
Key words: codebook, medical research support model, electronic research data collection.

Introduction

When initiating the research project, researchers need to create a study design that includes information about the hypothesis to be studied, the patient control group to be used in the study and its selection criteria, data field units, data logistics, and ethics (Ranganathan & Aggarwal, 2018). At the time of protocol creation, it should be clear how the required data will be integrated into the database, how they can be compiled, and whether they will be analysed. Questionnaires should be simple and understandable, and the questions should be designed to be directly relevant to the research objectives. Usually, at least the questionnaire outline will be drawn up in parallel with the research protocol design. Most grant review committees expect the pilot funding application to include such a presentation (Smith, Morrow, & Ross, 2015). It is important for the development of the protocol to take into account the development aspects of the database and the data recording mechanism, and for the shortcomings identified during the development of the database to be addressed by making the necessary adjustments to the research protocol.

The research protocol design process is time-consuming and not particularly feasible for researchers, who are not information technology specialists. There is also a high risk of data entry inaccuracies. After collecting data, it is also necessary to establish study database and document structures, which vary depending on the type of study. Each study can be carried out entirely independently, and each study can also define the data sets needed for itself with its own encodings and its data acquisition, processing, and analysis mechanisms. The set of tools for data collection, processing, and analysis to be used for research is mainly determined by the competencies and practices of specialists involved in the project, as well as by developments that allow research teams to carry out several studies of the same study area. However, it should be noted that such an approach

* Corresponding Author's email: signe.balina@lu.lv

DOI: 10.2478/plua-2022-0009
© 2022 S.Balina et al. This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

leads to technological stagnation by reducing the use of new methods in data extraction, processing, and analysis processes.

Although the use of one of the data coding schemes already defined in Latvia is not imposed on research projects, the use of nationally approved encodings makes it more likely to use data from existing data sources. For example, the classification of diseases can be defined as a code system defining the nosological units according to the criteria selected.

The study requires the authorisation of the Ethics Committee, the patient's signed consent form, as well as the confidentiality certificate, signed by the researcher. The General Data Protection Regulation lays down strict rules for the processing of data based on consent. The purpose of these provisions is to guarantee that a natural person understands what he or she agrees with. Consent must be given by affirmative action, for example by checking the box online or by signing the form. When a particular person agrees to his or her personal data being processed, they may only be processed for the purposes for which the consent was given.

Each project has its period of execution. Data collected within the framework of the project and not used should be destroyed. However, if there is a need to reuse this data in another project, the new consent of the patient and the authorisation of the Ethics Committee is required.

The data needed for the study project can be collected in the form of a survey (anonymous or non-anonymous) by conducting an electronic survey or a paper-based questionnaire. In the case of biomedical studies, sampling of biological material from the patient is carried out, or the data required for the study is requested from the relevant authority, which, in the case of Latvia, is the Centre for Disease Prevention and Control, which issues them electronically in specific format.

If a research project requires specific data that is not available in national databases, it is necessary to carry out a personalised data capture process within the framework of the study. The personalised data capture process is sufficiently complex and requires both serious preparation and accurate and consistent execution, based on the research protocol developed during the preparatory phase. When initiating the acquisition of the personalised data of the study, the responsible medical staff shall contact patients and provide them with the necessary information so that the patient can look into and understand what they are being asked for and why. In this context, it is important to go into the study itself, which is explained by the medical staff or specialist concerned.

It should be noted that the data sets from different sources, including data from patients, data from different types of studies and dynamic observation data are needed for the execution of the research project. All data acquisition and conversion costs are to be covered by the research project budget. Therefore, when planning a research project, the potential data sets and costs needed to obtain the desired data sets should be clearly identified. It is also necessary to evaluate whether the volume and quality of the data will enable one to reach the project objectives.

Research projects can be evaluated based on their budget and hypothesis to demonstrate "sufficient" data. Evidence of a study hypothesis can define the size of the data set, which, in the opinion of the researchers, is sufficient to demonstrate the hypothesis of the research project. By including a larger number of patients and more comprehensive data, the results of the study can be more accurate and qualitative, but, given the trends in cost growth, it is desirable to find the most economically viable compromise.

Another important aspect is the quality of the data. When collecting or enriching the data set required for the study, the relevance of the data sources to the needs of the project and the integrity of the other data included should be assessed. It should be noted that if the part of the data set to be added is not compatible with the data quality requirements of the research project, it might be necessary to rethink the need for this data set in the project or to consider obtaining these data in an alternative way.

It can be assumed that the use of data collection automation tools allows the effectiveness of the data collection process to be increased, significantly reducing the overall data collection stage length, researcher involvement volume and associated costs. These tools can provide IT support for the following stages of the process – (1) study initiation phase (the result of the phase is the research protocol), including generating templates and partly automated filling with available data, (2) registration of a study into IT infrastructure, (3) entry and storage of study data (includes personal data, survey results and data units from external systems).

Electronic Data Capture (EDC) applications provide new opportunities for reusing stored information. In 2011, a study of the effectiveness of EDC solutions compared to methods based on paper questionnaires (Walther *et al*., 2011) demonstrated both significant achievable human resource and time savings. The problem is topical, and other studies are being conducted around the world to improve the research process (Grunzke *et al*., 2019).

