Contemporary Clinical Trials 30 (2009) 300–316

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Comparison of paper-based and electronic data collection process in clinical trials: Costs simulation study

Ivan Pavlović [a,b], Tomaž Kern [c], Damijan Miklavčič [a,*]

[a] *University of Ljubljana, Faculty of Electrical Engineering, Tržaška 25, SI-1000 Ljubljana, Slovenia*
[b] *Institute for Project Management and Information Technology, Kotnikova 30, SI-1000 Ljubljana, Slovenia*
[c] *University of Maribor, Faculty of Organizational Sciences, Kidričeva cesta 55a, SI-4000 Kranj, Slovenia*

ARTICLE INFO

Article history:
Received 25 November 2008
Accepted 22 March 2009

Keywords:
Clinical trial
Data collection
Electronic data collection
EDC
Internet
Business process
Process models
Costs
Simulation study

ABSTRACT

An alternative to clinical trial paper-based data collection (PDC) is internet based electronic data collection (EDC), where the investigators over the internet enter data directly in the electronic database by themselves. In our study we considered clinical trial as a business process. Our objective was to model PDC and EDC process and to estimate the difference of the costs of PDC and EDC process for a sample clinical trial based on these models.

We used Extended Event-driven Process Chains (eEPC) modeling technique to model PDC and EDC process. In order to evaluate the costs of the processes we assigned costs functions to each process function which appears in the model. The parameters which appear in these functions include efforts, staff prices and data quality parameters. We estimated the values of all these parameters and performed costs calculations for a sample clinical trial.

Through an analysis and modeling efforts we identified sub-processes which contain main differences affecting duration and costs of the PDC and EDC process: data gathering at the research center; monitoring; and data management. The most significant model difference between PDC and EDC process appeared in data management sub-process. For the sample clinical trial considered in our simulation study and our parameters estimations the EDC process decreased data collection costs for 55%. For different scenarios of parameters variations we show that the EDC process may bring from 49% to 62% of savings when compared to PDC process.

1. Introduction

Clinical trial (CT) is “any investigation in human subjects intended to discover or verify the clinical, pharmacological and/or other pharmaco-dynamic effects of one or more medicinal product(s)” (for full definition see [1]). CT can be carried out in either one or multiple sites. One of the core documents in CT is the Case Report Form (CRF). The CRF is a form where the investigator enters all patients' clinical and non-clinical data related to the trial. This is like a CT dossier of the patient. Data collected in the CRF are both: data related to the patient health parameters; and data related to the medical procedure performed in the trial.

The data collected in the paper forms have to be entered in the electronic database in order to perform computer data analysis. For this purpose investigators usually send the copies of paper CRF to the data center where data managers enter these data into the database. This paper-based routine has many disadvantages which result in erroneous data in the database and longer duration of clinical trial (especially for the large multi-centre clinical trials) [2,22].

An alternative to paper data collection is internet based electronic data collection (EDC), where the investigators over the internet enter data in the electronic database by themselves. In this way the errors from copying data from paper forms to electronic database by the person who did not

[*] Corresponding author. Fakulteta za elektrotehniko, LBK, Tržaška 25, 1000 Slovenia. Tel.: +386 1 4768 456; fax: +386 1 4264 658.
E-mail addresses: ivan.pavlovic@fe.uni-lj.si (I. Pavlović), tomaz.kern@fov.uni-mb.si (T. Kern), damijan.miklavcic@fe.uni-lj.si (D. Miklavčič).

1551-7144/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.cct.2009.03.008

collect the data are avoided. Also, electronic Case Report Forms (eCRF) usually contain check routines which reduce erroneous data entry. Another advantage of EDC is that the data managers have continuous insight into the data and the data collection process and thus can manage the whole CT process better. Good overviews of the advantages of electronic data collection were provided by Welker [3] and Brandt et al. [4].

Despite the fact that the EDC tools have been available for more than two decades, clinical trials however are still mainly conducted using paper data collection as the primary tool (according to [5] in 2004 it was still over 75%). The reason for this was partially ascribed to the fact that technological applications often did not have adequate functionality to improve the data collection process as a whole and that applied technology was expensive to introduce and maintain. In recent years, a development of internet technologies and its availability reduced many technological obstacles. Many authors namely report successful implementation of internet based EDC solutions [6–14]. However, without being able to predict the costs, duration and quality of processes performed, a change from PDC to EDC is still a risky decision.

In order to facilitate decision between PDC and EDC use in clinical trial (CT), with respect to quality and cost of the process, in our study we considered CT as a business process. The CT as a business process is complex and includes different sub-processes and activities like: protocol development; protocol use and implementation in the CT experimentation; data collection; and the evaluation of the CT results. Each sub-process and activity has different objectives and is performed in different environments, carried out by its own agents and resources, and governed by specific rules. Some results of the efforts to develop a comprehensive model of the entire CT process were already published by Luzi et al. [15,16]. More specific models which deal only with CT data flow were published by PhRMA (Pharmaceutical Research and Manufacturers of America) [17]. The technological improvements may bring major changes to CT data collection process through the change from paper data collection (PDC) process to electronic data collection (EDC) process. Therefore, we focus our work on data collection process with its sub-processes (data gathering, monitoring and data management). In this paper we present the results of our modeling of PDC and EDC process as well as the estimations of the difference of the costs of PDC and EDC process for sample clinical trial.

2. Methods

There are several business process modeling techniques. Among others, these include eEPC (Extended Event-driven Process Chains), BPEL (Business Process Execution Language) and UML (Unified Modeling Language). We decided to use eEPC modeling technique because it is “user perspective” and allows to model both IT (Information Technology) supported functions as well as “manual” functions.

On Fig. 1 we present the main elements of an eEPC graph, which we used to model EDC and PDC process. This example shows that each function in the process is triggered by some event and must end with another event (which usually triggers another function). Different human or organizational resources may be involved in different ways in completing the function. Also some (input or output) documents as well as supporting computer systems may appear in the process. Beside the elements presented on Fig. 1, an eEPC graph also may contain some branching or merging elements like “∧” (AND), “∨” (OR) or “×” (XOR).

A detail description of eEPC model elements can be found in ARIS documentation [18].

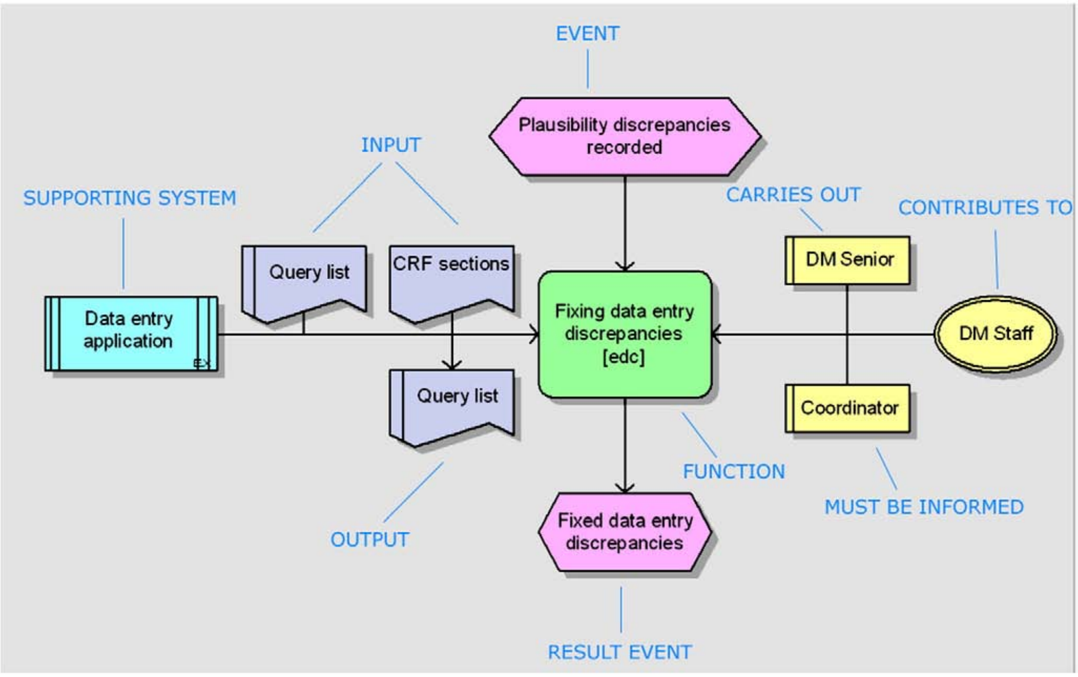

Fig. 1. eEPC graph elements.

2.1. Process models

We decided to model EDC and PDC process, because simple analysis would only yield specific results (for specific case). The model however gives us the opportunity to insert different values in a parameterization and sensitivity study.

We are aware of the fact that every clinical trial has its own characteristics and that each CT sponsor might have different

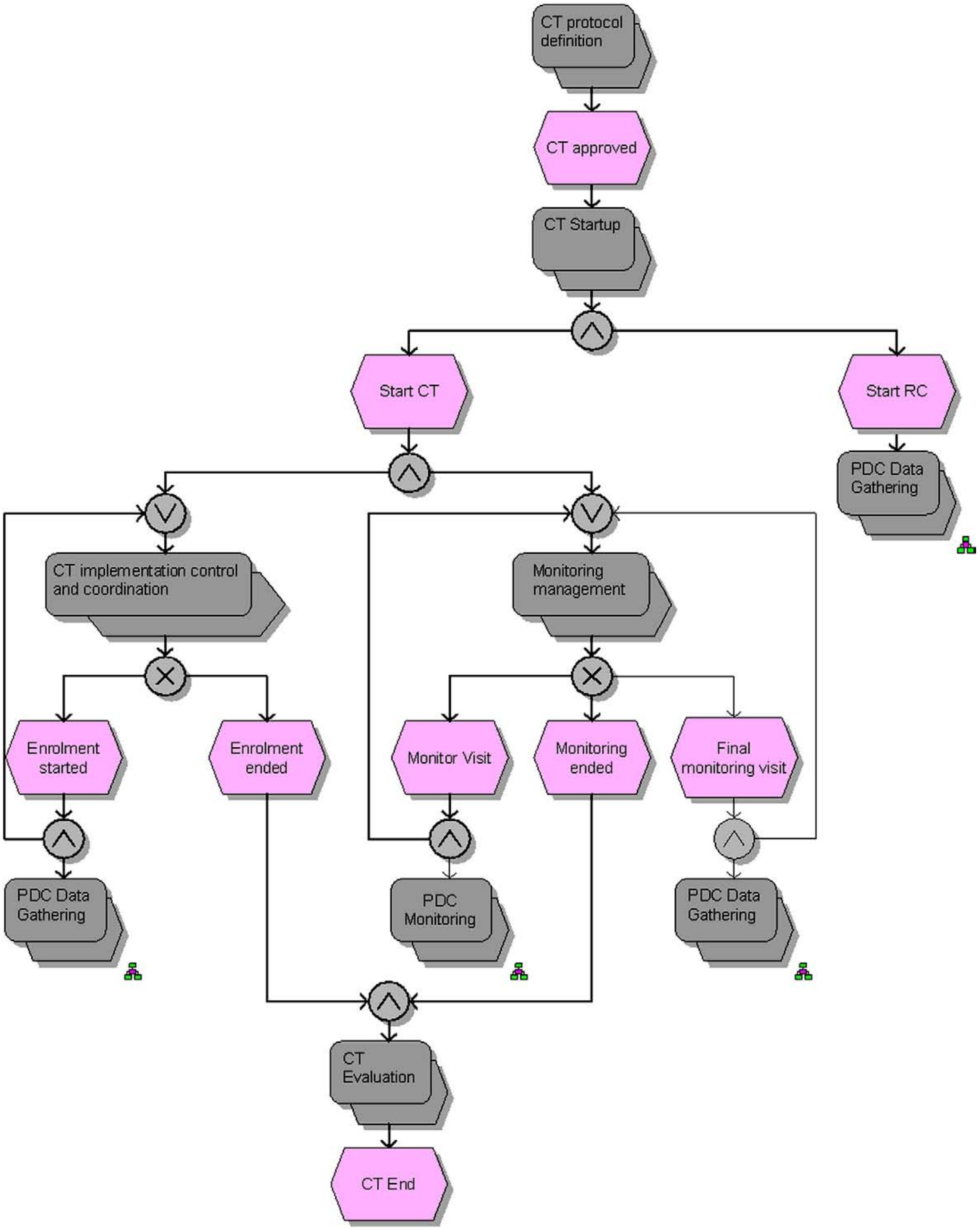

Fig. 2. PDC CT Management model. EDC model is the same as the presented PDC model, being different only in sub-processes. (CT — clinical trial, RC — research center).

approach to data collection. For paper-based data collection these include sending Case Report Forms (CRF) to sponsor by investigator after monitor (also known as Clinical Research Associate — CRA) verifies data, or sending CRF by monitor, or sending CRF by fax prior to monitoring, etc. However, we decided to model PDC and EDC processes which are widely accepted, having in mind that PDC and EDC processes both have to be designed to reach the same data quality.

