IT-Aided Business Process Enabling Real-time Analysis of Candidates for Clinical Trials

Matthieu-P. Schapranow, Cindy Perscheid, Hasso Plattner
Hasso Plattner Institute
Enterprise Platform and Integration Concepts
August–Bebel–Str. 88
14482 Potsdam, Germany
{schapranow|cindy.perscheid|plattner}@hpi.de

***Abstract*—Recruitment of participants for clinical trials is a complex task involving screening of hundreds of thousands of candidates, e.g., testing for trial-specific inclusion and exclusion criteria. Today, a significant amount of time is spent on manual screening as improper selected candidates have impact on the overall study results.**

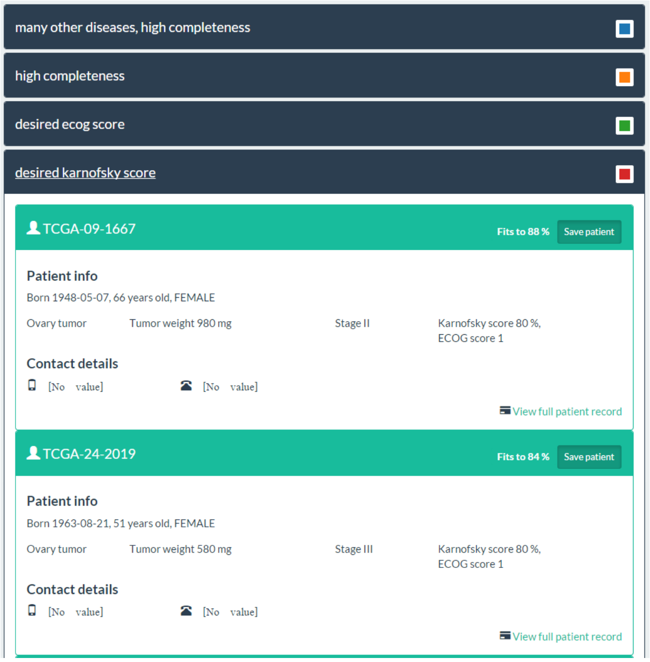
We introduce a candidate eligibility metric, which allows systematic ranking and classification of candidates based on trial-specific filter criteria in an automatic way. It is implemented as part of our web application, which enables real-time analysis of patient data and assessment of candidates. Thus, the time for identification of eligible candidates is tremendously reduced whilst additional degrees of freedom for assessing the relevance of individual candidates are available.

Keywords-Clinical Trials; In-Memory Technology; Data Analysis; Eligibility Metric; Clustering.

I. Introduction

The pharmaceutical industry heavily invests in clinical trials to improve existing drugs and introduce new ones every year [1]. The Guidelines for Good Clinical Practice (GCP) define details for conducting clinical trials on human subjects, which are widely adopted in certain countries, e.g., the United States of America (USA) and those of the European Union (EU) [2]. Amongst others, they define document artifacts, e.g., Informed Consent Form (ICF) or Case Report Form (CRF), as well as interaction between involved actors, e.g., investigators, sponsors, and trial participants. Before participation in a clinical trial, candidates need to be tested for certain trial inclusion and exclusion criteria.

In the given work, we introduce an integrated software-aided business process supporting the determination of eligible candidates for clinical trials building an additional source of confidence. Our software addresses principal investigators during the design of clinical trials and investigators during the identification of candidates. Our approach incorporates latest In-Memory Database (IMDB) technology to support real-time analysis of patient data during this phase. As a result, our approach contributes by reducing the overall time for determination of candidates, which consumes a significant amount of the overall clinical trial time today [3]. Figure 1 depicts the result view of our web application with four result clusters while a ranked candidate list for the cluster "desired Karnofsky score" is selected.



many other diseases, high completeness 🟦

high completeness 🟧

desired ecog score 🟩

desired karnofsky score 🟥

TCGA-09-1667 Fits to 88 % Save patient

Patient info

Born 1948-05-07, 66 years old, FEMALE

Ovary tumor Tumor weight 980 mg Stage II Karnofsky score 80 %, ECOG score 1

Contact details

[No value] ☎ [No value]

View full patient record

TCGA-24-2019 Fits to 84 % Save patient

Patient info

Born 1963-08-21, 51 years old, FEMALE

Ovary tumor Tumor weight 580 mg Stage III Karnofsky score 80 %, ECOG score 1

Contact details

[No value] ☎ [No value]

View full patient record

Figure 1: Result view of our web application showing four clusters of candidates each containing an automatically assessed list of candidates for review.