The aim of this study is to develop a research support model, the novel feature of which is the use of the codebook catalogue to provide a function of knowledge accumulation and transfer between

research projects, as well as to improve data quality by implementing data standardisation and validation mechanisms. The codebook catalogue with included codebook elements allows the formal recording and re-use of parts of the research project datasets based on clinical guidelines or nationally defined standards, with the possibility to include information about the potential source of data. The model will later be used as the basis for the definition of new electronic data collection tool requirements and prototype development.

Materials and Methods

Clinical data management should be seen as an essential phase of each clinical trial, and this phase should result in qualitative and sufficient data to carry out the study. This helps to drastically reduce the amount of time to make the diagnosis, time for treatment and in other clinical aspects. Various procedures, including questionnaire development, database design, data entry, validation, coding and extraction, non-compliance measures, should be executed when starting the trial.

The main objective of clinical data management is to obtain qualitative data by minimising the error rate and data insufficiency to the greatest possible extent, and by pooling the largest possible dataset for analysis. Medical coding is necessary in order to identify and classify the relevant medical terminology. Technically, this requires knowledge about medical terminology, diseases, and medications used, as well as an understanding of disease processes. Another required knowledge area relates to the structure of the medical dictionary to be used (Krishnankutty *et al*., 2012).

In human medicine, each research project requires its own risk analysis, analysis of innovation, knowledge and datasets, the consent of an ethics commission, a new form of patient consent, and the provision of data security. Similar to human medicine, in veterinary medicine research there are also ethical considerations (Cartlidge, 2012) to comply with - animal welfare as a whole, informed owner's consent, minimum reasonable number of animals to guarantee study consistency, necessary authorisations if using the invasive acts.

Existing experience shows that the research project procedure consists of several parts. When considering the experience at the beginning of the project, a medical research protocol is produced, and this protocol analyses the results of the treatment and post-operative complications of patients included in a biobank. Tracking of data in national registers is also performed. In order to create such a protocol, a multidisciplinary team representing specialists from different areas is required, so the basic task in the early stages of the protocol is to gather a work team that will be able to perform all of the following assignments (see Fig. 1).

There are many areas where it is possible to improve the data collection process by offering researchers some guidelines or a "road map" which, depending on the type of study, can lead the researcher along a certain path, reducing the probability of errors and the time needed to collect data. In general, data for research purposes fall into three major categories:

- Data obtained through the survey process. The project teamwork identified a formalised set of data which needs to be acquired through a patient survey. This does not mean that data extraction is not possible in any other way, but that the project team can identify potential shortcomings in an external source (such as inadequate data quality or non-matching encoding), or that it is not possible to provide an administrative process for data extraction, or that data extraction is less costly using the survey process.

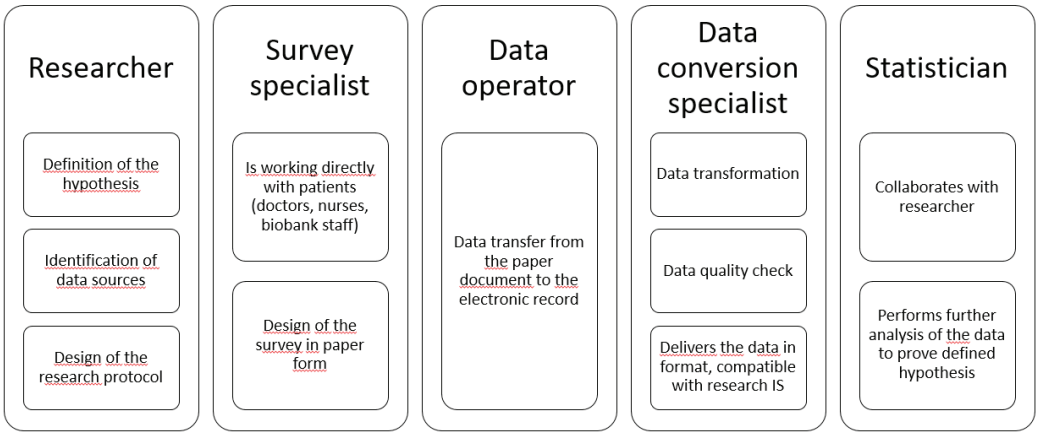


Figure 1. Research project team at data collection stage.

- Data derived from previous studies. In cases when the current study is a continuation of a previous study or a parallel sectoral study, the dataset obtained in the previous study may be reused directly or with an acceptable degree of transformation in the current study. The project team should decide on the possibilities for reuse, as well as on data compatibility and compliance with the data requirements of the current study.
- Data that can be obtained from information systems. Available data for patient datasets, for example, Electronic Patient Record in veterinary medicine or Electronic Health Record in human medicine. If the research project team has decided that the data needed for the research project can be obtained from an external data source, the process of outlining the administrative steps for obtaining these data shall be initiated.

