The World Health Organization’s Global Burden of Disease statistics identify cancer as the second largest global cause of death, after cardiovascular disease. Global cancer deaths are projected to increase from 7.1 million in 2002 to 11.5 million in 2030. However, new advances in cancer prevention, diagnostics, and treatment mean that one third of cancers are preventable while another third are curable through early detection and effective therapy. These new cancer therapies are subject to vigorous research and trials, including the application of new high-throughput biomedical technologies that generate large amounts of data accessible in public online registries.

In 2008, Dana-Farber Cancer Institute’s Cancer Vaccine Center (CVC) initiated a research project to investigate the competitive landscape of the cancer vaccine field and to help shape its strategy in the marketplace. This required studying data from 645 cancer vaccine clinical trials and analyzing statistics on cancer types, incidence, and survival rates. At first glance, this appeared to be a very difficult, time consuming task. To analyze information from multiple data sources, understand the relationships underlying the data, and identify trends and patterns would have required significant IT resources using a traditional approach.

However, using a visual analytics tool for data exploration and discovery, an approach was developed to rapidly extract complex cancer vaccines data from major clinical trial repositories. This application enables rapid analysis of information about institutions, clinical approaches, clinical trials dates, predominant cancer types in the trials, clinical opportunities, and pharmaceutical market coverage. Presentation of results is facilitated by visualization tools that summarize the landscape of ongoing and completed cancer vaccine trials. Summaries show the number of clinical vaccine trials per cancer type over time, by phase, by lead sponsors, as well as trial activity relative to cancer type, and survival data. From a single plot, cancers that are neglected in the vaccine field can be identified. The results were published in the journal Immunome Research.

Analysis Workflow
The data mining system consists of a back-end XML database, a front-end visualization interface, and an analysis component. The analysis workflow is shown in Figure 1. First, XML files for relevant cancer vaccine trials were downloaded from the ClinicalTrials.gov Web site and incidence and survival facts were downloaded from the National Cancer Institute (NCI).
A series of questions were defined to address using this system. Fields of interest contain information such as cancer type, phase of the trial, and recruiting status; these fields were extracted from the primary XML files. Additional fields of interest, such as technology platform, adjuvant usage, and therapy type, that provide information in a form suitable for database querying, were added manually and associated to each clinical trial record in Dana-Farber’s back-end database. These data were not available as separate fields in the ClinicalTrials.gov records, but could be derived from the descriptions and mapped.

Data visualization software was used to construct the environment for the Dana-Farber data mining application. The graphical user interface shown in Figure 2 facilitates graphing and tabulation through drag-and-drop actions.

### Cancer vaccine trials data mining questions
The data mining application yielded answers to questions such as “How has the cancer vaccine field evolved in the last ten years?” and “How many cancer vaccine trials have been conducted and how many of them are currently open in the United States?” and provide a historical view of the field. Similarly, answers to questions like “What cancer types are currently researched in clinical trials?” and “What phase are these trials?” offered an up-to-date view of the cancer vaccine space. In addition, this application helped answer more specific questions such as “How many breast cancer vaccine trials have been conducted by Dana-Farber Cancer Institute’s Cancer Vaccine Center and what types of vaccines were used for those trials?”

The versatility of this system enables the analysis of various dimensions of the clinical trials landscape, including clinical trials by timeline (Figure 3), type of cancer, lead institution, trials by disease prevalence, and/or specific vaccine technology visualized through dynamically generated graphs.

### Conclusion
By accessing comprehensive clinical trials information using next-generation software applications—like the Spotfire from TIBCO (Somerville, Mass.)—several mouse clicks provided access to knowledge that would otherwise require hiring of specialists or consultants. By combining public databases of clinical trials, data formatting by XML, and computational analysis and visualization, specific knowledge can be extracted rapidly from a large data set, summarized, and presented to the user. This data mining approach enabled rapid analysis of the hotspots of cancer vaccine activity and revealed hidden patterns, trends, and biases in the data. Summarization and visualization of these data represents a cost-effective means of making informed decisions about future cancer vaccine clinical trials.

---

**References**