The rest of the work is structured as follows: In Section II, our work is set in the context of related work whilst we share a real-world use case in Section III. We explain our incorporated methodology in Section IV, demonstrate how it can improve the clinical trial process in Section V, and introduce our IMDB approach in Section VI. In Section VII, we discuss our findings and our work concludes with an outlook in Section VIII.

ISBN: 978-1-61208-424-4

II. Related Work

The given contribution addresses a) a software solution improving filtering and assessment of patient data and b) a business process integrated the software. Selected related work addressing software solutions for candidate identification are discussed in the following.

The Veterans Health Information Systems and Technology Architecture (VistA) of the U.S. Department of Veterans Affairs is a Hospital Information System (HIS) combining data from distributed VA clinics and sites in a single data source [4], [5]. Although it contains data for candidate identification, it is not used this purpose as it does not provide required tools.

The Informatics for Integrating Biology and the Bedside (i2b2) system provides IT tools for clinical researchers for research purposes [6]. It provides a configurable and interactive query editor supporting a range of filters, including diseases, medications, laboratory tests, and doctor's visit details [7]. Our contribution provides additional enhancements, e.g., an eligibility metric to rank and cluster results, as defined in Section VI.

Dumas et al. identified Acute Myeloid Leukemia (AML) candidates for clinical trials using clinical data from of patients taken from the HIS of a German hospital and compared them to inclusion and exclusion criteria of ongoing clinical trials [8]. We also believe that the secondary use of existing clinical data is beneficial for candidate identification. Thus, our contribution defines an IT-aided business process for candidate identification for a wide range of diseases.

All aforementioned approaches lack support for assessment of results. In contrast, our approach systematically calculates a score for each candidate enabling ranking and clustering of the result set for the first time.

III. Use Case

We defined the following persona as a concrete example for our enhanced business process in the remainder of this work. Forrest G., male, 62 years old was an active smoker for a long period in his life. During one of his regular checkups, he was recently diagnosed with Non-Small Cell Lung Cancer (NSCLC). Now, Forrest worries about receiving the best available treatment for NSCLC. Thus, he learned about targeted therapies and is very interested in participating in clinical trials optimized for his personal disease. As a result, selected details of his Personal Health Record (PHR) are shared in a de-identified way to assess the individual eligibility for clinical trials. We refer to a fictive clinical trial targeting NSCLC requiring at least 150 g available tumor tissue for preliminary testing.

IV. Methods

In the following, we share details about our incorporated methodology: in Section IV-A we introduce our eligibility metric defining a unique key figure per candidate representative for its calculated trial applicability whilst the incorporated IMDB technology to leverage interactive data analysis of big patient data in our software artifact is introduced in Section IV-B.

A. Candidate Eligibility Metric

We define a vector of n candidate criteria $\vec{v} = (c_1, \ldots, c_k, \ldots, c_n)$ with $c_i \in [0,1], i \in \{1, \ldots, n\}$ while each of the vector components is calculated by an individual function, as described in Section VI-E. Thus, each candidate is represented as a point in a n-dimensional vector space where the most suitable candidate is defined by the vector $(1.0, \ldots, 1.0, \ldots, 1.0)$.

We define the eligibility score s_k of a candidate k as the normalized Euclidian distance $d(\vec{v_{1.0}}, \vec{v_k})$ between the vector of the most suitable candidate $\vec{v_{1.0}}$ and the vector of the individual patient $\vec{v_k}$, as defined in Equation 1.

$$
\begin{aligned}
s_k &= 1.0 - \frac{d\left(\vec{v_{1.0}}, \vec{v_k}\right)}{d\left(\vec{v_{1.0}}, \vec{v_{0.0}}\right)} && (1) \\
&= 1.0 - \sqrt{\sum_{i=1}^{n} \left(\frac{v_{1.0_{c_i}} - v_{k_{c_i}}}{v_{1.0_{c_i}} - v_{0.0_{c_i}}} \right)^2} && (2)
\end{aligned}
$$

B. In-memory Database Technology

We refer to IMDB technology as a toolbox of Information Technology (IT) artifacts, which enables processing of enterprise data in the main memory of server systems in real time [9]. Through the combination of IMDB database technology and analysis of available candidate data, we aim to achieve a speedup for the time-consuming identification of candidates, as described in Section I. In the following, selected building blocks of the IMDB technology are introduced.