We therefore made an overview of the literature and we performed a set of interviews with key participants in PDC and EDC process (investigator, monitor, study coordinator, data manager) followed by e-mail correspondence. In the models we included functions which assure data quality and improve performances of each of the two approaches (PDC and EDC).

Here we present eEPC (Extended Event-driven Process Chains) models that we developed in order to compare two approaches to clinical trial (CT) data collection: paper data collection (PDC) model and electronic data collection (EDC) model. For both data collection approaches CT management can be presented by the same model as presented in Fig. 2.

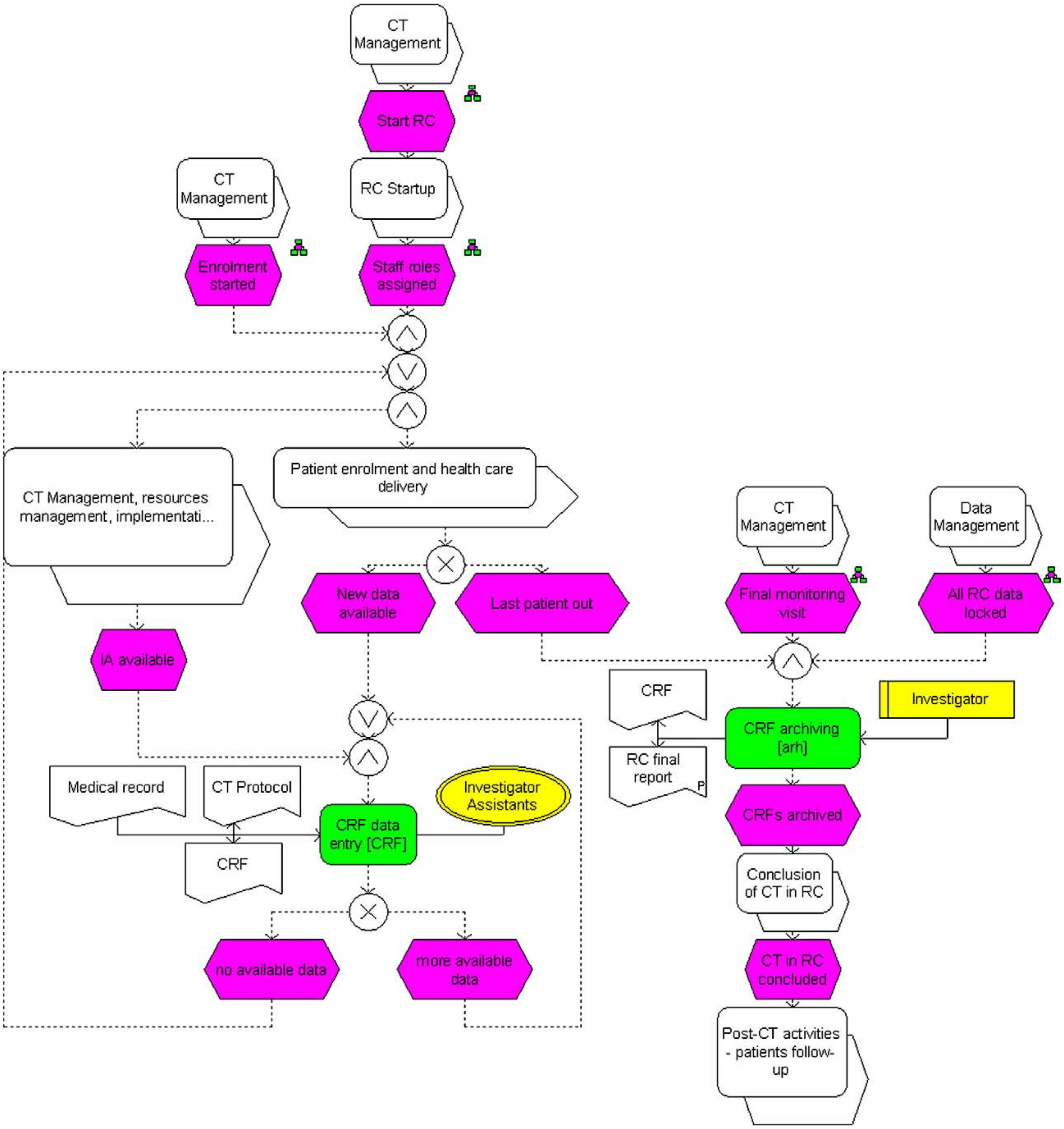

Fig. 3. PDC Data Gathering model (CT — clinical trial, RC — research center, CRF — Case Report Form, IA — investigator assistant).

In our study we focused on sub-processes which contain main differences affecting duration and costs of the processes. These are:

- data gathering at the research center (PDC Data Gathering vs. EDC Data Gathering);
- monitoring (PDC Monitoring vs. EDC Monitoring); and
- data management (PDC Data Management vs. EDC Data Management).

In the following subsections these sub-processes are described in details.

2.1.1. Data gathering at the research center

One of the key roles of the research center is to provide all the important information as required by CT Protocol and associated documentation. Data have to be gathered with high accuracy and delivered to the sponsor in a timely manner. With high level of generalization we may present the process of data gathering at the research site as depicted in Fig. 3 for PDC and Fig. 4 for EDC.

2.1.2. Monitoring

In order to improve the quality of data at the research center a sponsor assigns monitors to visit research centers

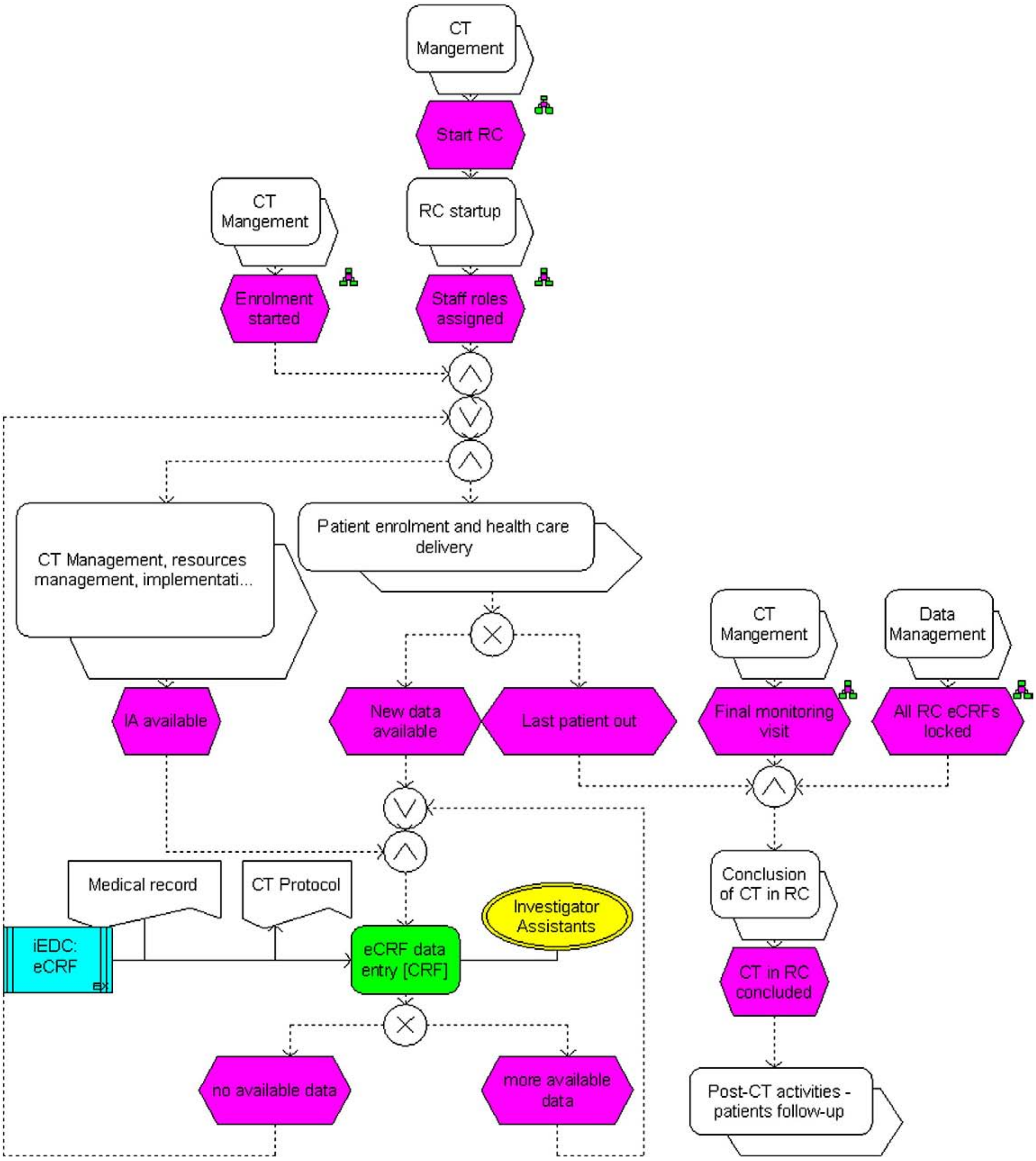

Fig. 4. EDC Data Gathering model (CT — clinical trial, RC — research center, eCRF — electronic Case Report Form, iEDC — electronic data collection application).

intermittently and, in addition to other tasks, check how the data collection is progressing with the aim of improving data quality. The change in the way data are collected (paper vs. electronic) also has considerable impact on monitoring process itself (Figs. 5 and 6).

