Published in final edited form as:
Cancer Invest. 2008 December ; 26(10): 1060–1067. doi:10.1080/07357900802272729.

Information Systems for Cancer Research

Michael F. Ochs[1] and **John T. Casagrande**[2]

1 *Associate Professor, Division of Oncology Biostatistics, Johns Hopkins Sidney Kimmel Comprehensive Cancer Center, Baltimore, MD, USA*

2 *Associate Professor, Department of Preventive Medicine, USC Kenneth Norris Jr. Comprehensive Cancer Center, Los Angeles, CA, USA*

Abstract

The last decade has seen a massive growth in data for cancer research, with high-throughput technologies joining clinical trials as major drivers of informatics needs. These data provide opportunities for developing new cancer treatments but also major challenges for informatics, and we summarize the systems needed and potential issues arising in addressing these challenges. Integrating these data into the research enterprise will require investments in 1) data capture and management, 2) data analysis, 3) data integration standards, 4) visualization tools, and 5) methods for integration with other enterprise systems.

Keywords

Medical Informatics; Computational Biology; Clinical Trial; Systems Biology; Vocabulary

INTRODUCTION

Comprehensive Cancer Centers focus on discovery of fundamental cancer biology, development and translation of this knowledge into improved therapy, and improved public health from a greater understanding of genetic and behavioral diversity. With new technologies, these research areas are undergoing rapid change and have increasing needs for comprehensive data management and analysis, leading to a large demand for informaticists in cancer centers. In this work, we introduce the technological components necessary for a comprehensive information system that can support modern cancer research, where the focus is on development of personalized treatment through an understanding of the molecular mechanisms underlying the disease in a patient. The systems required to drive this research include

- data capture and management systems, e.g., Clinical Trial and Laboratory Information Management Systems,
- data analysis pipelines, e.g., high-throughput systems capable of handling extremely large data sets,
- data integration structures, e.g., ontologies and interchange standards,
- visualization tools, and

Correspondence to: Michael F. Ochs, Johns Hopkins University, 550 North Broadway, Suite 1103, Baltimore, MD 21205, Email: mfo@jhu.edu OR John T. Casagrande, USC/Norris Comprehensive Cancer Center, 1441 Eastlake Avenue, Los Angeles, CA 90033, Email: casagrande_j@ccnt.usc.edu.

- a methodology for integration with other enterprise systems (e.g., IRB, grant management, clinical systems).

We summarize these systems in Figure 1, and we refer to this figure as we discuss specific components of the informatics infrastructure. First, we briefly discuss some non-informatics issues of importance, including the inevitable mixing of cultures required to create these systems, the biomedical research driving the need for informatics, and the likely costs and return on investment.

Cultures in Cancer Research

While we focus here on technical issues, successful creation and/or deployment of cancer research information systems ultimately rests upon the successful mixing of multiple "cultures" involved in the research endeavor. The first group, basic scientists, must take steps into the unknown, often ignoring conventional wisdom and sometimes therefore mistaking an artifact for reality. Only with such exploration do advances in our understanding of fundamental cancer biology become possible. The second group, clinical researchers, must rely on careful testing of treatments and a reluctance to leap into the unknown. Only with a reliance on evidence and statistical validity can patients be guaranteed the best care. The third group, biostatisticians, provide a check on these two groups, carefully testing the results and differentiating artifacts from facts. The fourth group, information technologists, must carefully design, develop, and test systems that can be used in the real world to provide appropriate information for biostatisticians, researchers, and clinicians to test hypotheses. Only by applying software engineering principles developed over the last several decades can reliable, scalable systems be developed and integrated. These cultures must collaborate together to optimize cancer research. Mutual respect for the talents of each group will be a prerequisite to success in developing useful cancer research information systems. In addition, clinical trial participants, patients, and patient advocates interact with and support the research enterprise as both the ultimate beneficiaries of the work and the source of information driving many aspects of the discovery process.

Bringing these cultures into a cohesive group that is greater than the sum of the parts is beyond the scope of this paper, however it is an important aspect to consider in establishing a successful informatics effort at a cancer center. Each center is likely to address the issue in a way that meshes with its present culture. For instance, matrix centers will need to work within existing departmental structures, perhaps establishing cross-departmental working groups and a reward structure that encourages efforts outside the department of the researcher. Independent cancer centers may wish to establish groups created specifically to bring together these diverse groups, which simplifies the establishment of promotion tracks. An example of a system built with the involvement of these groups is given in the CAFÉ sidebar, which describes a system, developed by one of us (jtc), at the USC/Norris Cancer Center.