1) Column-Oriented Data Layout: Most modern relational database systems fall into the category of transactional databases and store their data in a row-oriented format, i.e., all attributes of a record are stored in adjacent blocks. This is advantageous if the complete data of a single row has to be processed. On the other hand, analytical database systems store and process their data column-wise, i.e., all entries of a column are stored in adjacent blocks, which is beneficial if only selected attributes of a data set need to be accessed. When filtering patients based on a study's criteria, only certain parts of their data need to be read. The types of queries to be expected in our prototype can therefore benefit from this data layout.

2) Lightweight Compression: Lightweight compression refers to a data storage representation that consumes less space than its original pendant [9]. Storing data column-wise facilitates lightweight compression techniques, such as run-length encoding, dictionary encoding, and difference encoding [10]. The diverse nature of patient data results in heterogeneous data, i.e., many NULL values, facilitating the potential to save space through encoding.

3) Partitioning: Our incorporated IMDB provides vertical and horizontal partitioning [11]. The former addresses large database tables and splits up a database table in multiple column-wise subsets that can be distributed on individual servers while the latter divides a long database table in smaller subsets of data [12]. Splitting data into equally long horizontal partitions supports parallel search operations and improves scalability [9]. For our use case, partitioning enables the use of multiple sources of candidates to increase the reach of the system [13].

4) Multi-Core and Parallelization: Modern system architectures are designed to provide multiple CPUs with each of them having separate cores. This capacity needs to be fully exploited by parallelizing application execution to achieve maximum processing speed. Internal tools of IMDBs are implemented to benefit from parallelization. By using the capabilities of our database, we can speed up the process of clustering candidates.

5) Bulk Data Load: For candidate identification, large amounts of clinical data have to be collected and stored. The bulk load capabilities of IMDBs support this through the use of the CSV format and parallel processing of the data to insert.

C. Realistic Patient Data

We consider the use of realistic patient data for development and testing as the foundation to optimize software for real-world use cases. As a result, we incorporate patient data from The Cancer Genome Atlas (TCGA) program in the course of our development [14].

D. Design Thinking

Together with subject matter experts from research and industry, we derived design decisions for our web application and the enhanced process. For that, we incorporated the design thinking methodology, which helped to stratifying the cooperation in an interdisciplinary team [15]. Consequently, we performed regular user interviews sharing our research artifacts to constantly improve our approach.

V. The Clinical Trials Process

We define the following phases in the context of clinical trials as depicted in Figure 2:

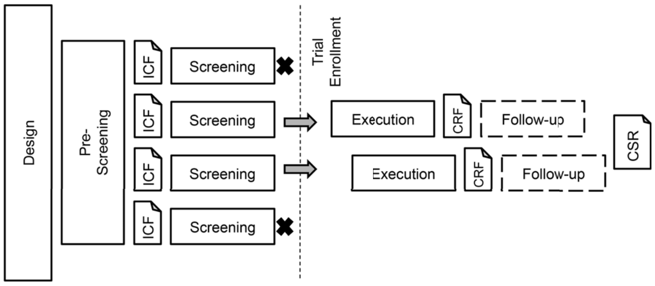


Figure 2: Phases of a clinical trial.

- **Design**: Inclusion and exclusion criteria need to be defined during this phase. Reviewing a list of potential candidates supports investigators in selected criteria to include a representative population.
- **Pre-Screening**: Candidate's data is scanned and they are contacted if they fulfill the criteria listed in the clinical trial synopsis.
- **Screening**: After a comprehensible explanation of the trial's purpose, candidates need to consent their participation by signing the ICF. A doctor, e.g., by conducting medical tests, checks them. Iff a candidate fits all required criteria of the trial definition, she/ he gets enrolled.
- **Execution**: The actual trial execution starts, e.g., intake of pharmaceuticals or placebos. After the participant completes the trial the individual case report is created. The Clinical Study Report (CSR) is created after the overall trial is finished and the trial database was locked.
- **Follow-up**: Study participants are followed up with satisfaction surveys and further assessments, e.g., regular health checks.