Data reuse also has a significant impact on research process effectiveness and project costs. According to Kerby (2015), over two-thirds of research projects do not include it. The most common type of reuse (25%) is using the data from a colleague. Some articles about research projects (Djoko & Pardo, 2013) pointed to reusing data from the researcher's own previous studies, for example, where the existing data from some study was analysed differently within another projects.

Some researchers (Holden, McDougald, & Hoonakker, 2014) have also noted other challenges that may hamper the process of successful data collection. It may be the lack of trust and understanding between researchers and the patients or their family members; cognitive limitations, problems with logistics, travel distances, recruitment and compensation, scientific quality and interpretation, issues of data validity and reliability, as well as combining data from multiple sources.

While some of these challenges cannot be successfully addressed using IT tools, others may be solved and/or reduced by implementing survey forms which are customised for the particular research project, enabling online data entry with the help of tablets and other mobile devices or by reusing datasets from previous research projects. To improve the effectiveness of data collection, efforts were directed into two major areas – codebooks for survey research and data enrichment from external data sources.

In order to create a research database, first it is necessary to understand what kind of data will be included in it. The data must be described in accordance with the research project protocol data definition, and a codebook (Lavrakas, 2008) is created for this purpose. The codebook provides information about the structure, content, and layout of the data. A well-documented codebook describes each variable included in the database. Although codebooks vary greatly in their quality and the amount of information provided, a typical codebook includes the location and width of the columns for each variable – definitions of different types of records, response codes for each variable, codes used to indicate non-response and missing data, the exact questions used in the survey and the samples omitted, and other indications of the size and characteristics of each variable. The codebook describes the structure of the research data file.

The nosology (Arslan, 2019) data can be divided into three parts. The first part shall include all of those variables which are potentially repeated in several nosologies: for example, the patient's basic data, gender, age, weight, height. The second part includes data such as habits, allergies, genealogical history, where different nosologies may use differently formulated questions or where a different amount of information should be obtained or the answer choices are different. The third part represents the data that are unique to a particular nosology, for example, in the research of lung cancer and head and neck cancer, which were used as a reference in the research.

The core idea is that when a set of data is defined in a research project, it is reasonable to stick to the codebook elements that are already defined historically (both in the context of nosology and apart from it), and to only create a new codebook element or a new version of the existing codebook element if a codebook element for the specific dataset is not found.

When examining the existing research project practices, we have used the data about actual ongoing or recently completed research projects. One of the selected projects was currently running the oncosurgery patient research at the Clinical and Preventive Medicine Institute in collaboration with the Oncosurgery Clinic of Riga East Clinical University Hospital (marked as Project #1 in Table 1). The results of treatment of patients were included in the biobank. The data, including post-operative complications associated with diseases, symptoms, provision of medical care and medical manipulation, are analysed according to the defined research protocol. The analysis process also includes tracking of data in national registers, for example, analysis of the data included in the cancer register and the pathology archive. The aim of the project is to obtain information about the patients, treated at Riga East Clinical University Hospital with certain diagnoses (malignancy and control group), and their data was included into the biobank. This information allows data about medical manipulations and treatment results to be summarised.

The project uses the patient data, already accumulated in the biobank, and all patients should give their written informed consent. Data from all participants in the study are analysed - both data already included in the biobank and the data that will be included in the biobank by the end of the study. The study is conducted and implemented according to the applicable regulations of the European Union and the Republic of Latvia regarding the protection and security of personal data. Personal information is kept strictly confidential and is not publicly available. The following personal data are collected, stored and processed - date of birth, nationality, sex, diseases diagnosed. The source of personal data is the data of the patient questionnaires and data present in the medical documents.

The data from two other projects ("Gastric cancer screening with breathing volatile organic compounds using a hybrid sensing approach" (marked as Project #2 in Table 1) and "Optimisation of the H. pylori eradication scheme for mass gastric cancer prevention measures" (marked as Project #3 in Table 1)) were also available for the purposes of analysis.

These projects are a typical example of this field of research and can contribute greatly to the detailed understanding of the current approach to the conduct of the research project, focusing on the acquisition process and the establishment of a research protocol.

On the basis of these three projects, a typical data collection process diagram was produced (see Fig. 2).

The conduct of the study can be divided into several areas. Looking at the experience of study initiation, a research protocol is being developed to analyse the results of the treatment and post-operative complications of patients included in the biobank. To create a protocol, a multidisciplinary team, capable of representing specialists from different areas, is required. Thus, the basic task of the early stage of the protocol is to establish a work team that will be able to perform all the activities described above.

The research support model, described in this paper, addresses two key parts of the data collection process – protocol design and creation of the database, codebooks and data registration mechanism. To create a study database, it is necessary to understand what data will be included in it. The codebook approach was selected for this purpose. An important step of this process is to evaluate whether all the data is compatible with any of the predefined codebooks.

