Semantically-enabled Intelligent Patient Recruitment in Clinical Trials

Vassiliki Andronikou, IEEE Member, Efthamios Karanastasis, Efthymios Chondrogiannis, Konstantinos Tserpes, Theodora Varvarigou, IEEE Member
Department of Electrical and Computer Engineering,
National Technical University of Athens
Athens, Greece
e-mail: vandro@mail.ntua.gr

Abstract—This paper presents and analyzes the data requirements within the clinical trial design process with particular focus on the selection of patients who are eligible to participate in the specified clinical study. The latter comprises an extremely time-consuming process which requires considerable budget and effort, whereas the resulting recruited subjects determine both the success of the clinical study and the validity of the clinical study results significantly. Hence, a novel approach based on Service Oriented Architecture mechanisms and the incorporation of a great number of ontologies enabling the semantic linking between clinical research and clinical care and the interpretation of numerous distributed heterogeneous data sources is presented and analyzed followed by the expected impact of its application in the clinical study design and implementation processes.

Keywords—clinical trial design, subject recruitment, semantic interoperability

I. INTRODUCTION
Translation of basic research findings into clinical therapy has to overcome substantial barriers at the preclinical and clinical levels. The market numbers are quite indicative; according to empirical studies, the number of medicines introduced worldwide containing new active ingredients has dropped from an average of over 60 per year in the late 1980s to 52 in 1991, only 31 in 2001 [12] and around 20–25 new licensed drugs per year over the past years [4]. It is interesting to notice that the success rate from first dose in man to registration is only about 11% across indications [6]. The drop of the overall number of new active substances undergoing regulatory review as well as the low success rates stem from a variety of financial, scientific and business reasons. On average, pharmaceutical companies spend more than $800 million per drug candidate, with clinical testing involving about one third of these costs. What is really interesting to notice is that half of these costs are highly related to the temporal needs of the clinical trials. The average duration of phase II and phase III trials is 25.7 and 30.5 months respectively [3], whereas the new drug development timeline is on average 11.3 years (including 4.3 years for discovery, pre-clinical research and development and 7 years for clinical trials’ conduction and final approval).

Given the ever-increasing research and development (R&D) costs in the pharmaceutical industry and the low success rate in these efforts, over the past years pharmaceutical companies show a tendency towards launching modified versions of clinically validated drug targets against novel targets with new chemical matter. This approach, which is known as drug re-positioning, drug re-purposing or therapeutic switching, quite often introduces incremental innovation, but it also presents reduced risk of failure and can generate generous profits [8]. Moreover, it extends to the ongoing attempts by pharmaceutical companies to extend the period of time under patent protection for a given drug and its associated family of products.

In the meanwhile, quite often great deviation between clinical research findings and the outcome of the treatments’ application in clinical care is recorded [11], limiting this way the external validity of the clinical trials’ findings; in other words, the understanding - on behalf of the doctors - of how widely these findings could be applied in clinical care while ensuring the patients’ safety. This deviation is driven by numerous factors the main of which include poor clinical trial design, lack of access to and linking with important data, such as Electronic Health Records (EHR) and related non-clinical and clinical research findings, and a non-well balanced resulting patient sample recruited in the clinical trials.

In this paper we will analyse the clinical trial design process and we present a novel ICT-based approach for enabling the automatic selection of patients eligible to participate into clinical trials by taking into consideration a number of factors related to the patients’ safety, the expected efficacy of the clinical trial and the related cost and how a wide range of data technologies promises to boost and advance such an effort.

II. BACKGROUND
Ensuring that a representative patient sample has been recruited in the clinical trials comprises a tremendously time consuming and an extremely difficult process which still needs to overcome the barriers of limited access to required patient data. In fact, subject recruitment accounts for
approximately 30% to 40% of clinical trial costs [7], only 15% of clinical trials conclude on schedule, while the other ones face tremendous delays due to low participant accrual. And what is more, 60% - 80% of trials do not meet their temporal endpoint because of recruitment issues, whereas 30% of trial sites fail to recruit even a single participant [9].

The adoption of the EHR aims at bringing together this information from the different entities within healthcare by providing an electronic umbrella for this information. However, proper linking of this information with clinical research information systems is required so that the flow of medical information between these two areas will be enabled.