A. Design of Study Protocols

Inclusion and exclusion criteria are defined during design of a study protocol to identify adequate candidates representing the population [16]. Improper design decisions may result in complex or inadequate selection of candidates affecting the quality of the study results [17], [18]. Our contribution supports trial investigators, e.g., they can examine all candidates and evaluate defined inclusion and exclusion criteria by simulating their impact on candidate identification using real data.

B. Identification of Eligible Candidates

The analysis of patient-specific data and their matching with study-specific inclusion and exclusion criteria is performed during the pre-screening phase, as defined in Section V. Nowadays, this involves time-consuming and manual analysis of individual trial criteria for hundreds of thousands of candidate profiles.

Our enhanced process improves the manual process by defining an eligibility score for each candidate, as defined in Section IV-A, representing the applicability in context of the study. The process consists of the following phases:

- Configuration of filter criteria,
- Automatic analysis of candidate data, and
- Review of results.

During configuration of filter criteria, investigators define criteria accordingly to the inclusion and exclusion of the clinical trial on the filter screen of our app. The specified filter criteria are translated into a database query matching the selected filter criteria.

During data analysis, individual candidate records are screened to meet defined filter criteria. This process is performed completely within the incorporated IMDB, which eliminates the need for any application-level filter and optimizes run time.

During the review phase, the investigator accesses the result screen, which consists of a list of clusters, each of them containing a ranked list of trial candidates with a specific matching score for the current trial. Candidates with a similar score but with different criteria are assigned to individual clusters, which outlines how candidates' criteria differ.

For each candidate, all accessible data can be directly expanded in the result view in order to assess follow-up questions directly during the manual review phase. The result view also provides interactive graphical analysis features using individual criteria, e.g., graphs and figures about age, gender, or tumor weight distribution. If a potential candidate is found, she or he is added to a persisted list of candidates to follow-up, e.g., by downloading the candidate list to inform them about their eligibility for the clinical trial.

VI. Contribution

In this section, we share implementation details of our software application for identification of trial candidates matching trial-specific filter criteria.

A. Database Schema

We defined an extendable star database schema for patient data optimized for data analysis [19]. It consists of the fact table `CLINICAL_PATIENT` containing unique master data about the patients, which is referred to by multiple dimension tables as depicted in Figure 3. For example, the fact table contains birth date, gender, and ECOG or Karnofsky score to quantify the health status of a person [20], [21]. Dimension tables store additional optional patient data using $n : m$ relations, e.g., the dimension table `DRUG` contains details about medication intake, such as duration and dosage.

For our research prototype, we incorporated TCGA data for indications of lung and ovarian cancer and consolidated it. The `LUAD` table contains details about lung cancer, e.g., tumor location or active smoking time, whilst the `OV` table contains ovarian-specific details.

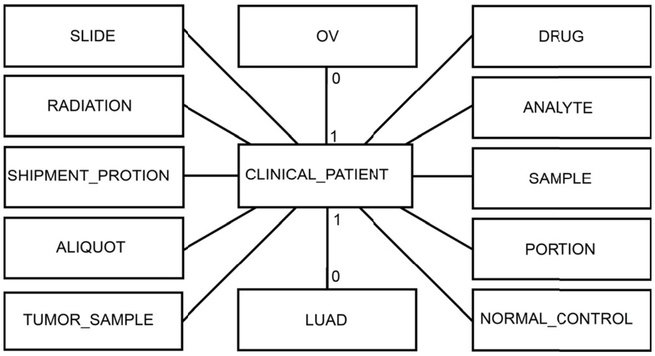
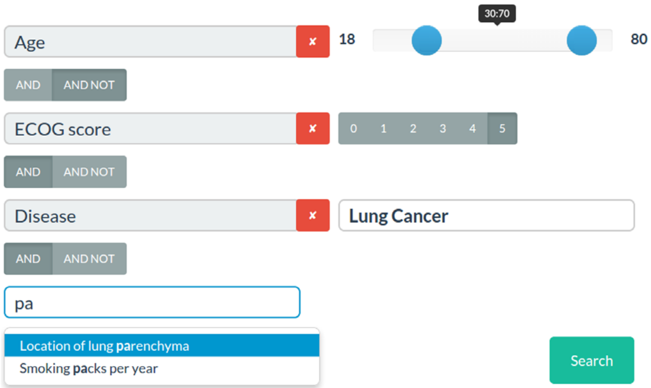