An important aspect during the automation of the research data collection process is its economic feasibility. Analysis of three independent studies of the Institute of Clinical and Preventive Medicine of the University of Latvia was carried out to determine the tasks that can be streamlined using the approach defined by the research support model. The positions where the research costs are most likely

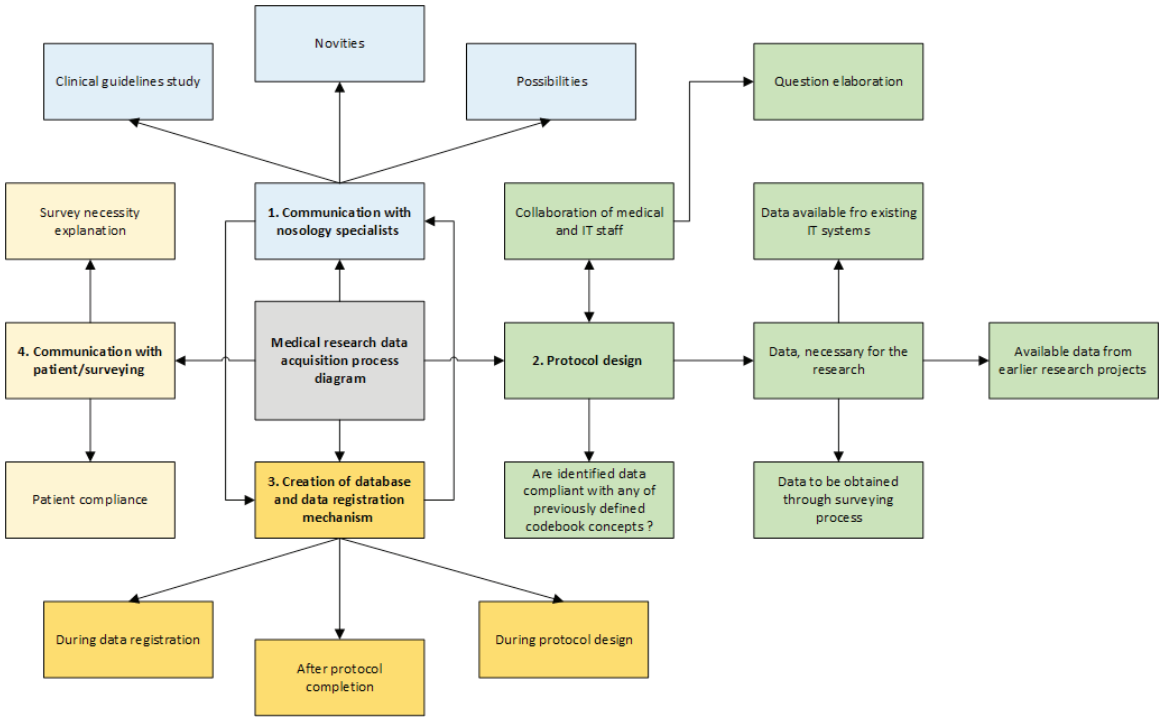


Figure 2. Medical research data collection process.

Table 1

Comparison of resource involvement

Protocol development steps	Project # 1	Project # 2	Project # 3
	hours	hours	hours
Research and establishment of the project team	248	68	20
Communication between protocol development team	208	120	16
Codebook development	80	56	20
Database development	320	68	28
Coordination of the draft protocol with the project partners	184	104	12
Development of necessary updates and re-coordination with project partners	96	56	0
Communications	1162	100	36
Total, man-hours:	2298	572	132

to be influenced are the collaboration between the protocol development team and the study database development.

Comparative resource involvement breakdown for all three projects is given in Table 1.

It should be clear that the number of hours for protocol development within each project is different and it is not possible to unambiguously describe and estimate the cost reduction as a result of using the research support model-based automation. Based on researcher remuneration data it can be concluded that in Latvia the total costs of activities related to the research protocol may vary from EUR 4,000 to EUR 50,000. Assuming that the implementation of an EDC tool may save at least 20–30% of the time spent, financial savings can easily exceed 10,000 euros in larger size projects.

Analysis of existing EDC applications was performed to assess their functional capabilities and to produce a set of optimal requirements for the development of a new tool, capable of significantly improving the level of data collection process automation.

Results and Discussion

Electronic data collection (EDC) software is gradually adopted for clinical trials and biomedical, pharmaceutical and healthcare research. EDC applications also provide new opportunities for the reuse of stored information.

A group of scientists from the US, along with their Korean and Chinese colleagues, developed a data-based system in order to get a clinical insight into virtually usable data of intervention radiology. A single, universal pre-processing data bank was developed where a database (Cohn *et al.*, 2020) was created to store the data collected by clinical researchers.

The data flow process between clinical source application systems and research application systems (RAS) is particularly important in order to avoid errors and inaccuracies. A publication (Camacho, Staubert, & Lobe, 2016) describes the method whereby clinical information is electronically imported from a LIMS (laboratory information management system) to RAS applications such as tranSMART and OpenClinica. This method makes it possible to transfer clinical data with minimal human participation. A 2011 study (Walther *et al.*, 2011) of the effectiveness of EDC solutions in comparison to paper-based methods demonstrated significant reductions in both time and cost effectiveness.

