

Visualizing Clinical Trial Data Using Pluggable Components

Jonas Sjöbergh, Micke Kuwahara, and Yuzuru Tanaka
Meme Media Lab
Hokkaido University
Sapporo, Japan
{js, mkuwahara, tanaka}@meme.hokudai.ac.jp

Abstract—We describe our system for visualizing data from clinical trials. The system is intended for clinicians with little or no knowledge of statistics, data mining, etc. The system is built using pluggable components, and it is easy to add more types of visualization. Currently the system supports interactive data exploration using for instance parallel coordinates, image maps, charts, and life tables. Grouping and filtering patients or subsets of data is easy and any changes are immediately reflected in all visualization tools currently used.

Keywords-visual analytics, information visualization, clinical trials

I. INTRODUCTION

Clinical trials are important for finding out what treatments work well against what medical problems. During clinical trials, huge amounts of data are collected. Computer support for clinical trials is of course very useful, and we have developed a system to support clinical trials, especially trials on cancer. In such trials, data is collected on many things, and a lot of it is not relevant to what results the treatment will have. What exactly has effects on the treatment is not known, of course.

The data quality is not always great. Many values in the data are “unknown”, we might for instance not know if anyone else in the patient’s family had a similar cancer for all patients. There are also many null values in the data base, and the meaning of these is sometimes different from the value “unknown” (though sometimes it just means that no such data was collected for this patient). For “Date of death”, patients who did not die could have null values while patients that did die but where we do not know the actual date may have “unknown”, for example. A lot of the data is not numerical or even ordered, so smoothing data or compensating for null values or other problems is not trivial.

Since the problem is not well formed, i.e. we do not have a good model of what data will have an effect on the cancer treatments, and since the data quality is not always great, automatic data mining or data analysis methods do not always give us all the insights we want.

We have created a system for clinical trials [1]. It supports all phases of a trial, designing the trial, running the trial, and analyzing the results of a finished trial. In this paper we describe the data visualization parts of our system. The data visualization in the system is aimed at supporting clinicians

so they can easily explore the data on their own. Clinicians may have intuitions or knowledge on what is likely to matter when it comes to cancer, so we would like to make it possible for them to explore the data and see if their intuitions are right or if they can get ideas for new things to examine.

The visualization has different types of components for visualizing the results of a trial, for example charts, parallel coordinates, life tables. It also has different components for selecting and filtering groups of patients or subsets of data to visualize.

II. SYSTEM

We use a system called the Trial Outline Builder [1], or TOB for short. It was developed to support clinicians when running clinical trials, and especially clinical trials on cancer. It supports designing the trial, data collection when running the trial, and data visualization of the data collected in a trial.

The system was developed in the EU FP 6 project “ACGT” (Advancing Clinico-Genomic Trials on Cancer: Open Grid Services for improving Medical Knowledge Discovery), and development is continuing in the FP 7 large integration project called “P-Medicine”, started in 2011.

The TOB runs as a plug-in inside another system, ObTiMA (Ontology based Trial Management Application) [2], that handles among other things data access restrictions (who is allowed to see what data). ObTiMA was also developed in the ACGT project.

The system is intended for users who are not very proficient with computers. It is a very visual system with direct feedback. The design of the trial layout is done by dragging and dropping medical events (such as radio therapy, randomization, surgery, etc.) from an event repository onto a work board. They can then be dragged around, copied, have their properties edited, etc. by simple interaction. Currently, we are told that most trials are designed using pen and paper, so even a simple work flow editor like this seemed helpful to the clinicians we showed it to.

When designing a trial you also attach data input forms to the events in the trial, and when running a trial it is easy to open the appropriate input forms just by clicking on the events in the patients treatment plan. The interface is