Emerging Technologies for Cancer Research

Cancer biology research has advanced considerably over the past several decades, providing us with an understanding of the impact of cell cycle regulation, apoptotic programs, cell-cell and cell-matrix interactions, and signaling processes on cancer etiology [1]. Historically, the methodologies employed in this research typically relied on molecular biology and biochemical assays, including measurements of small numbers of genes and proteins in limited cell or tissue types, with tools such as gels and radioactive labels providing limited quantification and resolution. With the development of microarrays in the mid-1990's [2,3], the handling of the data and the associated statistical issues involved in analyzing large, noisy data sets began to play important roles on the basic science side of cancer research. These issues continue to grow with the emergence of high-throughput sequencing and microscopy, SNP and exon arrays,

metabolomics, and even more novel technical developments, as well as increased use of high-throughput proteomics techniques.

These new methodologies are also playing an increasing role in clinical trials or population-based research studies, where genomics and proteomics measurements are becoming routine. Thus, in addition to capturing this data, there is a growing need for cancer centers to develop informatics strategies to provide appropriate cross-linkages, so that clinical outcomes and risk factors can be integrated with the results from these emerging "omics" methodologies. Recognizing these needs, the National Cancer Institute (NCI) initiated the Cancer Biomedical Informatics Grid (caBIG™) project, focusing on providing uniform tools to cancer centers, especially for clinical trials. With the caBIG™ desire to share or pool information among centers, there may also be a need to explicitly identify individuals who visit more than one center in the course of their treatment to avoid potential sources of bias.

Return on Investment in Cancer Informatics

There are significant costs in developing and deploying informatics for cancer research, and the return on investment (ROI) must be considered. While the costs of deploying informatics infrastructure can be estimated, the costs in misguided research or lost opportunities of improperly managing data are generally ignored. In one high profile example, it required substantial effort from researchers in the Department of Bioinformatics and Computational Biology at the MD Anderson Cancer Center to halt the marketing of an ovarian cancer screening test that was based on improperly designed experimental protocols and analysis [4]. The same group found errors due to improper data management in a high profile human cancer study [5] that was generating considerable drive for significant investment. The research community tends to follow high profile publications by expending effort and funds in extension of the work without waiting for independent confirmation. Therefore, poor data management can lead to substantial hidden costs.

The costs of deploying adequate infrastructure can be estimated by techniques used in deployment of information technology (IT) in other enterprises. An excellent recent review of the deployment of data management tools for high content screening, a research area with many similarities to other high-throughput biological disciplines, noted that a three year effort required 17 IT and informatics staff at a yearly cost of $5 – 8 million, with an additional $2 – 5 million per year for capital [6]. Although this supported 250 – 300 scientists, it is likely that this is a good ballpark estimate as much of the required infrastructure costs do not scale with the number of researchers. There would be some minor adjustment for direct support (help desk) and data storage, but such costs are relatively minor.

Substantial gains in efficiency will be obtained with the successful deployment of informatics infrastructure. For clinical trials, a Clinical Trial Management System (CTMS) is already a mandatory component of a cancer center's IT infrastructure. Ideally, a CTMS should include tools to aid in trial design, process review, and administrative procedures including billing, eligibility requirements, informed consent and HIPAA authorization, adverse event monitoring and reporting, trial recertification, and long-term participant follow-up. One difficulty in providing an adequate CTMS is that there is currently no system that provides all the needed functionality. The costs of the ideal CTMS are likely to be comparable to high-throughput data systems given the large regulatory and integrative needs.

DATA CAPTURE AND MANAGEMENT

In order to capture the data at the point of generation and thus minimize the potential for error and increase the value to the enterprise, information systems must be integrated into the natural research and trials workflow. As noted in a recent review on informatics in clinical cancer care,

"if the submission of data for research and monitoring purposes requires an extra step, … the process will likely fail" [7]. This is true for biological research data as well, since the rewards to researchers rarely increase linearly with good electronic data management or data sharing practices. For example, although there is substantial microarray data in the public domain, overall compliance with data submission, especially meta-data about protocols, remains limited. A successful system must therefore mesh seamlessly into the researcher's workflow and provide an advantage over simple documents and spreadsheets (such as automated submission of required data or generation of supplemental material or figures), as well as integrate into the larger research enterprise.