Towards this direction, many efforts are taking place at various levels in order to allow the information flow between two “isolated” but highly related areas; clinical research and clinical care. Thus, the ongoing effort for developing standards in clinical care and clinical research by international organizations such as HL7 Working Group [5] and CDISC [2] respectively is now complemented by an effort to harmonise the standards across the two domains. This work is incorporated into the Biomedical Research Integrated Domain Group (BRIDG) model [1], a clinical research domain analysis model aiming to harmonize and connect the entire spectrum of CDISC standards.

In the meanwhile, PCROM (Primary Care Research Object Model) is working towards the same direction. It has borrowed when required and possible from the work mentioned above, but it is driven by a different, though related, set of needs aiming at the development and implementation of a system able to support Randomised Clinical Trials in the primary care setting [10].

III. CLINICAL TRIAL MANAGEMENT ANALYSIS

As already mentioned above, data collection, analysis, mining and annotation pose a significant challenge for clinical investigators, due to the non-interoperable heterogeneous distributed data sources involved (published non-clinical and clinical research findings, Electronic Health Records, Drug Databases, etc). A great amount of medical information crucial to the success of a clinical trial is hidden inside a variety of information systems and data sources that do not share the same semantics nor adhere to widely deployed clinical data standards.

A. Patient Selection Process

One of the key goals of a clinical trial is to provide reliable and accurate clinical evaluation of an investigational drug for a target patient population suffering from a specific disorder (or disorders). Thus, a great challenge met at the clinical trial design phase is defining the sample patient population in such a way that it is homogeneous and representative so that it comprises the best possible approximation of the target patient population.

Within the clinical study protocol and through the determination of the inclusion and exclusion criteria, the clinical investigators describe a sample of the target patient population which is representative enough to control systematic error and, quite often, allow for a wide market share covering as much as possible the target population’s particularities, can be studied at tolerable costs (both monetary and temporal), and will allow for investigating upon their clinical study objectives and endpoints more efficiently and effectively.

Inclusion criteria describe the target patient population that the clinical study will aim at forming particular guidelines for subject participation. Exclusion criteria include features, personal characteristics, medical history, co-morbidities and particularities that a person of the target patient population may have which make them inappropriate for the clinical study. The reasons for setting the latter criteria include inability of the person to conform to the study protocol and safety issues due to comorbidities that are treated by medication severely interacting with the investigational drug, among others.

As these criteria determine the study population from which the subjects will be enrolled in the clinical trial, defining them in a precise, clear and well-documented manner comprises a very important issue. In fact, these criteria serve two distinct purposes: they form the basis for identifying the most “suitable” subjects for participating in the clinical trial through which the clinical investigator tries to find answers to their research questions as well as they comprise the “safeguards” of the subject’s medical safety.

Moreover, given that most often more than one different clinical sites and clinical investigators across the world organise clinical trials aiming at investigating and answering the same scientific question, the objective, scientifically-validated and unambiguous determination of the inclusion and exclusion criteria for the clinical studies comprises an essential basis for enabling the collaboration of researchers as well as the replication of studies for reaching results of high external validity.

The clinical investigators go through a patient “evaluation” process, during which they try to select patients eligible to participate in the clinical trial that they have designed. In order for the researchers to find potential patients to recruit within their study they advertise the opening of the recruitment process or they search within their clinic and/or hospital records, quite often following manual processes. These procedures, however, are extremely time-consuming, very restricting and, most of the times, insufficient. The advent of EHR adoption in hospitals brings a significant new opportunity for researchers to contact their recruitment needs to a much wider population.

B. Core Patient Profile

The term eligibility within our approach encapsulates three distinct concepts; clinical trial efficacy, patient safety and clinical trial cost. Hence, the basis for the enhancement of the patient recruitment process for drug-repositioning
clinical trials is the definition of a Core Patient Profile incorporating a set of sub-profiles;

- an efficacy sub-profile, which indicates how well the patient matches with the criteria the investigator has set for the patients who should participate in this clinical trial,
- a risk sub-profile, which denotes how safe the participation of the patient in the clinical trial will be,
- and a cost sub-profile, which is linked with the cost related to the participation of the patient in the clinical trial.