Figure 3: Entity-relationship diagram of our star database schema. It consists of the fact table `CLINICAL_PATIENT` holding general patient data and dimension tables holding additional data, e.g., drug intake or tumor sample. Cardinality of all relations is $1 : *$ unless marked.


Age 18 30:70 80

ECOG score 0 1 2 3 4 5

Disease Lung Cancer

pa

Location of lung parenchyma
Smoking packs per year

Figure 4: Filter criteria creation in our web application: Three filters have already been created, and the user is creating a forth one using the auto-completion.

We defined an Extract, Transform, Load (ETL) process for import of TCGA data into the database [22]. TCGA is downloaded and divided into multiple raw data files, which are transformed by a Python script into CSV files. The latter is loaded into our database using its bulk load capabilities, as described in Section IV-B5.

B. Backend and Frontend

Our web application backend consists of a lean Ruby server with Sinatra serving the static content of web pages only [23]. All analysis operations, e.g., clustering and ranking, are performed directly within the IMDB, which eliminates the need to transfer data through the application stack. We designed a web application exchanging user data between backend and frontend using Asynchronous JavaScript and XML (Ajax) [24].

C. User Interface

Our User Interface (UI) incorporates a responsive design, i.e., display optimized for individual device classes, e.g., desktop PCs and mobile devices.

Figure 4 depicts the filter definition view of our web app for the use case defined in Section III. On the left-hand side, inclusion and exclusion filter criteria are specified. Our app provides a list of relevant filter criteria using auto-completion while typing. Filter criteria are added either as Boolean operators `AND` or `AND NOT`. On the right-hand side, filter-specific values are defined.

Figure 6 shows a matrix of user-defined filter criteria for graphical exploration. The color of the dots indicates the cluster, where the value belongs to, e.g., the red cluster indicates candidates that are close to the desired ECOG score while having only a few other diseases. Candidates with a high result from the most suitable candidate function are most eligible from a data's point of view.

Candidates are grouped in similar clusters, where each cluster consists of a ranked list of candidates with specific information. Investigators can inspect all available candidate information with a click on the "View full patient record" link and store them in a follow-up list using the "Save patient" button.

D. Filter Types

Together with subject matter experts, we defined filter types supporting the identification of trial candidates. In the following, selected filter types of our web application are described in detail:

- **Single option filters** can have either of two possible values, e.g., gender: ♀ or ♂,
- **One of many options filters** describe one of multiple statuses, e.g., cancer stages I to IV,
- **Range filters** define a continuum of values, e.g., age between 18 and 45,
- **Threshold filters** must be above or below a certain value, e.g., a minimum tumor weight, and
- **Free-text filters** provide a text field with auto-completion for selection of various values, e.g., diseases or previous medications.

We distinguish hard and soft filters, where the former do not allow outliers and the latter consider incomplete or imprecise data points. Soft filters assign lower scores to outliers instead of removing them completely. If data relevant for a filter is missing, e.g., the tumor weight of a candidate is unknown, s/he will receive a lower score, but will not be removed from the result.

E. Ranking

For candidate ranking, we incorporate our vector-based eligibility metric, as described in Section IV-A. The most suitable candidate has the value one in all components of the vector. For each candidate in the result set, we calculate the distance between the candidate-specific and the most suitable patient's vector. The ranking process is based on the vector space model [25].

Table I: EXCERPT OF PATIENT DATA.

Candidate	Gender	Age	Diagnosis	Tumor Weight
1	♂	45	NSCLC	242 g
2	♂	51	NSCLC	n/a

For each vector component, an individual function implements the comparison of candidate-specific data with the data defined for the most suitable candidate and returns a decimal value within the interval $[0, 1]$.