2.1.3. Data management

Data gathered at the research center and verified by monitor are provided to data center where data management staff takes care of data provided by all the participating research centers and perform all the necessary data analysis. In order to be able

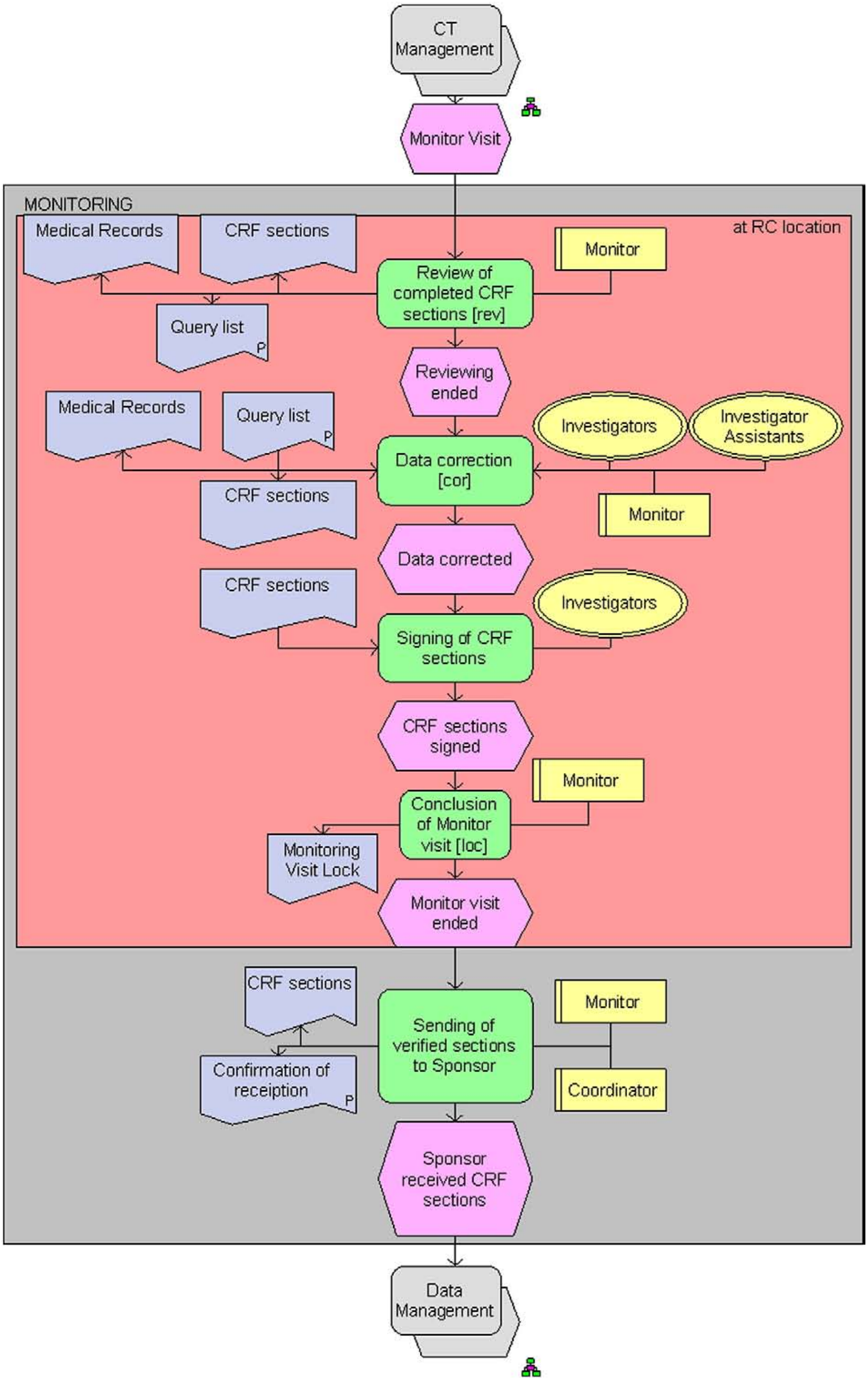

Fig. 5. PDC Monitoring model (CT — clinical trial, CRF — Case Report Form).

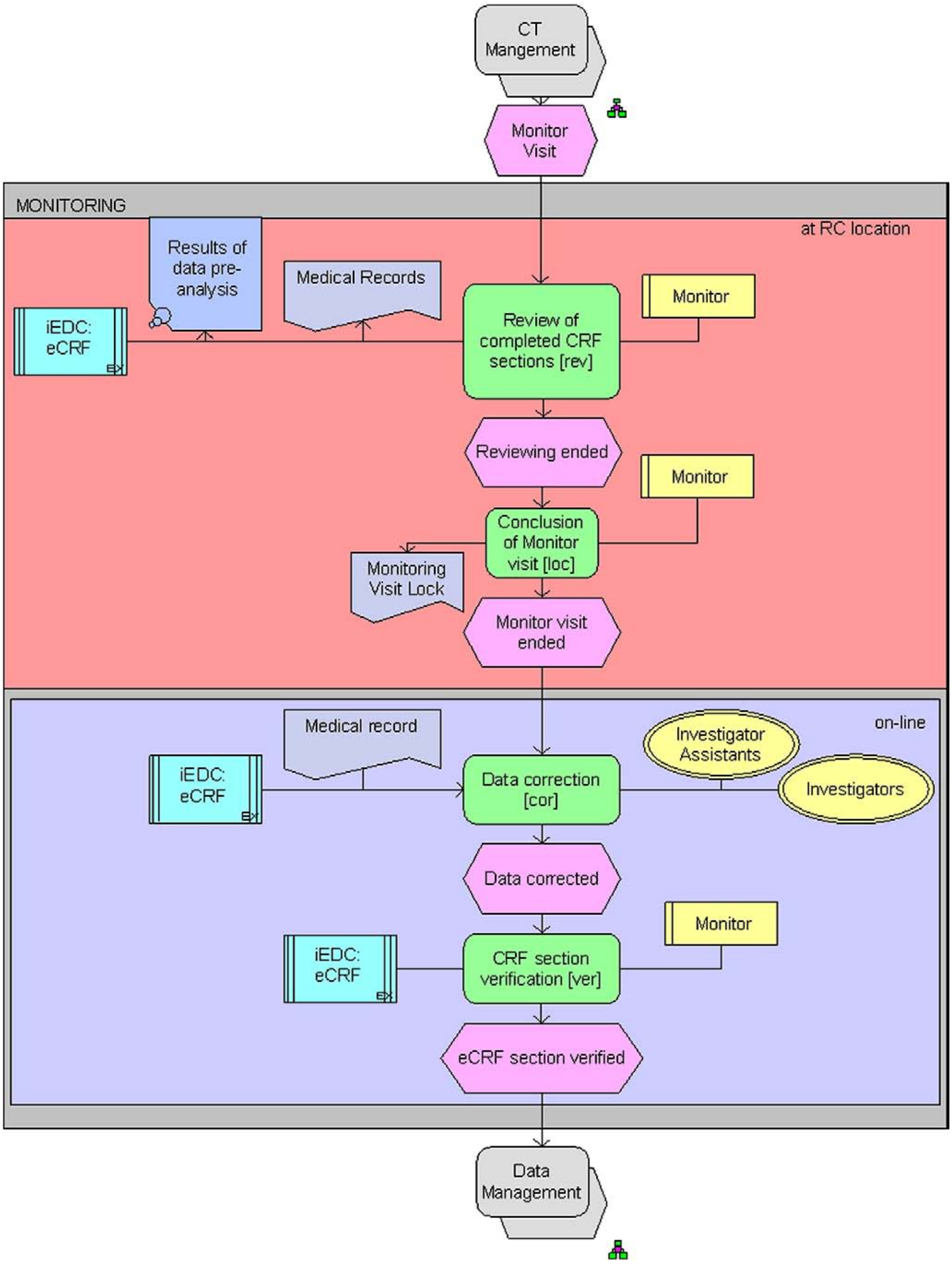

Fig. 6. EDC Monitoring model (CT — clinical trial, eCRF — electronic Case Report Form, iEDC — electronic data collection application).

to perform data analysis data have to be computerized, and are most often stored in central database. Therefore, PDC Data Management process must also include digitalization of data provided on paper forms while EDC Data Management may focus exclusively on data cleaning. The entire data management process is presented in Fig. 7 for PDC and in Fig. 8 for EDC.

2.2. Model parameters

The key property of PDC and EDC process which we wanted to compare was the cost of the process. The costs of the business process are difficult to evaluate as they are affected by the staff efforts, the price of the staff, the price and amortization of the other resources used (such as equipment, materials, etc…), and numerous other economical parameters.

In our analysis we do not consider the costs of the clinical trial process which are not directly related to data collection process (e.g. patient recruitment, or health delivery). Furthermore, we do not consider the costs of the technical resources used in the processes (e.g. the price of the EDC system) or the amortization costs. We focus only on the staff costs and try to recognize how these depend on the process