The latter are built based on advanced models and sets of complex rules which are formed and fed through the intelligent correlation of the various knowledge bases, including:

- Electronic Health Records (EHRs),
- disease data including epidemiology and pathophysiology of the disease,
- and clinical and non-clinical data related to the clinical study.

Especially in the case of drug repositioning clinical trials, drug profile data including pharmacokinetics (what a drug does to the body) and pharmacodynamics (what the body does to the drug) information comprises an additional valuable source.

The first sub-profile is strongly related to the objectives of the clinical study and the investigation of the effectiveness of the investigational drug on the study disorder (a.k.a. test of hypothesis). A great number of clinical studies examines the efficacy of a treatment by “shaping ideal conditions”; in other words, they select patients with no health problems other than the disorder under investigation. However, such conditions deviate strongly from everyday life in which most often patients suffer from more than one disorders.

The second sub-profile is connected with the safety of the patients and is built from rules formed by drug-drug interactions, special groups’ particularities (such as pregnant women, children, mentally-disorder people), as well as findings from relevant clinical studies outcomes.

The third sub-profile pertains to the actual cost of the clinical trial and mainly focuses on the required expenses related to the insurance coverage of the participants, their demographic data indicating need for traveling as well as the costs of the medical examinations for measuring the clinical trial endpoints.

C. A Novel Framework for Automated Patient Selection

In the paragraphs above, we have presented and analyzed the data requirements related to the design of a clinical trial with particular focus on the determination of the sample target population and the selection of the subjects which are eligible to participate in the clinical study. Moreover, various data sources which are available and highly related to clinical trial design, yet dispersed and heterogeneous, have been listed. This section presents an innovative framework incorporating an advanced infrastructure and highly intelligent decision support mechanisms for facilitating this process.

In Figure 1, the conceptual architecture of this solution is presented, including the main subsystems that are required as well as the various data sources. The key components of the solution include:

- Clinical Trial Management Tool: this component acts as the interface of the framework and allows the clinical investigators to perform queries within the “Cloud of Data” in order to define the inclusion and exclusion criteria for their clinical study as well as the rest of the clinical study parameters included in the clinical study protocol (such as clinical trial endpoints, dosage schedule, etc).
- Authentication/Authorization module: this part of the framework includes the authentication and authorization mechanisms that allows for the system user identification and control of access to services and data sources covering part of the confidentiality and privacy needs.
- Semantic Data Representation Layer: this layer is responsible for translating the queries and the data entered by the clinical investigator into the “system language”.
- Decision Support entity: this component is responsible for building the patients’ Core Profile
based on the inclusion/exclusion criteria specified by the investigator and the clinical trial endpoints. The functionality of this component is based also on the correlation of disparate data retrieved from the Ontology-Based Search Engine through intelligent queries. This data includes the drug pharmacological profile (including pharmacokinetics and pharmacodynamics information) and published clinical and non-clinical research findings related to the test of hypothesis. The output of this component which is fed by to the Clinical Trial Management Tool is a list of patients whose Core Profile indicates that they are eligible to participate in the clinical trial, with the term eligibility covering all three aspects defined in section B.

- **Ontology-Based Search Engine**: this entity incorporates innovative semantic data mining techniques for retrieving “answers” to the “queries” submitted by the medical researcher from published literature, available databases as well as EHRs. This component operates on top of a set of medical ontologies covering drug categorization and attributes (such as side-effects, interactions and absorption), disease classification and properties, geographical locations, available laboratory examinations.