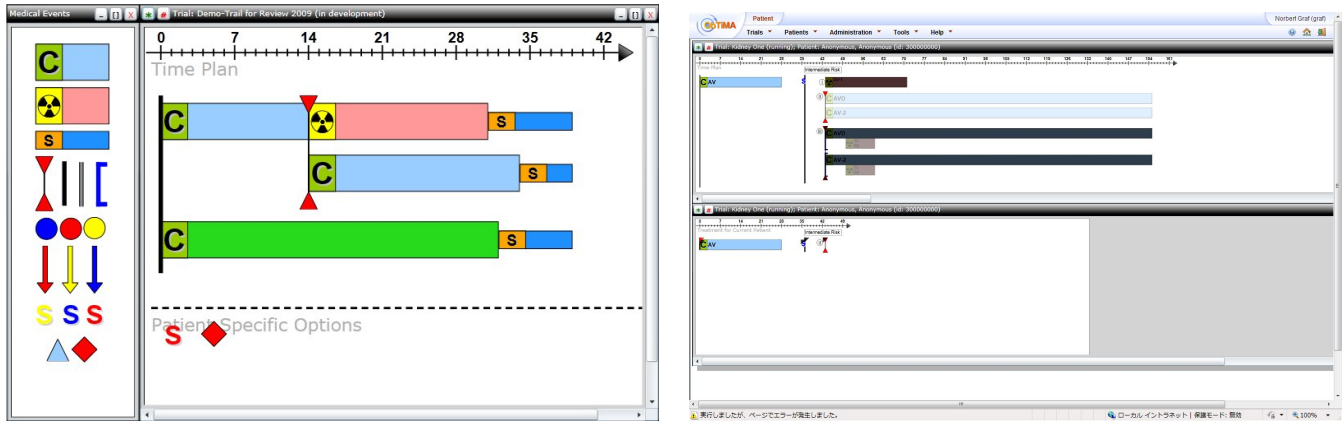


Figure 1. **Left:** The TOB when designing a trial. Medical events are dragged from the repository window on the left, and dropped into the treatment plan on the right. **Right:** TOB showing the overall treatment plan (top) together with the personal treatment plan of a specific patient (bottom). Paths in the treatment plan that the patient cannot reach are shaded out in black. Parts that are still in the future are shaded with white. Clicking on future events activates them and puts them in the patient's own treatment plan. Clicking on events in the patient's plan bring up input forms related to that event.

consistent throughout all phases of a trial, so the events and trial outline look the same when running the trial as they did during the design phase.

After the trial is finished, the system also supports visualization of the collected data. More on this in the next section.

The system is built using “meme media” [3], [4] pluggable components, a kind of intelligent software objects. Meme media is intended to make sharing and reuse of services and functionality as easy as sharing and reuse of data resources such as text and images is with standard technology now. This makes it easy to add new functionality to our system, by simply adding new pluggable components.

We developed several different components for selecting and grouping data, and these can be used separately, used together, or one type easily replaced by another, thanks to the pluggable component architecture. The same is true for the different visualization components that can display data in different ways. They can be used together, or easily interchanged, and they of course work together with the selection and grouping components.

Our system was developed with repeated feedback from clinicians. Adding new functionality when the clinicians came up with new things they would like to have in a clinical trial support system, such as adding Life Tables for visualization when showing a simple visualization tool, turned out to be easy. Very little extra work was needed to add new functionality to the running system, or to prototype new ideas, so the meme media concept worked well.

III. DATA VISUALIZATION, OVERVIEW

Data visualization is divided largely into three steps in our system. First you select what data to use from the huge amounts of data collected in the trial. This is done by simply clicking on an overview of the trial to bring up the data from

the part of the trial that you click, i.e. clicking a surgery event brings up all the data fields that were collected for the surgery. Selecting data fields from the trial outline brings components representing these data onto a work board.

On the work board, you can group the data, and filter out data that you are not interested in. You could for instance group patients into groups based on which treatment they received or what treatment group they were randomized into. You could then filter out patients with metastatic cancer, or look at only patients that share some genetic trait.

Finally, you can choose how to visualize the data. There are scatter plots, bar charts, etc. available. Clinicians usually prefer to see the data as Life Tables, plotting the number of patients still alive in each group as a function of the time from diagnosis.

IV. GROUPING/FILTERING: PARALLEL COORDINATES AND IMAGE MAPS

We use parallel coordinates [5], [6] both for visualizing data and for grouping and filtering. Parallel coordinates have been used to visualize medical data by others [7]. The parallel coordinate system has one dimension for each data field selected to be displayed on the work board. A patient is shown in the system as one line from each parallel coordinate to the next. Figures 2, 3, and 4 show parallel coordinates, and patients as lines between them.

Patients can be filtered out by sliding the triangular objects on the parallel coordinate line. Patients between the triangles are kept, and patients with values outside the triangles on any parallel coordinate are hidden. It is possible to have multiple sets of filter triangles, to keep and remove patients in non-continuous groups.