Electronic Data Capture (EDC) has significantly changed the research data collection process. However, spreadsheet software is still used on many occasions as a patient data storage tool. In one study (Staziaki *et al.*, 2016), the effectiveness of EDC was evaluated and compared to standard spreadsheets in the data collection time and data accuracy level, demonstrating that data collection was faster than using a spreadsheet, which made it easier to track the subjects.

Information sharing and data matching are essential for doctors and scientists as they gather and analyse data from different hospitals and other medical institutions in order to establish registries and carry out multicentre clinical research. This requires a standardised metadata exchange based on common

data models. An article (Breil *et al.*, 2012) describes the idea of creating a medical data model portal (MDM-Portal) based on ODM files. The portal provides a system-independent multilingual data model store in ODM format that can be used by doctors.

Just like in the field of collecting medical data where many applications have been introduced, in the field of electronic data collection there is also a debate on the need to establish a common framework for collecting data from medical studies. Taking into account the large number of clinical research conducted worldwide and the complexity of these studies, the need to share their results as well as their structure and metadata (Calbimonte *et al.*, 2017) has become very topical. This article presents MedRed ontology, which is widely used in displaying metadata for clinical research.

The problem of setting up EDC systems has been addressed for a long time. Already in 2007 it was concluded (Welker, 2007) that although the pace of progress in data collection methods is increasing overall, the transition to EDC is slow. The use of EDC systems should improve data integrity, and save the costs and time taken to study the conditions for creating a database.

Many research projects need to obtain and analyse data from different sources, often from different institutions and with differently coded information. One article (Hazlehurst *et al.*, 2015) describes the Comparative Effectiveness Research HUB (CER-HUB), a web-based informatics platform for research development and performance that combines electronic clinical data from a number of healthcare organisations. Comparative Effectiveness Research requires specific methods for obtaining and analysing complex multi-institutional clinical data. By offering a series of informatics tools integrated into the research framework using the Electronic Health Record (EHR) data, CER Centre provides a solution to the challenges of many institutions that use electronic medical documentation data. Although EHR collection generates a significant amount of electronic clinical data, these data are not available directly due to various social and technical reasons.

In the course of this research, based on the information available in producer resources and scientific publications, the four most popular EDC tools used in the field of medical research were analysed, comparing their characteristics, advantages and disadvantages (Table 2).

Table 2

EDC software feature comparison

Feature	Castor EDC	Dacima CS	OpenClinica	MedReD
Randomisation	X	X	X	X
Data capture	X	-	X	X
Offline data entry	-	X	-	-
Reports	Research statistics	Built-in reports	Memos, reports	Built-in reports
Data import and export	X	X	-	X
SOAP & RESTful services	-	-	X	-
Form designer	X	-	-	-
Research protocol designer	-	X	-	-
Advanced design	-	X	-	-
Patient surveys	X	-	-	-
Patient result reports	-	-	X	X
Clinical data management	-	-	X	-
Two-factor authentication	X	-	-	X
Source data validation	X	-	-	-
Real-time validation	-	X	-	-
Dynamic for activation	-	X	-	-
Codebook support	-	-	-	-

After examination of the MedRed tool, it was concluded that the cost would be very high at the stage of its deployment, and that the development process could take quite a long time. Similarly, preparing a package of documents (description of the necessary forms, translation, defining database fields and types) could take considerable time and human resources, and requires discussions with the developer on the timing of the integration of new forms (study/study protocols). A significant drawback of the tool is the need to transfer sensitive personal data outside the territory of the Republic of Latvia.

Castor EDS does not provide the ability to combine the dataset and its corresponding input form using the CodeBook elements. It is cloud-based and provides multi-tenant support. However, such an approach means that it is not possible for a research project team to control the location and access of the data.

Dacima Clinical Suite has fairly similar features to the previously mentioned Castor EDS, but with a significant difference - the solution has an on-premises version and can be installed on the research organisation servers. Decimal Clinical Suite also offers the reuse of previously created data entry forms and provides strong support for generating and managing data input forms. A fundamental drawback of this product is the fact that it does not support the use of existing CodeBook-type elements.

The OpenClinica tool provides the widest range of capabilities. Unlike the other solutions analysed, OpenClinica is closest to the implementation of the CodeBook concept, allowing one to prepare small sets of data (form template) in a timely manner and store them in the content library for reuse. As a negative point, it should be noted that the manufacturer has decided to only provide support after 2022 for a commercial version of the product, which will lead to significant costs for researchers of up to US $2,000 a month.

It was concluded that none of these tools has necessary features to be used in the case of a proposed Codebook-based research data collection and management scenario and the decision was made to work on defining requirements for the new tool, addressing the researcher's needs, and creating and validating its prototype. This work is scheduled for a later stage of this study.

Automated data collection tools can only improve part of the project implementation steps. Each research project requires its own risk analysis, analysis of innovation, knowledge and datasets, the consent of an ethics commission, a new form of patient consent and the provision of data security. However, if information technology can help to speed up the creation of a research protocol, it also helps to prepare an ethical application more quickly, thus saving working hours and reducing costs.