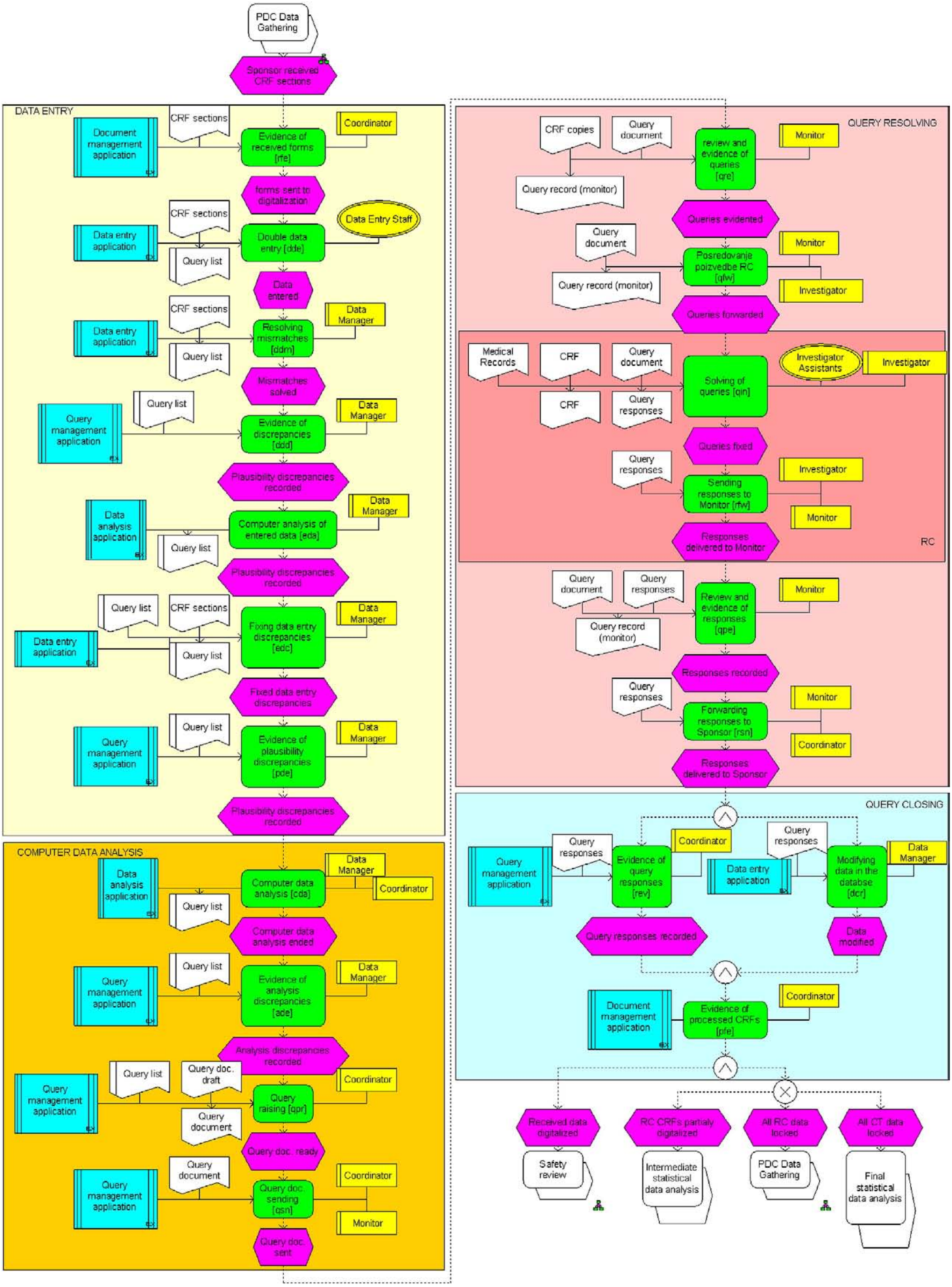
PDC Data Gathering
Sponsor received CRF sections
DATA ENTRY
Document management application
CRF sections
Evidence of received forms [rfe]
Coordinator
forms sent to digitalization
Data entry application
CRF sections
Double data entry [dde]
Data Entry Staff
Query list
Data entered
Data entry application
CRF sections
Resolving mismatches [ddm]
Data Manager
Query list
Mismatches solved
Query management application
Query list
Evidence of discrepancies [ddd]
Data Manager
Plausibility discrepancies recorded
Data analysis application
Computer analysis of entered data [eda]
Data Manager
Query list
Plausibility discrepancies recorded
Query list
CRF sections
Data entry application
Fixing data entry discrepancies [edc]
Data Manager
Query list
Fixed data entry discrepancies
Query management application
Query list
Evidence of plausibility discrepancies [pde]
Data Manager
Plausibility discrepancies recorded
COMPUTER DATA ANALYSIS
Data analysis application
Computer data analysis [cda]
Data Manager
Coordinator
Query list
Computer data analysis ended
Query management application
Query list
Evidence of analysis discrepancies [ade]
Data Manager
Analysis discrepancies recorded
Query management application
Query list
Query doc. draft
Query raising [qpr]
Coordinator
Query document
Query doc. ready
Query management application
Query document
Query doc. sending [qsn]
Coordinator
Monitor
Query doc. sent
QUERY RESOLVING
CRF copies
Query document
review and evidence of queries [qre]
Monitor
Query record (monitor)
Queries evidented
Query document
Posredovanje poizvedbe RC [qfw]
Monitor
Investigator
Query record (monitor)
Queries forwarded
Medical Records
CRF
Query document
Solving of queries [qin]
Investigator Assistants
Investigator
CRF
Query responses
Queries fixed
Query responses
Sending responses to Monitor [rfw]
Investigator
Monitor
Responses delivered to Monitor
RC
Query document
Query responses
Review and evidence of responses [qpe]
Monitor
Query record (monitor)
Responses recorded
Query responses
Forwarding responses to Sponsor [rsn]
Monitor
Coordinator
Responses delivered to Sponsor
QUERY CLOSING
Query management application
Query responses
Evidence of query responses [rev]
Coordinator
Data entry application
Query responses
Modifying data in the databse [dcr]
Data Manager
Query responses recorded
Data modified
Document management application
Evidence of processed CRFs [pfe]
Coordinator
Received data digitalized
RC CRFs partialy digitalized
All RC data locked
All CT data locked
Safety review
Intermediate statistical data analysis
PDC Data Gathering
Final statistical data analysis

Fig. 7. PDC Data Management model (CT — clinical trial, CRF — Case Report Form, RC — research center).

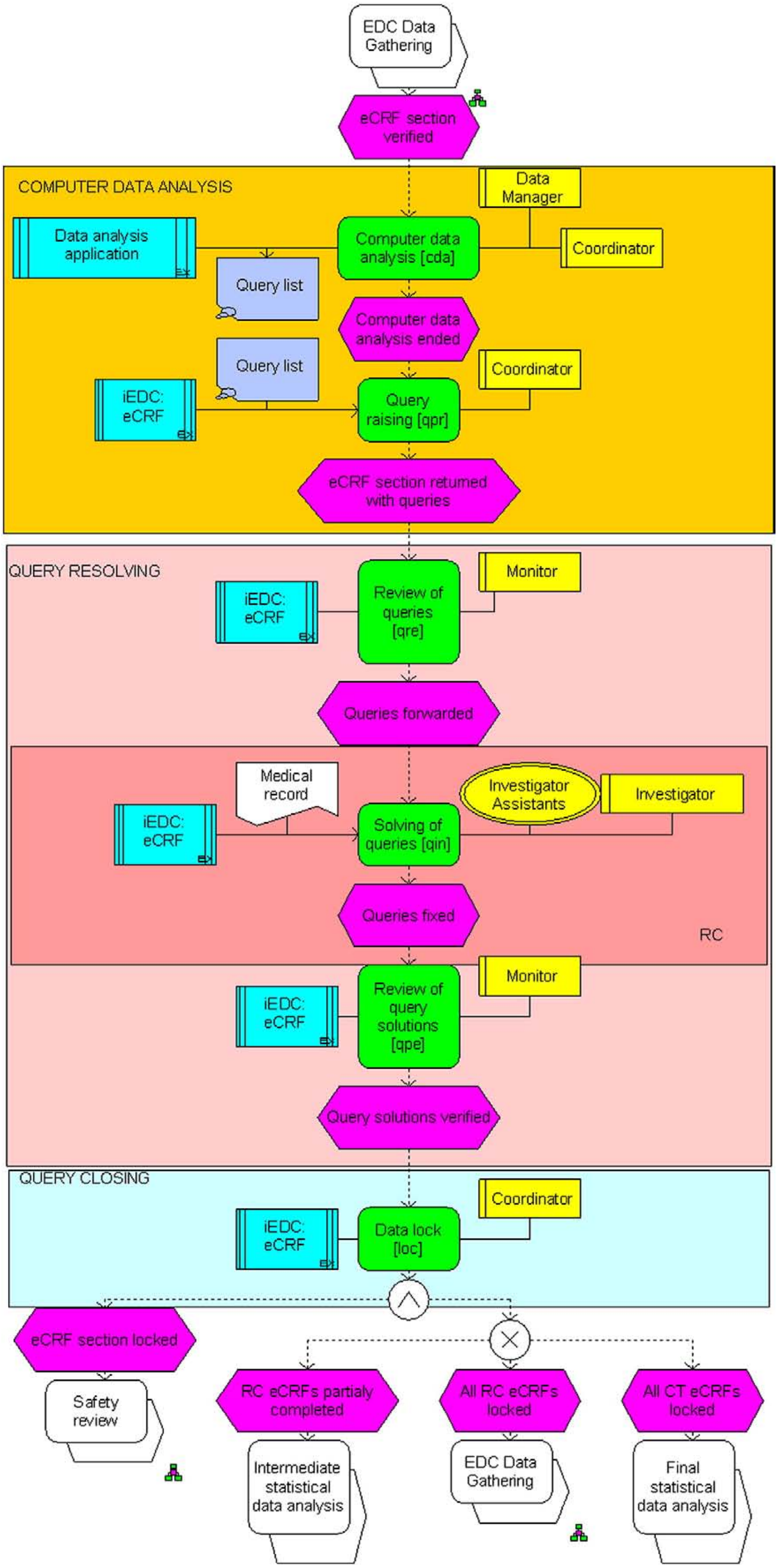

Fig. 8. EDC Data Management model (CT — clinical trial, eCRF — electronic Case Report Form, RC — research center, iEDC — electronic data collection application).

we choose, PDC or EDC. The staff costs depend on staff prices and the efforts required by each of the tasks included in data collection process. Therefore, in order to evaluate the costs we consider the following groups of parameters: efforts, staff prices and data quality parameters.

2.2.1. Efforts

The task (function) effort is the time that human resources (participants in the process) have to dedicate to the task in order to complete it. If we exclude all waiting times, this time includes: the preparation time; the execution time; and the conclusion time. In this analysis we simplify the calculation by considering all these three times together as task duration and trying to assess it for each task separately. In our equations the efforts are represented with the symbol F.

For each of the tasks, when calculating task duration, we took in consideration if any of the parameters of the clinical trial affects the task duration (e.g. the number of data in the CRF, or the number of CRF documents, etc.). Therefore, we introduced average efforts as a measure of time needed for task execution of one instance of the related parameter (e.g. time needed for writing down one data into paper CRF). In our equations we represent the average efforts with the symbol μ.

2.2.1.1. Data gathering at the research center. From the information obtained in interviews, we estimate that the average data entry speed is 5 data per minute for PDC and 10 data per minute for EDC. Faster EDC data fill in is possible due to the fact that EDC usually offers lists of values (from look-up tables/vocabularies/code lists) and can auto-calculate some fields and, in general, it is quicker for average computer user to type the text then to hand-write it. Therefore, we use the following estimations in our calculations:

- writing down data in CRF – $\mu_{\mathrm{DG}^{\mathrm{P}}_{\mathrm{CRF}}} = 12\,\frac{\mathrm{s}}{\mathrm{data}}$
- entering data in eCRF – $\mu_{\mathrm{DG}^{\mathrm{E}}_{\mathrm{CRF}}} = 6\,\frac{\mathrm{s}}{\mathrm{data}}$
- archiving paper CRF – $\mu_{\mathrm{DG}^{\mathrm{P}}_{\mathrm{arh}}} = 12\,\frac{\mathrm{min}}{\mathrm{doc}}$.

2.2.1.2. Monitoring. During data verification monitor has to note down all doubtful, or erroneous data and this is done differently for PDC and EDC (e.g. EDC application enables entering comment/warning next to doubtful data field, while in PDC monitor has to write down on a paper a notice which contains comment/warning together with the reference to the field in the form). Therefore we can consider that the effort for taking a note of possible error is twice bigger for PDC then for EDC ($\mu_{\mathrm{M}^{\mathrm{P}}_{\mathrm{rer}}} = 2\mu_{\mathrm{M}^{\mathrm{E}}_{\mathrm{rer}}}$).

Based on information obtained from interviews we use the following average efforts estimation in our calculations:

- noting down an error (PDC) – $\mu_{\mathrm{M}^{\mathrm{P}}_{\mathrm{rer}}} = 1\,\frac{\mathrm{min}}{\mathrm{data}}$
- noting down an error (EDC) – $\mu_{\mathrm{M}^{\mathrm{E}}_{\mathrm{rer}}} = 30\,\frac{\mathrm{s}}{\mathrm{data}}$
- checking a data – $\mu_{\mathrm{M}_{\mathrm{chk}}} = 30\,\frac{\mathrm{s}}{\mathrm{data}}$
- locking monitor visit – $\mu_{\mathrm{M}_{\mathrm{loc}}} = 2\,\frac{\mathrm{min}}{\mathrm{visit}}$
- correcting data – $\mu_{\mathrm{M}^{\mathrm{P}}_{\mathrm{cor}}} = \mu_{\mathrm{M}^{\mathrm{E}}_{\mathrm{cor}}} = 1\,\frac{\mathrm{min}}{\mathrm{data}}$
- signing CRF section (include verifying) – $\mu_{\mathrm{M}^{\mathrm{P}}_{\mathrm{sig}}} = 1\,\frac{\mathrm{min}}{\mathrm{doc}}$
- sending CRF section by monitor to sponsor – $\mu_{\mathrm{M}^{\mathrm{P}}_{\mathrm{snd}}} = 2\,\frac{\mathrm{min}}{\mathrm{doc}}$
- verifying CRF section – $\mu_{\mathrm{M}^{\mathrm{E}}_{\mathrm{ver}}} = 1\,\frac{\mathrm{min}}{\mathrm{doc}}$.