Patients can also be grouped and the groups displayed in different ways in for instance graphs. All patients passing

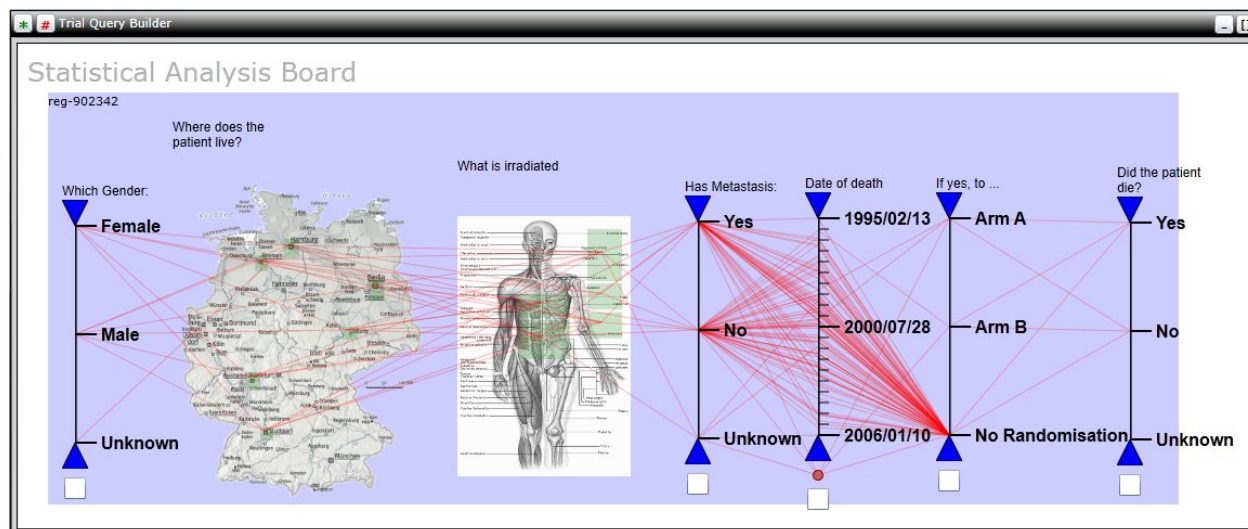


Figure 2. Two Image Map components together with Parallel Coordinate components. Each patient is represented by a line, so a female patient from Berlin and irradiated in the chest would be represented by a line consisting of line segments from the “Female” part of the first coordinate to Berlin on the map, and then a line segment from Berlin to the chest of the human figure, and so on.

through the same sets of triangles on all parallel coordinates are considered to be in the same group.

The order between the parallel coordinates is important when trying to see connections between different dimensions. In our system any parallel coordinate can easily be moved by just dragging it to a new position, thus enabling reordering of the coordinates.

New parallel coordinates can also be created by composition of other coordinates. For numerical data and dates, simple mathematical composition is possible. For instance creating an “Age when entering trial” coordinate by taking the difference between “Date of birth” and “Date of entering trial”. For textual data the values are just concatenated, though we plan to add things such as testing equality between values in the future.

Parallel coordinates work very well for numerical data or dates. For for example textual data, the parallel coordinates can still be used but the order on the parallel coordinate is arbitrary, and other visualization methods might be more appropriate. Another visualization component that works very much like the parallel coordinates is also available. It is an image map, where different data values correspond to different parts of an image. This can be used with for instance geographical data (data points correspond to points on a map). In Figure 2 data on where a patient was irradiated (abdomen, lungs, not irradiated, or unknown) is shown using an image of the human body, and there is also another image map that is showing where patients live on a map.

Patients can be filtered by clicking on areas of the image to add or remove patients whose values correspond to that area, and patients can be grouped by selecting sets of areas together.

These image maps can also be used together with the parallel coordinates, and the image maps behave like any parallel coordinate. This means for instance that lines in the coordinate system pass through the images too. The images can be dragged around to change the order between the coordinates, and they can be scaled.

V. VISUALIZATION: GRAPHS AND LIFE TABLES

Data can be visualized using many types of graphs and charts. All “standard” types of graphs have been wrapped into a graphical component, and it is easy to show for instance a bar chart showing how many patients had metastatic cancer and how many did not, or how many patients in each randomization group had metastatic cancer etc. Since these are standard visualization components, we will not dwell on them in this paper, though an example is shown in Figure 4.

Clinicians prefer to show the trial results using Life Tables. This means showing the number (or percentage) of patients still alive as a function of time into the trial (e.g. after a certain number of days from the date of diagnosing the cancer).

An example with a life table is shown in Figure 3, showing that patients in one group die to a much larger extent than in the other two groups, but if the patient is alive after about two years from the time of diagnosis, the likelihood of dying in the next ten years is very low for all groups.