Ranking functions for individual filters are implemented differently depending on their type: there are functions for hard and soft filters as well as a special implementation for threshold filters. Apart from filter-specific functions, the completeness and most suitable candidate functions are always executed per candidate.

1) Completeness: Patient data may be incomplete, i.e., some data is missing or unavailable in the candidate's profile. The completeness function assigns a higher score to candidates having all matching attributes available.

Let us consider the excerpt of patient data in Table I for the use case described in Section III. The completeness score is defined as mean of all available attributes, i.e., the number of available and matching facts divided by the number of requested facts. The completeness score is 1.0 for the first and $\frac{2}{3}$ for the second candidate. Thus, the completeness function ranks the first candidate higher than the second, who could still match all trial criteria although the tumor weight is unavailable.

2) Most Suitable Candidate: The most suitable candidate function defines a certain penalty for candidates with additional indications than the requested ones to take eventual issues into account. We defined an individual penalty score for each disease according to the degree of disease, e.g., having lung cancer is worse than having asthma. The score is calculated by subtracting the maximum of all disease penalties from 1.0.

Let us consider the use case described in Section III. NSCLC without any additional indications results in a score of 1.0, a candidate suffering from asthma receives a score of 0.8, and a candidate, who just recently suffered a heart attack, receives a score of 0.5.

F. Calculation of Eligibility Score

Let us consider the use case, as defined in Section III, using the completeness, general health, and threshold functions to define the eligibility score s, as defined in

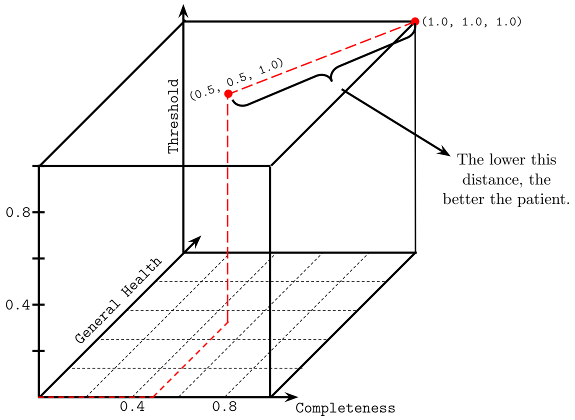


Figure 5: Example of calculating the ranking score for a candidate. The current candidate has the values (0.5, 0.5, 1.0). The most suitable candidate is located at the point (1.0, 1.0, 1.0). The normalized Euclidian distance between the current candidate and the most suitable candidate determines the total ranking score.

Section IV-A. Figure 5 depicts candidate k as $\vec{v_k} = (0.5, 0.5, 1)$ and the most suitable candidate defined by $\vec{v_{1.0}} = (1.0, 1.0, 1.0)$ as well as the distance from each other in the three-dimensional space. For candidate k not all required data is available, the health status differs from the requested one, and the requested threshold value is exceeded. Thus, the eligibility score s_k is $s_k = 0.5918$, i.e., candidate k is eligible for the clinical trial with 59.18% with regard to the selected criteria.

G. Clustering

Ranking candidates enables investigates to evaluate the eligibility of candidates with regard to trial criteria. However, two candidates may vary in different aspects while having the same eligibility score.

Thus, we perform a clustering of results, which assigns similar candidates to the same cluster. We selected the k-means clustering algorithm as it is the most appropriate algorithm for our purpose [26]. Clustering is directly executed within our IMDB, which eliminates the need for exporting the data, processing the data by third-party tools, and importing of results. As a result, we were able to leverage interactive data analysis and exploration of results for our application's end users.

While the target number of clusters must be configured, the algorithm can be applied to any number of dimensions, as the distance to the cluster centroids can be calculated in any multi-dimensional space. For our prototype, we configured the algorithm to return four clusters. This number is configurable and was chosen after consideration of our sample dataset. Because users of our application are allowed to vary the number of filters, the number of dimensions is not fixed.

In the frontend, the resulting list of candidates is grouped in clusters, which are given appropriate names based on frequent properties. Additionally, the clusters are visualized in a projection to each pair of two dimensions as depicted in Figure 6.