Clinical Trials Management Systems

For optimal efficiency clinical research data should be captured at the point of patient care as a byproduct of the normal clinical processes. However, this is seldom achieved due to inadequacies in the manual and electronic systems utilized for patient care and the lack of integration of research staff into the normal care process. This results in additional cost for the deployment of the systems (purchased or developed) needed to track research-specific information and the personnel needed to abstract research-specific data elements from the electronic or paper care records. In most cancer centers, this leads to two separate IT departments, one focused on supporting the business processes of the organization and a second focused on the research enterprise. There is a tremendous opportunity for cost savings and improved return on investment, as well as improved data reliability, if these separate functions are integrated within the center's IT infrastructure, but there are usually organizational, cultural, and social/political barriers that make this integration difficult.

From a purely research perspective, the CTMS should include multiple tools as summarized in Figure 2. An ideal CTMS provides tools allowing creation and/or capture of trial design, informed consent and HIPAA authorization, and data elements for identifying patients' eligibility for trials. For reporting it provides trial submission for local, regional, or national review and approval committees, adverse event reporting, and documentation for trial recertification. For tracking, tools for long-term follow-up of participants and trial status (open, accruing, closed) are needed. In addition, integrating with enterprise IT would permit the CTMS to capture and relay clinical and administrative (trial-specific billing issues, budgetary items, financial issues) to the enterprise billing systems. This includes integration with the Electronic Medical Record (EMR) as shown in Figure 1 in the top box.

A further complication for a CTMS arises from the portfolio of clinical trials having a variety of sponsors. Many of these sponsors provide proprietary tools, including paper or PDF case report forms (CRFs), laptops, and web-based or standalone applications, for data collection, so there is little incentive to capture a complete repository of trial data. However, since there is a need to quantify all trial activities for reporting purposes, there is a universal need to track trial accruals irrespective of the trial sponsor. For in-house trials, there is a need for more extensive data capture in the CTMS, since these trials will be analyzed by center staff.

Laboratory Information Management Systems

For high-throughput data (microarrays, SNP chips, proteomics, etc.), data capture is best accomplished with a Laboratory Information Management System (LIMS), as shown in the top box in Figure 1. These systems integrate with laboratory instrumentation and computers running instruments to capture data with minimal manual intervention, although they still need to be integrated into the natural laboratory workflow. In addition to the raw data, it is critical to capture meta-data defining phenotypes and experimental protocols that summarize how this data was generated, such as specific model organisms, reagents, and outcomes. It is likely that multiple LIMS will need to be deployed to capture the many different data types, although

where the culture permits, a single system offers advantages for data integration and management of cancer center facilities [8]. An example of LIMS that integrates into workflow, developed by one of us (mfo), is presented in the sidebar on the flowLIMS.

The need for LIMS are often overlooked given the high cost and apparent success of core facilities managing data using spreadsheets to store and transfer data. However, the return from the heavy investment in microarray and proteomic technology has not been impressive, and this reflects the difficulty researchers face in comparing experiments due to poor quality protocol information, repeated errors in data handling (as in the examples above), and, of course, unknown losses that never get reported. It is worth noting that the present system in use at many cancer centers to handle the data is completely unacceptable from a drug development viewpoint, and such systems would not lead to FDA approval of a therapeutic, *because of the potential mishandling of data and lack of traceability*.

CTMS and LIMS Integration

A logical next step for CTMS and LIMS is their integration, permitting capture of genomics data in the LIMS but with successful linking to CTMS patient-based information. Since a LIMS handles data both covered and not covered by HIPAA/IRB, a convenient method is to de-identify the data in the LIMS. This simplifies LIMS creation, as no protected health information is stored in the system. This can be automated by de-identification systems [9,10].

DATA INTEGRATION AND ANALYSIS

As an example of the need for data integration, imagine a clinical trial focused on the use of a targeted therapy, such as an antibody to a membrane receptor tyrosine kinase. Data would be collected on patient response, adverse events, long-term survival, as well as molecular information such as proteomic profiles, genotype from SNP chips, and microarray responses of the tumor during treatment. Alternatively, work in the clinic might demonstrate that only cancers arising from certain precursor cell types are responsive to the therapeutic, suggesting a shared modified pathway that the therapy targets and, potentially, specific other cancers that could respond to the same treatment. In such studies, linking patient data would permit classification of response, including adverse events. Linking chemical and structural data would identify similar compounds to the therapeutic, permitting leveraging of data from other trials and model organism studies. Linking across proteins and genes would tie responses measured on microarrays or through proteomics to pathways, potentially including rescue pathways activated in the tumor in response to treatment, which could provide potential additional therapeutic targets. Such data integration would also ease development of animal models used for testing potential adjuvant therapies that may be individually tailored based on each patient's response. While all these actions are possible without encoding the data, the manual effort required could easily lead to a failure to undertake a study due to cost and even due to an inability to gather information in time.