2.2.1.3. Data management

2.2.1.3.1. Data entry. We estimate average efforts needed for each step in data entry process as follows:

- evidence of received form – $\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{erf}}} = 1\,\frac{\mathrm{min}}{\mathrm{doc}}$
- entering data in database – $\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{dde}}} = 3\,\frac{\mathrm{s}}{\mathrm{data}}$
- resolving mismatches – $\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ddm}}} = 1\,\frac{\mathrm{min}}{\mathrm{data}}$
- evidence of query – $\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{pde}}} = 1\,\frac{\mathrm{min}}{\mathrm{data}}$.

2.2.1.3.2. Computer data analysis. We estimate that average effort for generating query document is twice higher for PDC than EDC, because, in the case of PDC, the coordinator has to write down both the question and to which field it is related. In the case of EDC the application enables the coordinator to enter the query right beside the field to which it relates.

Therefore, we estimate the average efforts as follows:

- evidence of query – $\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ade}}} = 1\,\frac{\mathrm{min}}{\mathrm{data}}$
- building query document (PDC) – $\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qpr}}} = 2\,\frac{\mathrm{min}}{\mathrm{data}}$
- building query document (EDC) – $\mu_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qpr}}} = 1\,\frac{\mathrm{min}}{\mathrm{data}}$
- sending query document by sponsor to monitor – $\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qsn}}} = 2\,\frac{\mathrm{min}}{\mathrm{doc}}$.

2.2.1.3.3. Query resolving. In the case of EDC no effort is needed to make evidence of queries and responses, as the EDC application takes care by itself of handling this information. Thus we consider that on average less effort is needed for review of each query in EDC process than in PDC (e.g. $\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qre}}} = 2\mu_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qre}}}$).

According to the interviewers the average effort for resolving one query is estimated as five times higher than effort needed to fill in data into CRF ($\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qin}}} = 5\mu_{\mathrm{CRF}^{\mathrm{P}}}$). When EDC application is used this can be estimated to be twice less ($\mu_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qin}}} = 2.5\mu_{\mathrm{CRF}^{\mathrm{P}}}$) as the investigator has all the information, apart from the data source documents, already available in the EDC application.

In our calculations we estimate the average efforts as follows:

- review and evidence of query (PDC) – $\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qre}}} = 1\,\frac{\mathrm{min}}{\mathrm{data}}$
- review of query (EDC) – $\mu_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qre}}} = 15\,\frac{\mathrm{s}}{\mathrm{data}}$
- forwarding query list by monitor to RC – $\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qfw}}} = 2\,\frac{\mathrm{min}}{\mathrm{doc}}$
- solving a query (PDC) – $\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qin}}} = 1\,\frac{\mathrm{min}}{\mathrm{data}}$
- solving a query (EDC) – $\mu_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qin}}} = 45\,\frac{\mathrm{s}}{\mathrm{data}}$.

2.2.1.3.4. Query closing. The related average efforts can be estimated as:

- evidence of query response – $\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{rev}}} = 1\,\frac{\mathrm{min}}{\mathrm{data}}$
- correcting data in the database – $\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{dcr}}} = 2\,\frac{\mathrm{min}}{\mathrm{data}}$
- evidence of processed form – $\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ecf}}} = 1\,\frac{\mathrm{min}}{\mathrm{doc}}$
- review of query response (EDC) – $\mu_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{loc}}} = 15\,\frac{\mathrm{s}}{\mathrm{data}}$.

2.2.2. Staff prices

The costs of the process depend on the staff prices. In our analysis we estimated staff prices as follows:

- Investigator: $\varphi_{\mathrm{I}} = 20\,\frac{€}{\mathrm{h}}$
- Investigator assistant: $\varphi_{\mathrm{IA}} = 10\,\frac{€}{\mathrm{h}}$
- Monitor (Clinical Research Associate – CRA): $\varphi_{\mathrm{CRA}} = 15\,\frac{€}{\mathrm{h}}$
- Data entry staff: $\varphi_{\mathrm{DE}} = 6\,\frac{€}{\mathrm{h}}$
- Data manager (senior): $\varphi_{\mathrm{DM}} = 20\,\frac{€}{\mathrm{h}}$
- Study coordinator: $\varphi_{\mathrm{C}} = 25\,\frac{€}{\mathrm{h}}$.

The prices can vary from center to center, from country to country. As needed, the above estimations can be replaced with specific prices for particular clinical trial.

2.2.3. Data quality

For most clinical trials data quality is set as a requirement. Certain level of data quality (low proportion of erroneous data) has to be assured in order to have reliable study results. There are several factors which affect the number of errors and these include:

- Error rate discovered by monitor — according to our interviewers for PDC $\varepsilon^{P}_{mon}=10\%$, and it is reported in [24] that for EDC ε^{E}_{mon} can be estimated as 20% ε^{P}_{mon}, which gives $\varepsilon^{E}_{mon}=2\%$.
- Error rate introduced by data entry staff (entry error rate — ε^{P}_{ent}) — according to literature [19–23] can be around 1%.
- Overall query rate (ε_{qry}) — according to our interviewers for PDC $\varepsilon^{P}_{qry}=2\%$; different sources (mentioned in [25]) pride EDC on 80% to 95% reduction on queries which let us estimate $\varepsilon^{E}_{qry}=0.2\%$.
- Plausibility discrepancies (data out of range, missing values, invalid combinations, etc.) — according to our interviewers and Spink [24] for PDC it represents 80% of all queries ($\varepsilon^{P}_{pla}=80\%\ \varepsilon^{P}_{qry}$).
- Analysis discrepancies (such as protocol violations or erroneous data, discovered by data analysis application and DM staff) — $\varepsilon^{P}_{ana}=20\%\ \varepsilon^{P}_{qry}$.
- Rate of data which have to be corrected after query resolving — according to our interviewers this can be estimated as half of all the queries $\varepsilon^{P}_{cor}=50\%\ \varepsilon^{P}_{qry}$.

The rates listed above are further discussed in the Results chapter and used to calculate the process costs.

2.3. Sample clinical trial

For the purpose of costs evaluation we decided to conceive a sample clinical trial. The characteristics of our sample clinical trial are:

- 10 research centers (RC)
- 100 patients per center (all together 1000 patients)
- CRF contains 10 sections
- Entire CRF contains 1000 data
- 24 months study duration
- Monitor visits each RC once a month.

Therefore we used the following values in our calculations:

- Total number of CRF forms (equal to the number of patients) — $N_{CRF}=1000$
- Total number of CRF sections — $N_{sec}=10{,}000$
- Total number of collected data — $N_{D}=1{,}000{,}000$
- Total number of monitor visits — $N_{M}=240$.

3. Results

3.1. Calculations

3.1.1. Data gathering at the research center

The difference in data capturing function (PDC: *CRF data entry*; EDC: *eCRF data entry*) and archiving function (PDC: *CRF archiving*) brings the differences in efforts and costs. If we generalize (for the purpose of estimation) that these processes are equal among all the participating research centers (RCs) for the sample clinical trial presented in Methods, a difference in costs of data gathering can be presented as:

$$\Delta\Phi_{DG}=\Phi_{DG^{P}}-\Phi_{DG^{E}}$$

$$\Phi_{DG^{P}}=F_{DG^{P}_{CRF}}\times\varphi_{IA}+F_{DG^{P}_{arh}}\times\varphi_{I}$$

$$\Phi_{DG^{E}}=F_{DG^{E}_{CRF}}\times\varphi_{IA}.$$

Total efforts for writing down data on paper CRF is presented as $F_{DG^{P}_{CRF}}=N_{D}\times\mu_{DG^{P}_{CRF}}$, where $\mu_{DG^{P}_{CRF}}$ is an average effort needed for writing down one data item. Total efforts needed for archiving all the paper CRF documents is presented as $F_{DG^{P}_{arh}}=N_{CRF}\times\mu_{DG^{P}_{arh}}$, where $\mu_{DG^{P}_{arh}}$ is an average effort needed for archiving one paper CRF.

On the other hand, total efforts for entering data into eCRF is presented as $F_{DG^{E}_{CRF}}=N_{D}\times\mu_{DG^{E}_{CRF}}$, where $\mu_{DG^{E}_{CRF}}$ is an average effort needed for entering one data.

We can calculate cost difference in PDC and EDC Data Gathering for our sample clinical trial as:

$$\Phi_{DG^{P}}\approx 33{,}500\,€$$

$$\Phi_{DG^{E}}\approx 16{,}500\,€$$

$$\Delta\Phi_{DG}\approx 17{,}000\,€.$$

3.1.2. Monitoring

The costs for PDC and EDC Monitoring processes can be estimated as:

$$\Phi_{M^{P}}=F_{M^{P}_{rev}}\times\varphi_{CRA}+F_{M^{P}_{cor}}\times(\varphi_{I}+\varphi_{IA}+\varphi_{CRA})+F_{M^{P}_{sig}}\times\varphi_{I}+F_{M^{P}_{loc}}\times\varphi_{CRA}+F_{M^{P}_{snd}}\times\varphi_{CRA}$$

$$\Phi_{M^{E}}=F_{M^{E}_{rev}}\times\varphi_{CRA}+F_{M^{E}_{loc}}\times\varphi_{CRA}+F_{M^{E}_{cor}}\times(\varphi_{I}+\varphi_{IA})+F_{M^{E}_{ver}}\times\varphi_{CRA}.$$

In the equations above the $F_{M^{P}}$ functions correspond to the PDC functions presented on Fig. 5 and $F_{M^{E}}$ functions correspond to the EDC functions presented on Fig. 6.