The research support model was introduced (Fig. 3) to increase study effectiveness. It is based on the typical medical data collection process (Fig. 2) and highlights the most important activities and the artefacts, which are necessary to carry out activities and results. The model development process represents a breakdown of the data collection process into detailed components, introducing the codebooks as a key element of the research protocol. According to the model, all study activities can be split into 4 stages - initial research, research protocol, data acquisition phase, data retrieval phase. The planned data collection automation tool should increase process effectiveness during the Research protocol phase and Data acquisition process phases.

Initial research within each project includes specialist knowledge, research opportunities, innovation, research into the latest clinical guidelines and also the generation of new ideas. Since evidence-based medicine is used in Latvia, it is important to constantly keep up with the latest guidelines, which are published once a year or less often. (Znotiņa I., 2014). Based on the guidelines of nosology, it is necessary to create a clinical survey for the patient, because it is important to understand what information needs to be obtained in order for further action in the study to be properly justified.

During development of the model, a need was identified for a component that performs the accumulation of existing experience and binding regulations and ensures its transfer to the study team. This solution should include a search function within the accumulated experience, prototyping, creation of new metadata artefacts and the replenishment and re-use of existing metadata artefacts. The component was named the "Codebook Catalogue" and is integrated into the Research protocol phase of the model. Some of the artefacts created in the research protocol phase represent the data set required for the data mining process. These artefacts must be formalised in order to ensure the generation of a database, the creation of a manual data entry interface. It is also necessary to provide a data retrieval mechanism by applying transformations in a way that the data to be retrieved are as close as possible to the data format required by the analytical solution. The prototyping feature of the codebook catalogue solution is expected to reduce misunderstandings during the research protocol phase and speed up the preparation of research protocol documentation.

The database development process includes the design of the project flowchart and definition of the data types, field lengths, name descriptions (*codebook*) for a particular project. The database development process

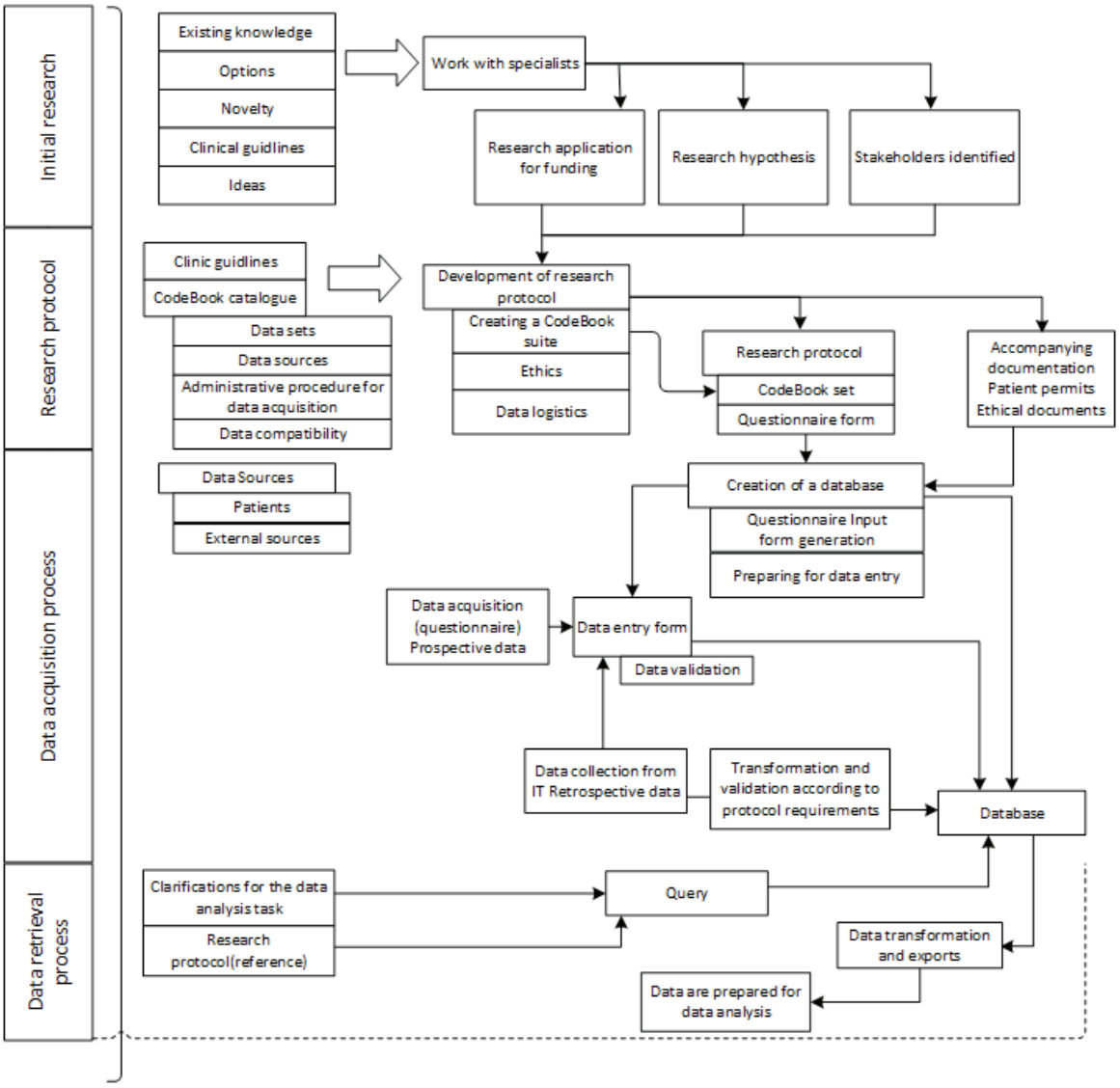


Figure 3. Research support model.