Integrating Data: Semantic and Syntactic Links

Ontologies and their use for encoding the data are essential, as they permit automated semantic integration of data. Ontologies, like controlled vocabularies, provide specific terms for describing each data element; however they also set up a hierarchy allowing data integration across different resolutions (e.g., pulmonary system, lung, alveoli). In medical informatics, there are multiple existing ontologies, collected and integrated within the Unified Medical Language System (UMLS) [11], and they have been utilized in some clinical systems for data integration [12]. For high-throughput biological data, ontologies are only beginning to be developed and utilized [13], however their implementation in systems during the first steps of development will significantly enhance the value of the data. In addition to enabling data

integration, ontologies enable semantic interoperability, permitting equipment and systems from different vendors to be fully integrated into the research enterprise. Thus, researchers are able to choose the best equipment for their research without incurring the loss of the ability for this data to be utilized in a larger study in the future.

While ontologies provide for semantic mapping between data sources, it is also necessary to provide interfaces that permit syntactic interoperability, i.e. a basic grammar for systems to communicate. The primary example within medical informatics is the HL7 messaging system [14], which permits clinical systems, billing and accounting systems, and third-party payer systems to interoperate. Each system retains its own internal structures and operations, but is able to gather data from and provide information to all other HL7 compliant systems. In the research community, the use of the eXtensible Markup Language (XML) is becoming standard. XML provides a syntactic framework permitting the creation and parsing of documents containing data elements and meta-data based on tags encoded within the text. These tags, if derived from appropriate ontologies, can also provide details for semantic integration. Use of caBIG™ metadata repositories such as EVS/caDSR to define common data elements and object models will enhance in-house and cross-center interoperability [15]. In addition, basing CTMS design on emerging models such as BRIDG, which rely heavily on XML and XMI (a metadata exchange model utilizing XML), will also simplify integration.

Since it is highly unlikely that a single vendor can provide suitable systems for all aspects summarized in Figure 1, interoperability is essential. While syntactic and semantic structures permit such interoperability, there are additional advantages to open-source systems, since these permit modification at the code level and often advance quickly to solve new problems, as a community of developers can emerge around the needs of the community of researchers. The best example of successful open-source development in biomedicine is the widely used R/Bioconductor system [16], for which tools are routinely developed and shared simultaneously with emerging technologies that generate new data types, often in large volumes.

Analyzing Data

Once data has been successfully integrated, the goal will be to apply analysis and data mining methods to discover knowledge, which will require development of both computational systems and analysis methodologies. As an integrated data set will comprise terabytes of data, the system will need to both manage large data footprints and provide significant computational power. It is unlikely to be economically feasible to provide adequate desktop systems to researchers, as both the processing power and memory requirements will be substantial. As has already occurred with recent Affymetrix GeneChips™ and the desktop R/Bioconductor statistical software [16], typical desktop computers will not have sufficient RAM memory for standard processing. The situation will grow rapidly worse with integrated data involving multiple data types. This suggests that new enterprise systems running on high powered servers and computer clusters will be required, as depicted in the middle box in Figure 1. In addition, these systems will require high-speed connections to the data resources to handle the large data transfer requirements. Such data would be stored in LIMS, CTMS, clinical EMR systems, and potentially data warehouse systems.

As knowledge discovery from integrated data is an emerging area, new algorithms will appear often and will need to be incorporated into the analysis system, making a flexible system that allows easy extension essential. Preferably extension should occur on the live system in order to minimize downtime and insure that long data mining operations are not interrupted. Such operations will typically seek the molecular bases of specific cancer development and successful treatment, which will likely require knowledge contained in national and international repositories, such as NCBI, EBI, PIR, PDB, CGAP, TCGA, etc. The inclusion of ontologies for encoding data will be essential to leveraging these resources.

While there are multiple potential technologies that can form the basis of these systems, the success of service oriented architectures for flexibility, web services and application servers for scalability, and computer clusters for computational throughput suggest a model using application servers that permit live extensions coupled to Beowulf clusters (as in Figure 1). Another potential approach is the use of grid computing technology to enable data and compute cycle sharing, however there are multiple issues to resolve, including adequate throughput for terabyte-size data sets and data security for sensitive data.