Furthermore, the number of notes that have to be taken depends on the number of errors discovered by monitor and this is considerably higher for PDC then EDC (as previously discussed in subsection: Methods: Model parameters: Data Quality).

$$F_{M^{P}_{rev}}=N_{D}\times\mu_{M_{chk}}+N_{D}\times\overset{P}{\varepsilon}_{mon}\times\mu_{M^{P}_{rer}}$$

$$F_{M^{E}_{rev}}=N_{D}\times\mu_{M_{chk}}+N_{D}\times\overset{E}{\varepsilon}_{mon}\times\mu_{M^{E}_{rer}}$$

We however consider that monitor visit lock effort is the same for PDC and EDC and can be presented as:

$$F_{M^{P}_{loc}}=F_{M^{E}_{loc}}=N_{M}\times\mu_{M_{loc}}.$$

The way data correction is performed and documented is also different for PDC and EDC. EDC data correction can be done through the EDC system which supports raising queries and sending messages with queries to investigators or their assistants. Furthermore, due to the validation routines on data entry when using EDC system, a number of errors among data is considerably lower for EDC process than PDC process and is reported in [24] to be $\varepsilon^{\mathrm{E}}_{\mathrm{mon}} = 20\%$ $\varepsilon^{\mathrm{E}}_{\mathrm{mon}} = 2\%$. This means that there will be less effort needed for EDC data correction than PDC.

$$F_{\mathrm{M}^{\mathrm{P}}_{\mathrm{cor}}} = N_{\mathrm{D}} \times \overset{\mathrm{P}}{\varepsilon}_{\mathrm{mon}} \times \mu_{\mathrm{M}_{\mathrm{cor}}},\ F_{\mathrm{M}^{\mathrm{E}}_{\mathrm{cor}}} = N_{\mathrm{D}} \times \overset{\mathrm{E}}{\varepsilon}_{\mathrm{mon}} \times \mu_{\mathrm{M}_{\mathrm{cor}}}$$

$$F_{\mathrm{M}^{\mathrm{P}}_{\mathrm{sig}}} = N_{\mathrm{sec}} \times \mu_{\mathrm{M}^{\mathrm{P}}_{\mathrm{sig}}}$$

$$F_{\mathrm{M}^{\mathrm{P}}_{\mathrm{snd}}} = N_{\mathrm{sec}} \times \mu_{\mathrm{M}^{\mathrm{P}}_{\mathrm{snd}}}$$

$$F_{\mathrm{M}^{\mathrm{E}}_{\mathrm{ver}}} = N_{\mathrm{sec}} \times \mu_{\mathrm{M}^{\mathrm{E}}_{\mathrm{ver}}}$$

By putting our estimations into the equations above we get the difference in costs between PDC and EDC Monitoring processes to be:

$$\Phi_{\mathrm{M}^{\mathrm{P}}} \approx 233,500\,€$$

$$\Phi_{\mathrm{M}^{\mathrm{E}}} \approx 140,000\,€$$

$$\Delta\Phi_{\mathrm{M}} \approx 93,500\,€.$$

3.1.3. Data management

Considering that the Data Management process for PDC is like on Fig. 7 and for EDC is like on Fig. 8, efforts for these two processes might be presented with the following equations:

$$\text{PDC Data Management}: F_{\mathrm{DM}^{\mathrm{P}}} = F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ENT}}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ANL}}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{QRY}}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{CLO}}}$$

$$\text{EDC Data Management}: F_{\mathrm{DM}^{\mathrm{E}}} = F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{ANL}}} + F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{QRY}}} + F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{CLO}}}.$$

3.1.3.1. Data entry. EDC data entry into the database is done by investigators (and their assistants). On the other hand, several data entry functions performed by data management staff are required in the case of PDC (Fig. 7). Thus, for PDC, Data Management costs considerably increase.

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ENT}}} = F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{erf}}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{dde}}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ddm}}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ddd}}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{eda}}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{edc}}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{pde}}}$$

The PDC data entry process begins with the evidence of received forms. RCs/monitors are sending entire CRF sections after being completed. Therefore all together $N_{\mathrm{sec}} = 10{,}000$ documents have to be evidenced.

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{erf}}} = N_{\mathrm{sec}} \times \mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{erf}}}$$

When double data entry approach is used to minimize data entry errors, data entry staff has to enter data twice.

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{dde}}} = N_{\mathrm{D}} \times \mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{dde}}} \times 2$$

Considering that average data entry error rate ($\varepsilon^{\mathrm{P}}_{\mathrm{ent}}$) is around 1% [19–23] for each staff member, this gives us nearly twice more mismatches in entered data produced by data entry staff.

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ddm}}} = N_{\mathrm{D}} \times \overset{\mathrm{P}}{\varepsilon}_{\mathrm{ent}} \times 2 \times \mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ddm}}}$$

Furthermore, data entry staff faces problems in reading some data written in CRFs and these have to be evidenced ($F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ddd}}}$). On the other hand some other data discrepancies are discovered by computer analysis of entered data and also evidenced ($F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{pde}}}$). Additional step ($F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{edc}}}$) to resolve these discrepancies needs to be introduced in the process in order to lower the error level. All these discrepancies we call plausibility discrepancies and we consider that they represent 80% of all the queries raised by sponsor ($\varepsilon^{\mathrm{P}}_{\mathrm{pla}} = 80\%\ \varepsilon_{\mathrm{qry}^{\mathrm{P}}}$; see Methods: Data quality section).

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ddd}}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{pde}}} = N_{\mathrm{D}} \times \overset{\mathrm{P}}{\varepsilon}_{\mathrm{pla}} \times \mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{pde}}}$$

Among plausibility discrepancies, thanks to double data entry, only few of these are eventually caused by data entry staff and in our calculation the effort for correcting these errors might be omitted ($F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{edc}}} \approx 0$). Also the effort for computer data analysis is relatively low comparing to all the other efforts and related costs as well and can be omitted from our calculation ($F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{eda}}} \approx 0$).

Considering average efforts estimations the difference in the costs of data entry process between PDC and EDC can thus be calculated as:

$$\Phi_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ENT}}} \approx F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{erf}}} \times \varphi_{\mathrm{C}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{dde}}} \times \varphi_{\mathrm{DE}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ddm}}} \times \varphi_{\mathrm{DM}} + \left(F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ddd}}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{pde}}}\right) \times \varphi_{\mathrm{DM}}$$

$$\Delta\Phi_{\mathrm{DM}_{\mathrm{ENT}}} = \Phi_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ENT}}} \approx 26,000\,€.$$

3.1.3.2. Computer data analysis. The PDC data analysis efforts can be presented as follows:

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ANL}}} = F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{cda}}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ade}}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qpr}}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qsn}}}.$$

In the case of EDC data analysis process is simpler as it does not require special efforts for query evidence and sending (this is done automatically through the EDC application):

$$F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{ANL}}} = F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{cda}}} + F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qpr}}}.$$

For example, after computer data analysis, the coordinator simply raises queries in the EDC application and unlocks the doubtful forms.

We can consider that the same efforts are put in computer data analysis for both PDC and EDC and for the simplification purposes we might estimate that these efforts are considerably low and might be omitted from our calculations.

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{cda}}} = F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{cda}}} \approx 0$$

On the other hand, different query rates for PDC ($\varepsilon^{\mathrm{P}}_{\mathrm{qry}}=2\%$) and EDC ($\varepsilon^{\mathrm{P}}_{\mathrm{qry}}=0.2\%$) lead to different efforts and costs for generating query documents.

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ade}}}=N_{\mathrm{D}}\times\left(\overset{\mathrm{P}}{\varepsilon}_{\mathrm{qry}}-\overset{\mathrm{P}}{\varepsilon}_{\mathrm{pla}}\right)\times\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ade}}}$$

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qpr}}}=N_{\mathrm{D}}\times\overset{\mathrm{P}}{\varepsilon}_{\mathrm{qry}}\times\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qpr}}}$$

$$F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qpr}}}=N_{\mathrm{D}}\times\overset{\mathrm{E}}{\varepsilon}_{\mathrm{qry}}\times\mu_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qpr}}}$$

Finally, as our sample clinical trial has all together $10\times24=240$ monitoring visits and each of them triggers data management process which results in query document, we considered that in principle 240 query documents ($N_{\mathrm{DQ}}=N_{\mathrm{M}}$) would be generated within the entire sample clinical trial.

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qsn}}}=N_{\mathrm{DQ}}\times\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qsn}}}$$

Considering average efforts estimations the difference in costs for PDC and EDC data analysis can be calculated as:

$$\Phi_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ANL}}}=F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ade}}}\times\varphi_{\mathrm{DM}}+F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qpr}}}\times\varphi_{\mathrm{C}}+F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qsn}}}\times\varphi_{\mathrm{C}}\approx18,000\,€$$

$$\Phi_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{ANL}}}=F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qpr}}}\times\varphi_{\mathrm{C}}\approx1000\,€$$

$$\Delta\Phi_{\mathrm{DM}_{\mathrm{ANL}}}\approx17,000\,€.$$

3.1.3.3. Query resolving. Different organizations have different approaches to query resolving. This can be done in direct communication between sponsor (coordinator) and investigators, or may include monitor in this process. We decided to take the approach where coordinator communicates only with monitor. Monitor then forwards queries (those which she/he cannot resolve by her/himself) to corresponding investigators. Corresponding investigators resolve these queries, and send the responses back to monitor who collects all the queries and respective answers and sends them to the coordinator. In this way monitor is aware of all eventual data changes.

From the models on Figs. 7 and 8 we obtain the following equations:

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{QRY}}}=F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qre}}}+F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qfw}}}+F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qin}}}+F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{rfw}}}+F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qpe}}}+F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{rsn}}}$$

$$F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{QRY}}}=F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qre}}}+F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qin}}}+F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qpe}}}.$$

Further estimation of costs is based on the following assumptions. The first is that the average efforts needed for review and evidence of query are equal to those needed for review and evidence of responses to queries. This allows us to simplify our calculations considering that $F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qpe}}}=F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qre}}}$ and $F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qpe}}}=F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qre}}}$, where:

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qre}}}=N_{\mathrm{D}}\times\overset{\mathrm{P}}{\varepsilon}_{\mathrm{qry}}\times\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qre}}}$$

$$F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qre}}}=N_{\mathrm{D}}\times\overset{\mathrm{E}}{\varepsilon}_{\mathrm{qry}}\times\mu_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qre}}}.$$

It is hard to predict to which research centers the queries raised by sponsor will belong. For the purpose of calculating $F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qfw}}}$ efforts we can consider that there are some queries raised after each monitor visit resulting in one query document for each center. Therefore the number of query documents that monitor has to forward is equal to $N_{\mathrm{DQ}}=240$.

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qfw}}}=N_{\mathrm{DQ}}\times\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qfw}}}$$

Third assumption is that the same effort is needed on the RC side to send the responses back to monitor ($F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{rfw}}}=FD_{\mathrm{M}^{\mathrm{P}}_{\mathrm{qfw}}}$).

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qin}}}=N_{\mathrm{D}}\times\overset{\mathrm{P}}{\varepsilon}_{\mathrm{qry}}\times\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qin}}}$$

$$F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qin}}}=N_{\mathrm{D}}\times\overset{\mathrm{E}}{\varepsilon}_{\mathrm{qry}}\times\mu_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qin}}}$$

Fourth and finally, we consider that effort needed by monitor to send responses back to coordinator is equal to efforts needed by coordinator to send the queries to monitor ($F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{rsn}}}=F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qsn}}}$).

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{rsn}}}=N_{\mathrm{DQ}}\times\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qsn}}}$$

Therefore, the costs for PDC and EDC query resolving process are calculated as:

$$\begin{aligned}\Phi_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{QRY}}}&=F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qre}}}\times\varphi_{\mathrm{CRA}}\times2+F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qfw}}}\times(\varphi_{\mathrm{CRA}}+\varphi_{\mathrm{I}})\times2\\&\quad+F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{qin}}}\times(\varphi_{\mathrm{I}}+\varphi_{\mathrm{IA}})+F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{rsn}}}\times\varphi_{\mathrm{CRA}}\\\Phi_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{QRY}}}&=F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qre}}}\times\varphi_{\mathrm{CRA}}\times2+F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{qin}}}\times(\varphi_{\mathrm{I}}+\varphi_{\mathrm{IA}}).\end{aligned}$$

Considering average efforts estimations the difference in costs of PDC and EDC query resolving is:

$$\Phi_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{QRY}}}\approx21,000\,€;\ \Phi_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{QRY}}}\approx1000\,€;\ \Delta\Phi_{\mathrm{DM}_{\mathrm{QRY}}}\approx20,000\,€.$$

3.1.3.4. Query closing. The last step in data management leads to locking the data in the database. Data are considered to be clean and reliable, and therefore should be locked to further changes.