The upper lines (that plane out to horizontal lines) show the optimistic estimate of the percentage of patients alive. Every time a patient dies and the data is entered in the database, these curves go down. If nothing else is known, patients are assumed to be alive.

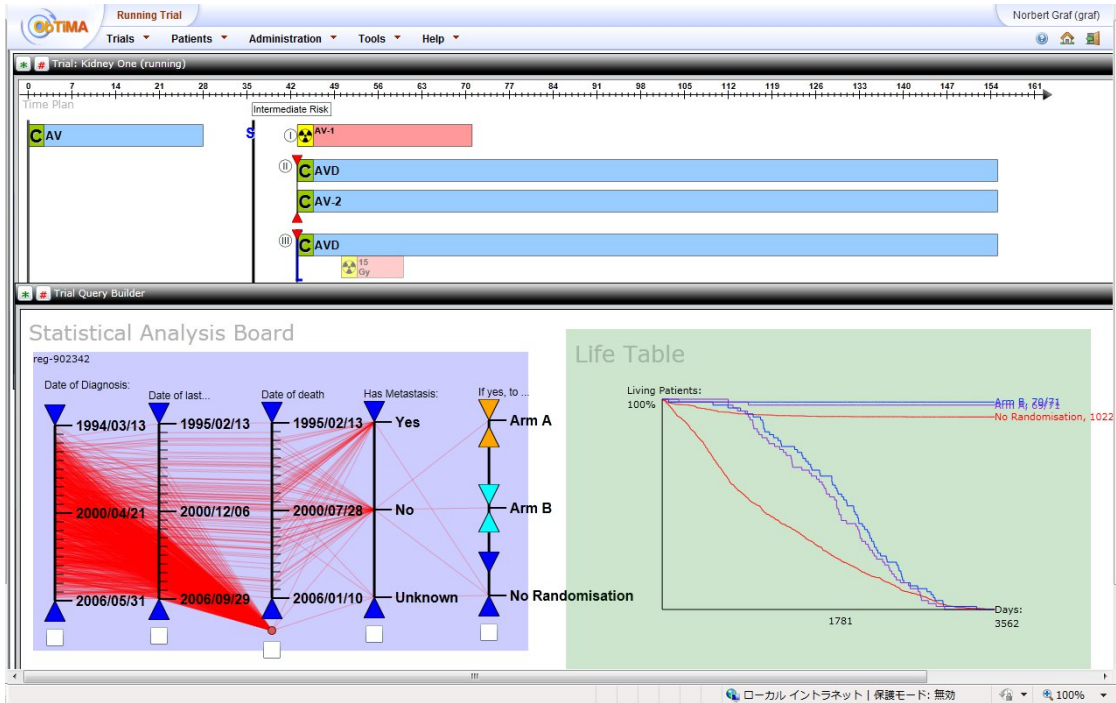


Figure 3. Parallel Coordinates and a Life Table. The top shows the treatment plan of the whole trial. The parallel coordinates have been used to group patients based on what treatment arm they belonged to, and a life table has been used to show the difference in successful treatment between these groups. The upper lines that flatten out towards the right show the percentage of patients in each group that are still alive as a function of the time from entering the trial (i.e. from the time the treatment started or the cancer was diagnosed). Two groups have very little difference, but one group has significantly lower survival rate. This was because all patients with metastatic cancer were placed in that group.

The lower three curves that go steadily down show a pessimistic estimate of the percentage of patients still alive. Every time we have no more data on a patient, the patient is assumed to be dead (there are many reasons for not having data, including the patient leaving the trial, the patient not having reached this stage in the trial (i.e. a patient 7 years into the trial has no data for being dead or alive after 10 years), the patient not being heard back from for whatever reason, etc.). The true percentage of patients still alive is thus somewhere between the upper and lower line for each group. In general, if a patient dies, the trial people will hear of it, and thus the upper (optimistic) estimate tends to be close to the true value.

In the example in Figure 3, the reason one group did so much worse was that all patients with metastatic cancer (thus more likely to die) had to be allocated to this group.

Life tables can also show the number of patients who are relapse free or some other important factor other than just being dead or alive. Our system supports displaying the data in various types of life tables too.

When visualizing data using charts or life tables, groups of patients to visualize are selected using the parallel coordinates and image maps. Changing the filters on the parallel coordinates or regrouping the patients is reflected in the charts and life tables. These are updated in real time. An

example screen shot using all components together is shown in Figure 4.