can be automated, and this viewpoint is justified by an ideologically technological approach to "CodeFirst" (https://www.entityframeworktutorial.net/code-first/what-is-code-first.aspx). This approach provides that physical database development takes place on the basis of a data model, i.e., using a well-established logical database model to automatically generate the build-up of a physical database. This approach has gained widespread popularity in the programming process and not only allows you to generate database objects, but also provides a database interface layer by providing "CRUD" operations (Create Read Update Delete). Consequently, data sets in a research project can be used to generate a database if they are properly recorded and presented.

Once the entire project structure has been approved, a questionnaire data entry form is generated. If the data are already collected in tables or spreadsheets, the information is appropriately described and prepared for import into the database. The processing of the information in the tables depends on both the type of data source and the type of data. Before entering data, it is important to check its adequacy, especially if the data originates from several sources. The data retrieval process involves exporting the data stored in the database for further processing. Before exporting the data, queries are created in which the data necessary for the execution of the research protocol is inserted.

The data can also be exported and used for statistical analysis or patient situation prediction using deep machine learning and the usage of artificial intelligence solutions. Potential application areas there are tumour segmentation, histopathological diagnosis,

tracking tumour development, and prognosis prediction. (Londhe & Bhasin, 2019).

Preparation of research project data is one of the most labour-intensive phases, as demonstrated by the analysis of three typical research projects at different levels.

Inventing the research process support model provides an opportunity to accumulate the experience of existing research projects in a formalised way, which then makes it possible to use this experience in future projects to reduce the resources required for project execution. Using research data support tools will also make it easy to follow changes in the data used when data is collected from external data sources that will be able to monitor data in real time with the tools. Re-use of data will be improved.

In the implementation of clinical medicine research projects, data acquisition and preparation is one of the most labour intensive stages, as evidenced by the analysis of typical research projects of various scales in the field of oncology. Improvement and acceleration at this stage by using IT support tools can help to achieve significant cost and time savings. Traditional coding methods are prohibitively expensive and do not use modern digital infrastructures, which allow for the efficient collection of comprehensive drug data from participants (Lockery *et al*., 2019).

The use of the research support model allows one to formalise the research project process and to accumulate the experience gained during the project for its re-use in future research projects. The model is universal and it can be successfully applied to medical research projects in different areas, in both human and veterinary medicine; the main variable components of the model are codebooks, which may be specific for each study area. Therefore, this phase of operation improvement and acceleration through IT support tools can provide important cost and time savings. The research support model will serve as the basis for the specification of EDC tool prototype software requirements.

Conclusions

The research produced IT support model for medical research data collection automation. This model is based on the data collected from several medical research projects in the field of oncology and describes all the key elements of the data collection process.

Analysis of actual human resource involvement in the data collection process has shown that depending on the overall project size, 0.54 to 6.7 person months or 35 to 60 percent of the total time spent developing the study protocol relates to the activities, which may be sped up by using software tools for the design of surveys, data collection, database design, data validation and team communications.

The assessment of several well-known existing electronic data collection tools leads one to the conclusion that there are significant deficiencies, which do not allow significant savings to be achieved in terms of researcher involvement time. The most common and important deficiencies are a lack of support for data entry and storage templates or codebooks, a cloud-based approach, which makes it difficult to properly handle the sensitive personal data or high costs, associated with the adaptation or use of particular software.

The created model includes the use of codebooks as a new and important element of research protocol design and data acquisition phases and addresses regarding both the data collection process within the current research project and data acquisition from external data sources (local or national registers, other research projects), including monitoring of changes in these data.

The model was later used to define software requirements for the new data collection tool and building the tool prototype. Validation of the prototype has shown that it can allow the researchers to spend less time collecting the data, achieving 20 to 30 percent time economy depending on the project.

References

Arslan, R. C. (2019). How to Automatically Document Data with the Codebook Package to Facilitate Data Reuse. *Advance in Methods and Practices in Psychological Science.* 2(2), 169-187. DOI: 10.1177/2515245919838783.

Breil, B., Kenneweg, J., Fritz, F., Bruland, P., Doods, D., Trinczek, B., Dugas, M. (2012). Multilingual Medical Data Models in ODM Format: A Novel Form-based Approach to Semantic Interoperability between Routine Healthcare and Clinical Research. *Applied Clinical Informatics.* 3(3), 276-89. DOI: 10.4338/ACI-2012-03-RA-0011.