When PDC queries are resolved and responses sent back to the data management team, responses have to be evidenced and, if necessary, data in the database have to be corrected. Finally, all the processed forms (CRF sections) have to be evidenced as well. For these the efforts are:

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{CLO}}}=F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{rev}}}+F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{dcr}}}+F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{epf}}}$$

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{rev}}}=N_{\mathrm{D}}\times\overset{\mathrm{P}}{\varepsilon}_{\mathrm{qry}}\times\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{rev}}}$$

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{dcr}}}=N_{\mathrm{D}}\times\overset{\mathrm{P}}{\varepsilon}_{\mathrm{cor}}\times\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{dcr}}}$$

$$F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{epf}}}=N_{\mathrm{sec}}\times\mu_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{ecf}}}.$$

On the other hand, in the case of EDC, if necessary, investigators have already entered any data changes and clarifications to queries directly in the EDC application.

Therefore the coordinator has just to look through the responses and lock the database.

$$F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{CLO}}} = F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{loc}}} = N_{\mathrm{D}} \times \varepsilon^{\mathrm{E}}_{\mathrm{qry}} \times \mu_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{loc}}}$$

Considering average efforts estimations we can calculate the difference in costs for PDC and EDC query closing as:

$$\Phi_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{CLO}}} = F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{rev}}} \times \varphi_{\mathrm{C}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{dcr}}} \times \varphi_{\mathrm{DM}} + F_{\mathrm{DM}^{\mathrm{P}}_{\mathrm{epf}}} \times \varphi_{\mathrm{C}} \approx 19,000\,€$$

$$\Phi_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{CLO}}} = F_{\mathrm{DM}^{\mathrm{E}}_{\mathrm{loc}}} \times \varphi_{\mathrm{C}} \approx 0\,€$$

$$\Delta\Phi_{\mathrm{DM}_{\mathrm{CLO}}} \approx 19,000\,€.$$

Finally, considering the calculations above the entire difference in costs between PDC and EDC Data Management processes for our sample clinical trial is:

$$\Delta\Phi_{\mathrm{DM}} = \Delta\Phi_{\mathrm{DM}_{\mathrm{ENT}}} + \Delta\Phi_{\mathrm{DM}_{\mathrm{ANL}}} + \Delta\Phi_{\mathrm{DM}_{\mathrm{QRY}}} + \Delta\Phi_{\mathrm{DM}_{\mathrm{CLO}}}$$

$$\Delta\Phi_{\mathrm{DM}} \approx 82,000\,€.$$

In conclusion, the total PDC costs (ϕ^{P}), EDC costs (ϕ^{E}) and costs differences ($\Delta\phi$) between PDC and EDC process, affected by implementation of EDC application and related changes of data gathering, monitoring and data management processes, are estimated as:

$$\Phi^{\mathrm{P}} = \Phi_{\mathrm{DG}^{\mathrm{P}}} + \Phi_{\mathrm{M}^{\mathrm{P}}} + \Phi_{\mathrm{DM}^{\mathrm{P}}} \approx 350,000\,€$$

$$\Phi^{\mathrm{E}} = \Phi_{\mathrm{DG}^{\mathrm{E}}} + \Phi_{\mathrm{M}^{\mathrm{E}}} + \Phi_{\mathrm{DM}^{\mathrm{E}}} \approx 158,500\,€$$

$$\Delta\Phi = \Phi^{\mathrm{P}} - \Phi^{\mathrm{E}} \approx 192,500\,€.$$

From the results presented above we can see that just due to the savings in overall staff efforts (without consideration of savings due to shortening of data collection process duration) the EDC process costs are 55% lower than PDC costs.

3.2. Sensitivity analysis

We modeled paper data collection (PDC) and electronic data collection (EDC) processes in order to evaluate the difference in costs of these two approaches in clinical trial. Furthermore, we estimated the parameters of our models that affect the costs and calculated PDC and EDC costs as well as costs differences for a sample clinical trial. The exact values depend both on the models (processes) and the values of the parameters used to calculate the costs. Here we mainly considered different estimations of error/query rates and average efforts for each of the process elements, as well as staff prices. However, in this section we will discuss our estimations of parameter and try to understand how variations of these may affect our results.

Our estimations of parameters are based on available literature and the information obtained in interviews. Despite the fact that these estimations are not highly reliable (more reliable would be experimental measurements of these parameters) we believe that our results reflect the PDC and EDC costs differences well.

Some of our estimations of error and query rates are based on information provided by interviewers. These include PDC error rate discovered by monitor which is around 10% and PDC query rate which is 2%. The last is also supported by some other studies [24]. The EDC error and query rates we estimated by taking in consideration PDC rates and published studies [19–25] on comparison between PDC and EDC error and query rates. These definitely prove that electronic data collection reduces number of errors and queries. The reduction is highly dependent on the quality of EDC application which must be designed carefully and user friendly. If the EDC application is under-designed and EDC error and query rates are thus higher, this will lower the difference in costs. However, if we double the EDC error and query rates this will raise EDC costs from 158,500 € to 173,000 €, which is 8.5% increase. As a result the overall costs savings of switching to EDC process decrease from 55% to 45%. This however is still a considerable saving.

The size of clinical trial also affects the savings considerably. Namely if we consider twice smaller sample clinical trial having 5 research centers with 50 patients each and 500 data per CRF the entire costs of PDC decrease from 350,000 € to only 44,000 €. Despite the fact that the savings ratio is still 55% the 24,000 € of EDC savings would probably not justify the investment into EDC application.

We believe that it is of key importance to measure the average efforts for all the process tasks and then recalculate the equations and get more reliable costs. For example, if we consider that the average efforts for EDC tasks are equal to the PDC tasks (e.g. that entering data in to eCRF is as fast as writing down data into paper CRF) the results slightly change. In that case the EDC costs rise for 13% from 158,500 € to 179,000 €, which results in EDC savings lowering from 55% to 50%. However, this might be still considered as a considerable reduction in costs. We intend to measure the parameters and use measured parameters to make more reliable calculations. We also encourage others to try to use their estimations of average efforts and error rates to calculate the costs differences according to our process models (PDC and EDC).

We set PDC and EDC models that might be challenged. For example, it is not so unusual that the sponsor decides to decrease monitoring costs by verifying just a sample of

Table 1

Data gathering and monitoring costs for different estimations of parameters ($\phi_{\mathrm{DG}^{\mathrm{P}}}$ — PDC Data Gathering costs; $\phi_{\mathrm{DG}^{\mathrm{E}}}$ — EDC Data Gathering costs; $\Delta\phi_{\mathrm{DG}}$ — difference between PDC and EDC Data Gathering costs; $\phi_{\mathrm{M}^{\mathrm{P}}}$ — PDC Monitoring costs; $\phi_{\mathrm{M}^{\mathrm{E}}}$ — EDC Monitoring costs; $\Delta\phi_{\mathrm{M}}$ — difference between PDC and EDC Monitoring costs).

Scenario	$\phi_{\mathrm{DG}^{\mathrm{P}}}$	$\phi_{\mathrm{DG}^{\mathrm{E}}}$	$\Delta\phi_{\mathrm{DG}}$	$\phi_{\mathrm{M}^{\mathrm{P}}}$	$\phi_{\mathrm{M}^{\mathrm{E}}}$	$\Delta\phi_{\mathrm{M}}$
Standard values as presented in Methods and Results	33.660 €	16.500 €	17.160 €	233.559 €	140.093 €	93.466 €
Doubled EDC error/query rates	33.660 €	16.500 €	17.160 €	233.559 €	152.612 €	80.947 €
Smaller clinical trial (12 months; 5 RC × 50 CRF × 500 data)	4.290 €	2.063 €	2.228 €	29.210 €	17.526 €	11.683 €
Equal EDC and PDC average efforts	33.660 €	33.000 €	660 €	233.559 €	142.592 €	90.967 €
Only 50% data are checked by monitor	33.660 €	16.500 €	17.160 €	120.984 €	71.358 €	49.626 €

Table 2
Data management costs for different estimations of parameters ($\phi_{DM^{P}_{ENT}}$ — PDC data entry costs; $\phi_{DM^{P}_{ANL}}$ — PDC data analysis costs; $\phi_{DM^{E}_{ANL}}$ — EDC data analysis costs; $\Delta\phi_{DM_{ANL}}$ — difference between PDC and EDC data analysis costs; $\phi_{DM^{P}_{QRY}}$ — PDC query resolving costs; $\phi_{DM^{E}_{QRY}}$ — EDC query resolving costs; $\Delta\phi_{DM_{QRY}}$ — difference between PDC and EDC query resolving costs; $\phi_{DM^{P}_{CLO}}$ — PDC query closing costs; $\phi_{DM^{E}_{CLO}}$ — EDC query closing costs; $\Delta\phi_{DM_{CLO}}$ — difference between PDC and EDC query closing costs; ϕ^{P}_{DM} — PDC Data Management costs; ϕ^{E}_{DM} — EDC Data Management costs; $\Delta\phi_{DM}$ — difference between PDC and EDC Data Management costs).

Scenario	$\phi_{DM^{P}_{ENT}}$	$\phi_{DM^{P}_{ANL}}$	$\phi_{DM^{E}_{ANL}}$	$\Delta\phi_{DM_{ANL}}$	$\phi_{DM^{P}_{QRY}}$	$\phi_{DM^{E}_{QRY}}$	$\Delta\phi_{DM_{QRY}}$	$\phi_{DM^{P}_{CLO}}$	$\phi_{DM^{E}_{CLO}}$	$\Delta\phi_{DM_{CLO}}$	$\phi_{DM^{P}}$	$\phi_{DM^{E}}$	$\Delta\phi_{DM}$
Standard values as presented in Methods and Results	26.099 €	18.034 €	835 €	17.199 €	20.713 €	1.000 €	19.713 €	19.125 €	209 €	18.917 €	83.971 €	2.044 €	81.928 €
Doubled EDC error/query rates	26.099 €	18.034 €	1.670 €	16.364 €	20.713 €	2.000 €	18.713 €	19.125 €	417 €	18.708 €	83.971 €	4.087 €	79.884 €
Smaller clinical trial (12 months; 5 RC × 50 CRF × 500 data)	3.262 €	2.279 €	104 €	2.175 €	2.673 €	125 €	2.548 €	2.391 €	26 €	2.365 €	10.605 €	255 €	10.350 €
Equal EDC and PDC average efforts	26.099 €	18.034 €	1.670 €	16.364 €	20.713 €	1.500 €	19.213 €	19.125 €	417 €	18.708 €	83.971 €	3.587 €	80.384 €
Only 50% data are checked by monitor	26.099 €	18.034 €	835 €	17.199 €	20.713 €	1.000 €	19.713 €	19.125 €	209 €	18.917 €	83.971 €	2.044 €	81.928 €

gathered data instead of all the gathered data. For example, this could be done by checking only critical data (e.g. applied therapy, responses, and adverse events). Therefore, if we recalculate the costs with the assumption that monitor checks only half of all the gathered data the costs of PDC will decrease for 32% (from 350,000 € to 238,500 €) as well as the costs of EDC for 43% (from 158,500 € to 90,000 €). The savings amount will decrease from 192,500 € to 148,500 €, but this will actually raise the saving rate from 55% to 62%.