- **“Cloud of Data”**: the “Cloud of Data” includes all the data mechanisms that enable the transparent and secure access to the data sources which are also part of it. Thus, it includes a “network of ontologies”:
  - **Drug Knowledge Base**: it stores and manages the drug ontologies that enable the proper interpretation of the data retrieved from the data sources. These ontologies need to analyse the drug terms and include a description of the drug mechanism as well as the pharmacodynamics and the pharmacokinetics information. Hence, for example, the term pharmacokinetics is further analyzed into absorption, distribution, metabolism, excretion and liberation, whereas drug action needs to be described such as depressing, stimulating, destroying cells, irritation, replacing substances.
  - **Disease Knowledge Base**: this base stores and manages the disease ontologies which allow for the semantic search of the data sources for disease-related data. These ontologies aim at providing a structure of the medical knowledge related to diseases, such as the epidemiology of diseases, mortality, mechanism.
  - **EHR Knowledge Base**: it stores and manages ontologies which describe the content of Electronic Health Records (EHR) with its main aim being the interpretation of the EHR records in the hospitals connected to the system and their mapping to the system.
  - **Clinical Trial Knowledge Base**: it stores and manages ontologies which describe the clinical trial protocol in order for the data retrieved from the published clinical data sources to be mapped into the system knowledge.
  - **Knowledge Base Gateway**: it acts as a gateway between the various knowledge bases presented above and the system as well as the “Cloud of Data” and the system. Moreover, it serves the heavy task of mapping the several ontologies stored in the “Cloud of Data” which each other. This task comprises one of the most important tasks in the “Cloud of Data” as most ontologies are based on different encodings of data posing several restrictions and limitations when applying more than one in answering a “query”. as well as:
    - **(Standardized) Interoperability Layer**: it enables transparent access to underlying heterogeneous data resources by virtualizing them through a set of web service interfaces.
    - **Data sources**: which mainly include Disease Bases, Drug Bases, Clinical and Non-Clinical Research Data Sources and EHR data sources which are connected with the system and communicate with the latter as part of the “Cloud of Data” in a SOA-oriented manner.

When the investigator is designing a clinical trial through the presented system, they perform queries through the Clinical Trial Management tool using keywords such as “mortality”, “aspirin”, “myocardial infraction”. Then, the system through the “Ontology-Based Search Engine” analyzes this query by using the disease and drug ontologies managed by the Disease and the Drug Knowledge Bases in the “Cloud of Data”. The analysis results are fed as queries through the Standardized Interoperability Layer to the various data sources (Clinical and Non-Clinical Research Findings, Disease Knowledge Bases and Drug Databases) and returns the results to the investigator following the same path reversed.

The investigator can further search upon related clinical studies which have examined or are currently investigating the safety and/or the efficacy of the investigational drug on the study treatment in order to navigate through inclusion/exclusion criteria and endpoints that the rest of the research community working on the same test of hypothesis is interested in. In order to avoid reducing the sample population to a small number of patients among which the recruitments effort will focus, every time the researchers set a new inclusion and/or exclusion criteria, the system automatically and rapidly filters the available EHRs in order to have an indication of the available sample population. In case a selection of a specific criterion tremendously reduces the target population, the system prompts a notification to the investigators indicating this reduction in order for them to be able to re-adjust their decision (if possible) for achieving broader validation of the clinical study results.
After the design of the clinical trial (including setting the inclusion/exclusion criteria, clinical study objectives and endpoints, expected study duration and periods) is finalized, the system filters the available EHR data sources in order to find patients eligible to participate in the specified clinical trial. The rules for reaching such a decision include:

- the inclusion criteria,
- the exclusion criteria,
- correlation of comorbidities and/or allergies and medical history of persons satisfying the criteria above and the profile of the investigational drug in order to identify possible side-effects which could be deterrent concerning their participation for safety reasons,
- mortality data (if any) from other clinical studies that are semantically searched upon based on the test of hypothesis of the current clinical study.

The investigator is prompted with a list of pseudonymized subject data of patients eligible to participate in their clinical trial. Pseudonymization is aimed at for data confidentiality reasons. As health data comprise sensitive personal data, special data management mechanisms need to be applied in order for ensuring that patient privacy is not comprised. The process for the investigator approaching the patients for informing them about the potential of their participation in their clinical trial strongly depends on the legal framework of the country(-ies) sharing EHR data with the system.

IV. IMPACT

During the patient selection process the clinical profiles of a great number of patients are automatically processed and analyzed based on the clinical trial design specifications set by the investigators, resulting in tremendous reduction of the time required for deciding upon the patients to be recruited, related costs and effort saving and great acceleration of the clinical research agenda. Moreover, providing proper linking of clinical trial systems with the wealth of patient information within healthcare provides representative samples of patients eligible to participate in the clinical trials and thus the validity of the clinical studies findings is boosted and the convergence between clinical trials and real world application of the researched treatments is significantly increased, leading to a great mitigation of patient safety risks stemming from the application of the new treatments.