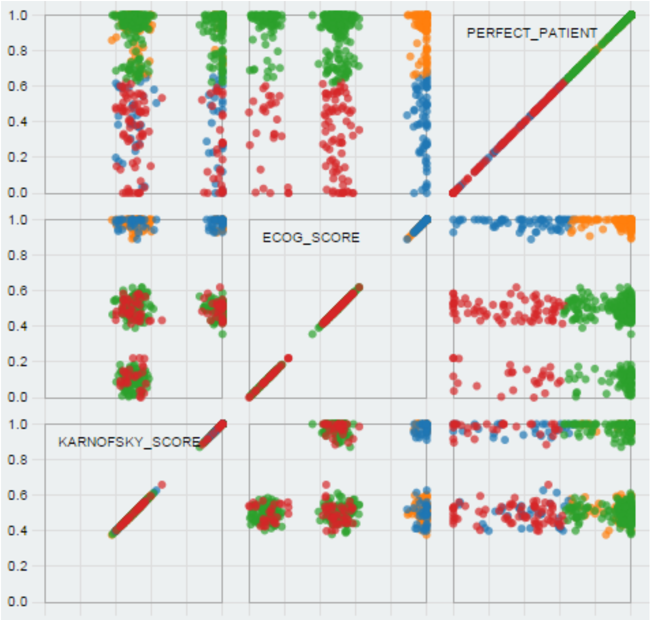


Figure 6: Visualization of multi-dimensional clustering results. Each block shows all candidates, projected onto two dimensions. For example, the bottom right square shows the position of all candidates with regard to Karnofsky score and the most suitable candidate function, as defined in Section VI-E2. Clusters are indicated by a point's color.

VII. Evaluation and Discussion

We configured our candidate eligibility metric for a concrete clinical trial addressing lung cancer patients, as described in Section IV-A. Our database was populated with real patient data taken from TCGA including various indications, such as lung cancer and ovarian cancer. Together with our experts from industry, we defined specific filter criteria to retrieve relevant subsets of candidates eligible for the specific clinical trial.

During our conducted user interviews, we received promising feedback that our enhanced business process will reduce the overall time to identify candidates for clinical trials.

Our enhanced process combines the automatic analysis of patient data with a manual review phase performed by a human expert. Thus, routined work is optimized through our software-aided web application while the accuracy of the outcome is guaranteed by the incorporated human reviewer.

VIII. Conclusion and Outlook

In the given work, we shared research results of our interdisciplinary cooperation of pharma experts, clinical trial teams, and software engineers. Applying the concrete use case of a clinical trial for lung cancer patients, we defined an enhanced business process incorporating IT artifacts, which enable the integrated analysis of patient data. Today, the identification of candidates requires manual interpretation and matching of patient and trial data, which results in a time-consuming and error-prone process, e.g., in course of phase III and IV clinical trials with more than thousand participants.

We defined a generic candidate eligibility metric in Section IV-A that is configurable per trial to reflect specific inclusion and exclusion criteria of the study synopsis. Furthermore, we introduced an enhanced IT-aided business process for identification of eligible candidates for clinical trials in Section V. Consequently, we shared in Section VI design decisions and implementation details of our web application supporting the analysis of large pools of patient data in real time.

Our future work focuses on the adaption of our candidate eligibility metric to further indications and the use of additional criteria. Furthermore, we are working together with our partners to establish our enhanced process as standard operating procedures for cost-effective identification of candidates for clinical trials.

References

[1] Y. Feyman, “Shocking Secrets of FDA Clinical Trials Revealed,” http://www.forbes.com/sites/theapothecary/2014/01/24/shocking-secrets-of-fda-clinical-trials-revealed/ (last accessed: Jun 1, 2015), 2014.

[2] World Health Organization, “Handbook for Good Clinical Research Practice (GCP),” http://apps.who.int/prequal/info_general/documents/GCP/gcp1.pdf (last accessed: Jun 1, 2015).

[3] S. J. Projan, “Why is Big Pharma Getting Out of Antibacterial Drug Discovery?” Current Opinion in Microbiology, vol. 6, no. 5, 2003, pp. 427 – 430.

[4] D. Hynes, G. Joseph, and C. Pheil, “Veterans Health Information Systems and Technology Architecture as a Research Tool,” VIReC Ins, vol. 3, no. 1, 2002, pp. 1–8.