The costs and savings from all the different scenarios and estimations mentioned above are presented in the Tables 1–3. Namely, we present the costs for standard values as presented in Methods and Results and the following variations:

- doubled EDC error/query rates;
- smaller clinical trial (12 months, 5 research centers with 50 patients in each center contributing 500 data per patient);
- equal EDC and PDC average efforts; and
- only 50% data checked by monitor.

4. Discussion

From the results presented in Tables 1–3 we can conclude that the EDC process brings savings when compared to PDC process for all five scenarios. The savings primarily are due to lower error and query rates which reduce work that have to be done on data cleaning. Another important parameter that influences CT costs is the size of CT (number of centers, patients and collected data). It is obvious that for smaller CT savings might be insufficient to justify the investment in EDC application. However, by using our models and own parameters' assessments the sponsor can calculate the savings of EDC comparing to PDC process and consider this information when making a decision on investment in an EDC application.

The models which we presented in this paper are limited to the data collection related processes. However, switch to EDC brings many changes in clinical trial organization and implementation. EDC remarkably reduces spending on paper. Furthermore, CT management may be driven by up-to-date information from EDC application. For example, monitoring visits might be appointed when sufficient data is submitted by research center, instead of having a priori scheduled visits. Serious Adverse Events (SAE) can automatically trigger SAE reporting mechanisms and enhance patients' safety. Finally, greatest savings EDC brings through shortening of last patient out to database lock time which eventually decreases drug's time to market.

There are also disadvantages which EDC brings, comparing to PDC. For example there is a need for extended costs due to hotline and maintenance of the EDC system. Also there are hardware and network constraints (e.g. firewall, low or no network availability, or slow network at some research centers, etc.) which might appear and which may also require additional costs for solving these issues. Another disadvantage of EDC is inconvenience of electronic data entry during outpatient visit, as well as during some demanding clinical activities. Further problems may appear due to the regulations related to eSource [26]. The last, but not the least, a switch from paper-based process to electronic process may be a demanding organizational challenge as well. An extended discussion on advantages and disadvantages of EDC over PDC was recently presented by Welker [3].

In our study, we considered two "extreme" approaches — fully paper-based data collection and fully electronic data collection. In many cases, a number of scenarios may be necessary for the same trial. In the same clinical trial, some

Table 3
Total costs and savings for different estimations of parameters (ϕ^{P} — PDC costs; ϕ^{E} — EDC costs; $\Delta\phi$ — difference between PDC and EDC costs).

Scenario	ϕ^{P}	ϕ^{E}	$\Delta\phi$	EDC savings
Standard values as presented in Methods and Results	351.190 €	158.637 €	192.554 €	55%
Doubled EDC error/query rates	351.190 €	173.199 €	177.991 €	51%
Smaller clinical trial (12 months; 5 RC × 50 CRF × 500 data)	44.105 €	19.844 €	24.261 €	55%
Equal EDC and PDC average efforts	351.190 €	179.179 €	172.011 €	49%
Only 50% data are checked by monitor	238.615 €	89.902 €	148.713 €	62%

centers might perform PDC and other EDC. Data center should handle data coming from both sources, making data collection models more complex. It may appear that EDC savings cannot justify all the complexity that handling hybrid PDC/EDC clinical trial brings.

All the disadvantages mentioned above have to be balanced with the advantages of EDC and savings that EDC process may bring. Namely only in this way the switch from PDC to EDC can be justified. Nevertheless, our study shows that modeling of processes together with accurate estimations of process parameters clearly identifies where in the data collection process the costs savings come from. Having in mind that data management (site setup, monitoring and closing including CRF processing) costs may be estimated as almost 30% of Phase III clinical trial costs (according to published Fast Track Systems example $16,000,000 of $58,400,000 [28]), optimization of processes and introduction of IT solutions in order to reduce the costs of clinical trial may be a reasonable decision. Considering that clinical trial costs represent 47.9% (Phase III — 28.1%, Phase II — 13.1%, Phase I — 6.7%) of the entire drug R&D investments (according to PhRMA Membership Annual Report 2008 [27]) we can conclude that lowering clinical trial costs consequently reduce drug-to-market costs.

5. Conclusion

In our report we demonstrated an example of how a change from paper-based data collection (PDC) to internet based electronic data collection (EDC) affects the costs of data collection and its sub-processes. We developed eEPC (Extended Event-driven Process Chains) models for PDC and EDC sub-processes (data gathering, monitoring and data management) and simplified them to the extent that allowed us to calculate the related costs. We estimated the values of all the parameters which appear in the models, such as clinical trial size, error and query rates, average efforts and staff prices. We based these estimations on available literature and information obtained in interviews. The results show that most benefit comes from reducing monitoring and data management costs. The exact value depends on estimation of parameters which affect the calculations. For example, the variations in clinical trial size significantly affect the savings that EDC brings, while the costs are not that sensible to changes in average efforts for particular tasks. These results are not surprising, but with our approach we offer the way to quantify them. With more reliable estimation of average costs and error and query rates and considering specific values we can get more reliable results and use them to decide on switching from paper data collection to electronic data collection. We however have to emphasize that our models do not include all the aspects of organization and implementation of clinical trial which are affected by the change in data collection approach. The same approach that we presented in our study can however be used in assessing the costs of the clinical trial through modeling of the entire clinical trial as a business process.

Acknowledgments

This research was in part supported by the Slovenian Research Agency and Slovenian Ministry of the Economy.

References

[1] European Parliament and the Council of the European Union. Directive 2001/20/EC. Off J Eur Communities May 1 2001;121:34–44.

[2] Paul J, Seib R, Prescott T. The internet and clinical trials: background, online resources, examples and issues. J Med Internet Res Mar 16 2005;7(1):e5.

[3] Welker JA. Implementation of electronic data capture systems: barriers and solutions. Contemp Clin Trials 2007;28:329–36.

[4] Brandt CA, Argraves S, Money R, Ananth G, Trocky NM, Nadkarni PM. Informatics tools to improve clinical research study implementation. Contemp Clin Trials 2006;27:112–22.

[5] Alschuler L, Bain L, Kush RD. Improving data collection for patient care and clinical trials. Sci Career Mag Mar 26 2004 http://sciencecareers.sciencemag.org/career_magazine/previous_issues/articles/2004_03_26/noDOI.5622907321165187916. Archived at: http://www.webcitation.org/5cGF6Xmf8.

[6] Pavlovic I, Miklavcic D. Web-based electronic data collection system to support electrochemotherapy clinical trial. IEEE Trans Inf Technol Biomed Mar 2007;11(2):222–30.

[7] Edwards RL, Edwards SL, Bryner J, Cunningham K, Rogers A, Slattery ML. A computer-assisted data collection system for use in a multicenter study of American Indians and Alaska Natives: SCAPES. Comput Methods Programs Biomed Apr 2008;90(1):38–55.

[8] Proctor SJ, Wilkinson J. A web-based study concept designed to progress clinical research for 'orphan' disease areas in haematological oncology in the elderly: the SHIELD programme. Crit Rev Oncol/Hematol 2007;61:79–83.

[9] Formica M, Kabbara K, Clark R, McAlindon T. Can clinical trials requiring frequent participant contact be conducted over the internet? Results from an online randomized controlled trial evaluating a topical ointment for herpes labialis. J Med Internet Res 2004;6(1):e6.

[10] Avidan A, Weissman C, Sprung CL. An internet web site as a data collection platform for multicenter research. Anesth Analg 2005;100:506–11.

[11] Lopez-Carrero C, Arriaza E, Bolanos E, Ciudad A, Municio M, Ramos J, et al. Internet in clinical research based on a pilot experience. Contemp Clin Trials 2005;26:234–43.

[12] Lallas CD, Preminger GM, Pearle MS, Leveillee RJ, Lingeman JE, Schwope JP. Internet based multi-institutional clinical research: a convenient and secure option. J Urol May 2004;171(5):1880–5.

[13] Rangel SJ, Narasimhan B, Geraghty N, Moss RL. Development of an internet-based protocol to facilitate randomized clinical trials in pediatric surgery. J Pediatr Surg 2002;37(7):990–4.

[14] Marks R, Bristol H, Conlon M, Pepine CJ. Enhancing clinical trials on the internet: lessons from INVEST. Clin Cardiol 2001;24(supplV):V17–23.

[15] Collada AL, Fazi P, Luzi D, Ricci FL, Serbanati LD, Vignetti M. Toward a model of clinical trials. Proceedings of the 5th international symposium ISBMDA; Nov 18–19 2004. p. 299–312. Barcelona, Spain, http://www.springerlink.com/content/y3yxru6pxhfpw015/.

[16] Luzi D, Ricci FL, Serbanati LD. E-clinical trials supported by a service-oriented architecture. Proceedings of Mednet 2006: 11th world congress on the internet in medicine — MEDNET; Oct 13–26 2006. Society for the Internet in Medicine, Toronto, Canada, http://www.mednetcongress.org/fullpapers/MEDNET-137_DanielaLuziA4_e.pdf. Archived at: http://www.webcitation.org/5cIrbshXk.

[17] Clinical Trial Electronic Data Capture Task Group. PhRMA Biostatistics and Data Management Technical Group. US PhRMA's EDC position paper, revision 1. PhRMA eClinical Forum; May 2005. http://www.eclinicalforum.com/content/Knowledge/Articles/EDC%20Revision%201%20-%20Final%20Version.pdf. Archived at: http://www.webcitation.org/5cGFm0gTX.

[18] Scheer AW. ARIS — business process modeling. Berlin, Germany: Springer-Verlag; 2000.

[19] King DW, Lashley R. A quantifiable alternative to double data entry. Control Clin Trials 2000;21:94–102.

[20] Wahi MM, Parks DV, Skate RC, Goldin SB. Reducing errors from the electronic transcription of data collected on paper forms: a research data case study. J Am Med Inform Assoc 2008;15:386–9.

[21] Kleinman K. Adaptive double data entry: a probabilistic tool for choosing which forms to reenter. Control Clin Trials 2001;22:2–12.

[22] Kawado M, Hinotsu S, Matsuyama Y, Yamaguchi T, Hashimoto S, Ohashi Y. A comparison of error detection rates between the reading aloud method and the double data entry method. Control Clin Trials 2003;24:560–9.

[23] Day S, Fayers P, Harvey D. Double data entry: what value, what price? Control Clin Trials 1998;19:15–24.

[24] Spink C. Electronic Data Capture (EDC) as a means for e-clinical trial success. IBM Global Services. Pharmaceutical Clinical Development; 2002 (Mar).

[25] Bart T. Comparison of electronic data capture with paper data collection — is there really an advantage? Bus Brief Pharmatech 2003:1–4.

[26] The eClinical Forum and PhRMA EDC/eSource Taskforce. The future vision of electronic health records as eSource for clinical research. The eClinical Forum and PhRMA EDC/eSource Taskforce; Sep 14 2006. http://www.eclinicalforum.com/content/Knowledge/Articles/Future%20EHR-CR%20Environment%20Version%201.pdf.Archived at: http://www.webcitation.org/5cTvNZOVC.

[27] PhRMA. Pharmaceutical industry profile 2008. Washington, DC: PhRMA; 2008 (Mar), http://www.phrma.org/files/2008%20Profile.pdf. Archived at: http://www.webcitation.org/5f0RbcGI4.

[28] BIO-IT World. Where the millions go.; May 9 2003. The YGS Group, 1808 Colonial Village Lane, Lancaster, PA. http://www.bio-itworld.com/archive/050903/data_sidebar_2450.html. Archived at: http://www.webcitation.org/5f0UtE4yS.