Proper clinical trials’ design comprises a key success factor for the clinical trials’ outcome. The presented system enables well-designed clinical trials by incorporating findings of other related clinical and non-clinical research activities, published clinical evidence and the drug and disease profile, as well as by offering instantaneous access to this information to the medical researchers both during clinical trial design and clinical trial execution.

The automatic selection of representative samples of patients eligible to participate in well-designed clinical trials forms the basis for achieving enhanced clinical trial design efficiency. The system facilitates faster and more reliable decisions to either kill or advance compounds. Hence, this solution aligns directly with the goal set by drug regulatory organizations to ensure that basic scientific discoveries translate more rapidly and reliably into new and better medical treatments, resulting in improved healthcare and increased patients’ safety. This way, the new therapy may reach thousands of patients much sooner than nowadays.

As already mentioned, clinical trials are accounted for about one third of the costs of drug development, with almost one half of these costs being attributed to the time required for the completion of the clinical trials. Thus, this reduction in the timeline of a drug reaching the market combined with the boosted validity of the clinical trials findings through the platform will subsequently lead to major reduction of costs. In the meanwhile, early kill decisions will allow the sponsors to reallocate people and money to other development programs promising more benefit.

Moreover, the pre-clinical research costs represent the largest opportunity costs for firms in drug development. The platform will have great impact towards the reduction of these opportunity costs, through the early identification of failure, reduction of the time spent on pre-clinical phase and improvements of success rates. Thus, costs related to missed market opportunities for potential blockbusters that pharmaceutical companies and medical research organizations are facing the past years due to great operational delays will be significantly reduced, whereas these resources can then be transferred to other medical research areas which are otherwise put in lower priority due to lack of funding.

Reduction of duplicated data entries through proper information linking, acceleration of the patients - eligible for participating into the clinical trial - identification process, accurate recruitment to clinical trials, faster translation of research discoveries into new and improved healthcare and eventually reduced healthcare costs of tomorrow and more affordable medication are only some aspects of the platform that will enable cost reduction within clinical trial conduction. In fact, cost-effectiveness comprises one of the key aspects of the platform. For this reason, the selection process of patients’ eligible to participate in the clinical trials will also take into account the risk profile of each patient, thus further reducing the clinical trials cost.

Overall, the platform is expected to have great impact on clinical trials conduction – an important, time consuming and extremely costly process within drug development – towards the increased validity of their findings and significant cost and operational delays reduction which will in turn offer a high return to the European society through improved and more cost-effective healthcare systems, significant mitigating of patient safety risks, enhanced quality of life, healthier
European citizens and eventually an expanding economy of the various European medical industries.

V. CONCLUSION

This paper presented an innovative solution within the clinical research domain which aims at bridging the gap between clinical research and clinical practice by enabling the semantic interoperability between these two highly related, although currently isolated, domains. The approach is based on the fact that dispersed, heterogeneous information, which is to a great extend publicly available, can prove to be extremely valuable for the clinical study design process. Moreover, it considers the Electronic Health Record to be an electronic umbrella of medical information with the potential of speeding up considerably and advancing in terms of sample representativeness and patient safety the subject selection and recruitment process. The solution incorporates a SOA-oriented approach combined with the exploitation of a network of ontologies which forms a first “intelligence” layer for interpreting and analyzing existing, yet not easily or efficiently searched upon with current means, important information. The social, financial and technological impact of this solution is presented which indicates that the latter, by allowing for clinical study results of high external validity and faster and more reliable go/no-go decisions in the drug development process, the pharmaceutical sector finds a new way out of the current shrinking in drug development, whereas new therapies will be available for the citizens much sooner than nowadays.

ACKNOWLEDGMENT

This work is being supported by the PONTE project and has been funded by the European Commission’s IST activity of the 7th Framework Programme under contract number 247945. This paper expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this paper.

REFERENCES