VI. EVALUATION

Our system has not been formally evaluated yet. We have had about six clinicians play with the system, and we have shown it at the big final review of the EU FP 6 project “ACGT” (the reviewing committee included clinicians). All clinicians who have seen the system so far were very impressed and expressed interest in using the system if it was made available. We would like (and plan) to do proper evaluation of the system, but the preliminary feedback so far at least seems very promising.

The uniform look and feel throughout all modes of the system seems to be popular, and it is similar to the paper based handling that the intended users seem to use currently. Most of the positive comments we have received have been with regards to the data visualization parts, though. The fact that a clinician can explore the data by himself, without contacting statisticians to do the work for him, is very much appreciated. That all the visualizations can be set up first and are then updated in real time when subsets of patients are added or removed so that changes can be seen immediately seems to give the strongest impression. While a statistician could probably explore similar data with his usual tools, the

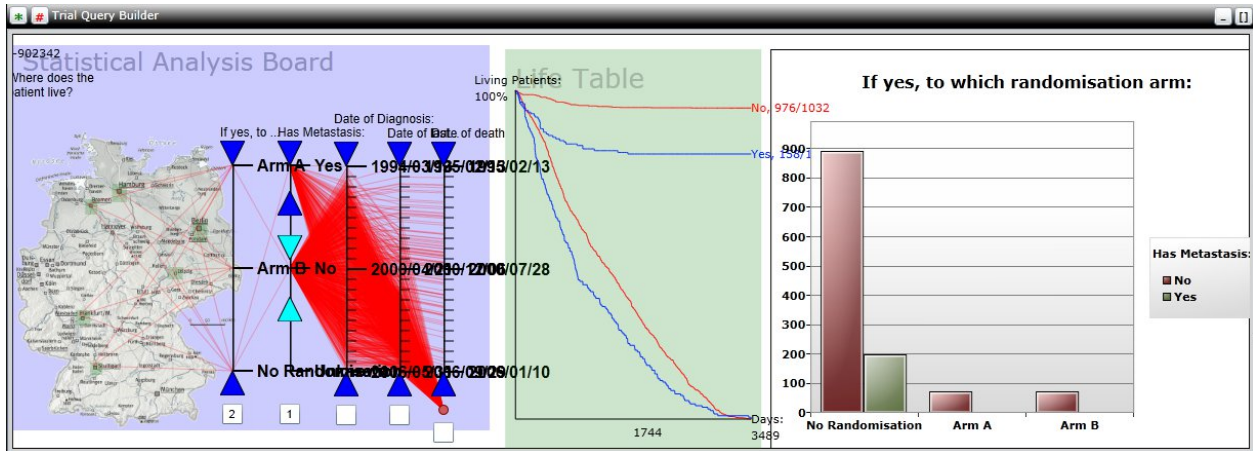


Figure 4. Combining an image map, parallel coordinates, a life table, and a bar chart. The parallel coordinates have been used to group the patients into patients with and without metastatic cancer. Patients where it is “unknown” if they had metastatic cancer or not have been removed. The life table shows that patients with metastatic cancer have lower survivability (as expected), and the bar charts show that such patients were never randomized, while two small groups of patients were randomized to tracks A and B.

clinicians are not statistically trained and they appreciate a tool with a simple interface that allows them to do the things they commonly want to do. Currently they have to ask a statistician to do it for them, even for fairly simple analyses.

VII. RELATED TECHNOLOGIES

There are other systems for clinical trial support. Examples include AI based systems that check constraints on trials and warns the chairman if the knowledge base indicates the current trial has problems, such as using a not recommended medicine for the control group [8]. The TOB does not currently support any automatic validation of the created trials, though it does indicate possible mistakes such as medical events with no assigned input forms to the chairman. Adding validation to the TOB has been planned, but is not done yet.

There are also systems for visualizing the trial flow during design [9], [10], similarly to TOB. We have not seen any system that supports all phases of the work flow, though, only the design phase.

There are also many systems for exploratory data mining of medical data. One example very similar to the analysis part of TOB is described in [7]. It uses several visualization methods, for example 3D parallel coordinates, and allows real time filtering of the data etc. This system is not for use specifically with clinical trials, it has a broader usage. Unlike TOB it does not allow for direct access to the data through the graphical trial plan etc.

VIII. DISCUSSION

TOB was designed in close cooperation with an expert on cancer clinical trials. The idea was to first design a system based on what clinicians would like to have, without

considering how to implement it. After building a system based on these design ideas, more iterations ironing out misunderstandings and presenting ideas that are possible that the clinicians had not even imagined, were also performed. After a number of such iterations, a system that seems to please clinicians very much was created.