[5] United States Department of Veterans Affairs, “VistA Health System,” http://www.vistahealth.com/vista-health-system/myhealthhomepatientportal.aspx (last accessed: Jun 1, 2015), 2014.

[6] S. N. Murphy et al., “Serving the Enterprise and Beyond with Informatics for Integrating Biology and the Bedside (i2b2),” American Medical Informatics Assoc., vol. 17, no. 2, 2010, pp. 124–130.

[7] Partners Healthcare, “i2b2 Software,” https://www.i2b2.org/software/index.html (last accessed: Jun 1, 2015), 2015.

[8] M. Dugas, M. Lange, W. Berdel, and C. Müller-Tidow, “Workflow to Improve Patient Recruitment for Clinical Trials within Hospital Information Systems: A Case-Study,” Trials, vol. 9, no. 2, 2008.

[9] H. Plattner, A Course in In-Memory Data Management: The Inner Mechanics of In-Memory Databases, 1st ed. Springer, 2013.

[10] P. Svensson, “The Evolution of Vertical Database Architectures: A Historical Review,” in Proceedings of the 20th Int’l Conf on Scientific and Stat Database Mgmt. Springer-Verlag, 2008, pp. 3–5.

[11] S. S. Lightstone, T. J. Teorey, and T. Nadeau, Physical Database Design: The Database Professional’s Guide to Exploiting Indexes, Views, Storage, and more. Morgan Kaufmann, 2007.

[12] J. M. Hellerstein, M. Stonebraker, and J. Hamilton, Architecture of a Database System, Foundation and Trends in Databases. now Publishers, 2007, vol. 1.

[13] H. Plattner and M.-P. Schapranow, Eds., High-Performance In-Memory Genome Data Analysis: How In-Memory Database Technology Accelerates Personalized Medicine. Springer-Verlag, 2014.

[14] National Cancer Institute, “The Cancer Genome Atlas Data Portal Overview,” https://tcga-data.nci.nih.gov/tcga/ (last accessed: Jun 1, 2015), 2015.

[15] H. Plattner, C. Meinel, and L. Leifer, Design Thinking Research, ser. Understanding Innovation. Springer, 2012.

[16] V. Cottin et al., “Small-Cell Lung Cancer: Patients Included in Clinical Trials are not Representative of the Patient Population as a Whole,” Annals of Oncology, vol. 10, no. 7, 1999, pp. 809–815.

[17] L. Lovato, K. Hill, S. Hertert, D. Hunninghake, and J. Probstfield, “Recruitment for Controlled Clinical Trials: Literature Summary and Annotated Bibliography,” Controlled Clinical Trials, vol. 18, no. 4, 1997, pp. 328–352.

[18] L. Marks and E. Power, “Using Technology to Address Recruitment Issues in the Clinical Trial Process,” Trends in biotechnology, vol. 20, no. 3, 2002, pp. 105–109.

[19] W. A. Giovinazzo, Object-Oriented Data Warehouse Design: Building a Star Schema. Prentice Hall Upper Saddle River, NJ, 2000.

[20] M. M. Oken et al., “Toxicity and Response Criteria of the Eastern Cooperative Oncology Group,” American Journal of Clinical Oncology, vol. 5, no. 6, Dec. 1982, pp. 649–656.

[21] D. A. Karnofsky, “The Clinical Evaluation of Chemotherapeutic Agents in Cancer,” Evaluation of Chemotherapeutic Agents, 1949.

[22] J. Singh, Understanding ETL and Data Warehousing: Issues, Challenges and Importance. Lambert Acad. Publishing, 2011.

[23] A. Harris and K. Haase, Sinatra: Up and Running. O’Reilly Media, 2011.

[24] A. Mesbah and A. van Deursen, “Migrating Multi-Page Web Applications to Single-Page Ajax Interfaces,” in Proceedings of the 11th European Conf on Software Maintenance and Reengineering. IEEE, 2007, pp. 181–190.

[25] G. Salton, A. Wong, and C.-S. Yang, “A Vector Space Model for Automatic Indexing,” Commun. ACM, vol. 18, no. 11, 1975, pp. 613–620.

[26] T. Kanungo, D. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Wu, “An Efficient K-Means Clustering Algorithm: Analysis and Implementation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, 2002, pp. 881–892.