Calbimonte, J. P., Dubosson, F., Hilfiker, R., Cotting, A., Schumacher, M. (2017). *The MedRed Ontology for Representing Clinical Data Acquisition Metadata*. Retrieved September 15, 2021. From https://iswc2017.ai.wu.ac.at/wp-content/uploads/papers/MainProceedings/205.pdf

Camacho, R. J., Stäubert, S., Löbe, M. (2016). Automated Import of Clinical Data from HL7 Messages into OpenClinica and tranSMART Using Mirth Connect. *Studies Health Technology and Informatics*. 228, 317-21.

Cartlidge, H. (2012, July). *Collecting data and designing experiments*. Retrieved February 14, 2022. From https://www.theveterinarynurse.

com/review/article/collecting-data-and-designing-experiments

Cohn, D. M., Mabud, T. S., Arendt, V. A., Souffrant, A. D., Jeon, G. S., An, X., … Rubin, D. L. (2020). Toward Data-Driven Learning Healthcare Systems in Interventional Radiology: Implementation to Evaluate Venous Stent Patency. *J Digit Imaging*. 33(1), 25-36. DOI: 10.1007/s10278-019-00280-6.

Djoko, S. S., Pardo T. A. (2013). Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly.* 30 (1), S19-S31. DOI: 10.1016/j.giq.2012.06.011.

Grunzke, R., Hartmann, V., Jejkal, T., Kollai, H., Prabhune, A., Herold, H., ... Nagel, W. E. (2019). The MAS repository service - Comprehensive, metadata-driven and multi-community research data management. *Future Generation Computer Systems*, 94, 879–894. DOI: 10.1016/j.future.2017.12.023.

Hazlehurst, B. L., Kurtz, S.E., Masica, A., Stevens, V. J., McBurnie, M. A., Puro, J. E. … Sittig, D. F. (2015). CER Hub: An informatics platform for conducting comparative effectiveness research using multi-institutional, heterogeneous, electronic clinical data. *International Journal of Medical Informatics*. 84 (10), 763-773. DOI: 10.1016/j.ijmedinf.2015.06.002.

Holden, R. J., McDougald, A. M., Hoonakker, P. L. T., Carayon, P. (2014). Data collection challenges in community settings: Insights from two field studies of patients with chronic disease. *Quality of Life Research*. 24, 1043-1055.

Kerby, E. (2015). Research Data Practices in Veterinary Medicine: A Case Study. *Journal of eScience Librarianship*. 4(1): e1073. DOI: 10.7191/jeslib.2015.1073.

Krishnankutty, B., Bellary, S., Kumar, N. B. R., Moodahadu, L.S. (2012). Data management in clinical research: An overview. *Indian Journal of Pharmacology*. 44 (2): 168-72. DOI: 10.4103 / 0253-7613.93842.

Lavrakas, P. J. (2008). *Encyclopedia of Survey Research Methods.* Retrieved September 15, 2021. From https://methods.sagepub.com/reference/encyclopedia-of-survey-research-methods/n69.xml

Lockery, J. E, Rigby, J., Collyer, T. A, Stewart, A. C, Woods, R. L., McNeil, J. J., … Ernst, M. E. (2019). Optimizing medication data collection in a large-scale clinical trial. *PloS One*. 14 (12), e0226868. DOI: 10.1371/journal.pone.0226868.

Londhe, V. Y., Bhasin B. (2019). Artificial intelligence and its potential in oncology. *Drug Discov Today*. 24 (1), 228-232. DOI: 10.1016/j.drudis.2018.10.005.

Ranganathan, P., & Aggarwal, R. (2018). Study designs: Part 1 - An overview and classification. *Perspectives in clinical research*. 9 (4), 184–186. DOI: 10.4103/picr.PICR_124_18.

Smith, P. G., Morrow, R. H., Ross, D. A, editors. (2015). *Field Trials of Health Interventions: A Toolbox*. 3rd edition. Oxford (UK): OUP Oxford. DOI: 10.1093/med/9780198732860.001.0001.

Staziaki, P. V., Kim, P., Vadvala, H. V., Ghoshhajra, B. B. (2016). Medical Registry Data Collection Efficiency: A Crossover Study Comparing Web-Based Electronic Data Capture and a Standard Spreadsheet. *Journal of Medical Internet Research*. 18(6):e141. DOI: 10.2196/jmir.5576.

Walther, B., Hossin, S., Townend, J., Abernethy, N., Parker, D., Jeffries, D. (2011). Comparison of Electronic Data Capture (EDC) With the Standard Data Capture Method for Clinical Trial Data. *PLoS One*. 6 (9): e25348. DOI: 10.1371/journal. pone.0025348.

Welker, J. A. (2007). Implementation of electronic data capture systems; barriers and solutions. *Contemporary Clinical Trials*. 28(3), 329-36. DOI: 10.*1016*/j.cct.2007.01.001.

What is Code-First? Entity Framework Tutorial. Retrieved May 4, 2022 from https://www.entityframeworktutorial.net/code-first/what-is-code-first.aspx.

Acknowledgement

The research leading to these results was funded by the European Regional Development Fund project "Competence Centre of Information and Communication Technologies", (1.2.1.1/18/A/003), individual research project 1.5 "Development of research support model for medical data collection process improvement".