The resulting direct manipulation interface for creating a master plan is very visual and intuitive. The same interface is then used for data input and data exploration too.

From an implementation point of view, the use of plug-gable components has been very convenient. When adding new visualization tools etc. it has been easy to just plug in new functionality.

In the future, we would like to add analysis of genomic data too. This will have to be based on yet another component, that allows selection and visualization of the very large data sets even the genomic markers for a single patient produce. Even simple sets of genomic markers that seem to predict various cancer types consist of for instance 95 different markers. Adding a new component will be necessary to deal with this, but plugging in a new component once you know what component you want is straightforward.

The use of parallel coordinates and other tools for direct manipulation of the data to select subsets of patients based on combinations of data features allows for easy exploration of the data. Setting up life tables or charts to show some performance measure and then immediately see the changes when changing the selected set of patients is an intuitive way to look at different aspects of the data. A visual analytics approach like this to visually an exploratively search the very high dimensional space of data is useful since what parameters affect cancer in what ways is very difficult to model formally.

IX. CONCLUSIONS

We described a system for visualizing data collected during clinical trials. The system is built using pluggable software objects, so it is possible to add, combine, or replace visualization components with other types of visualization. Currently the system supports parallel coordinates, clickable image maps, many types of charts, and life tables.

The interface is built on direct manipulation, and aimed at non-experts in statistics or data mining (the intended users are clinicians). It is easy to select data to visualize and to set up a visualization scenario such as showing the survival rate of two groups of patients receiving different treatments. Once it has been set up, selecting subsets of patients, e.g. only patients of a certain age or only patients who had a certain genetic predisposition, shows how these changes reflect the original visualization immediately, i.e. the same charts or life tables set up using different selection criteria are updated in real time to reflect the new selection restrictions. It is thus easy to interactively explore the multi dimensional data to see if for example a clinician's intuition about some parameter being important under certain circumstances is supported by the data.

Clinicians that have tried the system have been very positive, and they especially like that they can explore the data themselves, and that they can see the changes their choices make in real time.

REFERENCES

- [1] J. Sjöbergh, M. Kuwahara, and Y. Tanaka, "Using Web-based Meme media technologies to create an integrated visual environment for clinical trials," in *The IET International Conference on Frontier Computing 2010*, Taichung, Taiwan, 2010, pp. 366–371.
- [2] G. Weiler, M. Brochhausen, N. Graf, A. Hoppe, F. Schera, and S. Kiefer, "Ontology based data management systems for post-genomic clinical trials within an European grid infrastructure for cancer research," in *Proceedings of the 29th Annual International Conference of the IEEE EMBS*, Cité Internationale, Lyon, France, 2007.
- [3] Y. Tanaka, *Meme Media and Meme Market Architecture*. Piscataway, NJ; USA: IEEE Press, 2003.
- [4] M. Kuwahara and Y. Tanaka, "Webble world – a Web-based knowledge federation framework for programmable and customizable Meme Media objects," in *The IET International Conference on Frontier Computing 2010*, Taichung, Taiwan, 2010, pp. 372–377.
- [5] A. Inselberg and B. Dimsdale, "Parallel coordinates: a tool for visualizing multi-dimensional geometry," in *VIS'90: Proceedings of the 1st conference on Visualization '90*, Los Alamitos, CA, USA, 1990, pp. 361–378.
- [6] E. Wegman, "Hyperdimensional data analysis using parallel coordinates," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 664–675, 1990.
- [7] G. Falkman, "Information visualisation in clinical odontology: Multidimensional analysis and interactive data exploration," *Artificial Intelligence in Medicine*, vol. 22, no. 2, pp. 133–158, 2001.
- [8] S. Modgil and P. Hammond, "Decision support tools for clinical trial design," *Artificial Intelligence in Medicine*, vol. 27, no. 2, pp. 181–200, 2003.
- [9] R. D. Shankar, S. B. Martins, M. J. O'Connor, D. B. Parrish, and A. K. Das, "Epoch: an ontological framework to support clinical trials management," in *HIKM'06: Proceedings of the international workshop on Healthcare information and knowledge management*. New York, NY, USA: ACM, 2006, pp. 25–32.
- [10] P. Y. Wong and J. Gibbons, "On specifying and visualising long-running empirical studies," in *ICMT'08: Proceedings of the 1st international conference on Theory and Practice of Model Transformations*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 76–90.