VISUALIZATION

With the complexity of cancer etiology and treatment, it is unlikely a statistical tool that provides details in a single generated table will provide the greatest insight. Therefore there is a need for visualization tools capable of providing insight into patterns in high-dimensional data. At its most trivial, the problem is similar to finding something significant in a spreadsheet summary of a microarray experiment where relative expression levels for 40,000 probesets across tens of conditions appear. Early visualization in this field utilized the now (in)famous red-green heatmaps of clustered genes, while more recently the ability to incorporate additional information has led to pathway- and ontology-centric analyses [17,18]. As dimensionality increases and multiple data types (e.g., genotype, expression, protein levels, protein states, phenotypes, etc.) all need to be visualized simultaneously, novel methods will be required.

For visualization to be meaningful, it is essential that sound statistical methods underlie the analysis, highlighting the importance of involving the biostatistics groups present in cancer centers in planning. Where necessary, biostatistics groups should be augmented with specialists in genomic data analysis. These individuals will complement those working on clinical trials and associated methodologies, so that the biostatistics group will emerge as a multidisciplinary team who together have the necessary background to handle the analysis of integrated data sets emerging in translational research. These groups will need to work closely with bioinformaticists and clinical trials informaticists to integrate statistical techniques into the analysis systems. While there is a tendency to segregate these groups by titles, in reality both the work and the skills form a continuum.

At this time, visualization is not a high profile field in cancer informatics. However, it will be essential for aiding researchers in understanding the results of analyses and in forming a feeling for the organism, which has been so essential to our growth in understanding to this point. In general, researchers, especially within basic science, wish to interact with their data, explore it in ways that cannot be foreseen, so that visualization tools are needed to enable this form of discovery science.

INTEGRATION WITH SUPPORT SYSTEMS

In addition to research data systems, cancer centers need a variety of support systems to fulfill their mission, such as facility billing and order tracking, grant management and publication tracking, and web systems for institutional development or marketing. Most cancer centers have several “cores” or “shared facilities” supporting research activities, where web-based order entry, result delivery, and billing/activity databases have been used to effectively replace historic “log” books. As a center’s research productivity is measured in a variety of ways, including the publication records of researchers, a web-service based retrieval system of member’s publications from PUBMED that categorizes publications by the members’ center program affiliation can be a major improvement in the preparation of reports and funding submissions (see [19] for an example). It can also be used to showcase research activity on the center’s public website by integration with institutional development and communications systems.

A number of enterprise systems could support research and gain efficiency by being linked to research systems. Authentication and authorization systems, including key and ID badge distribution, can be used to provide a single password and token for systems. The most widely used is the Lightweight Directory Access Protocol (LDAP) that provides an open-source infrastructure to allow a single password across multiple systems. With the need to work across cancer centers, authentication and authorization between cancer centers becomes an issue, and work within the caBIG project on a system to allow this is ongoing. Systems for tracking work order requests, for bug tracking in systems, and for processing orders and tracking delivery can also be integrated to enhance a center's efficiency.

In addition, many systems would gain from gathering data from research systems. For clinical trials, linking the CTMS to billing systems and report systems can automate the recovery of funds from trial sponsors and generation of progress reports and adverse event reports respectively. Linking both CTMS and LIMS with grant management can aid in providing details for grant submissions and reports, including use of shared resources for NCI CCSG funded institutions. A major gain in efficiency involves linking of CTMS and LIMS to IRB and data monitoring systems to provide automated information necessary for proper trial oversight. In the future, linking CTMS and LIMS to emerging clinical systems such as EMR, pharmacy, and clinical laboratory systems will aid in data integration for translational research.

SUMMARY

Figure 1 provides an overview of the systems and interactions needed to handle clinical trials and high-throughput data in research. A distinct advantage in creating these systems is the ability to leverage several generations of work in medical informatics, so that research informatics can begin with "third generation" systems that use distributed processes, structured data, and XML [20]. These systems will ideally interact gracefully with institutional systems for administration and clinical care and will utilize institutional IT infrastructure and expertise. While the cost and effort required to deploy such systems should not be underestimated, the potential return on investment is substantial, both in real dollars and in the improvement in the likelihood that personalized medicine will become a reality.

Since there are many required systems and interactions, a well formulated plan for purchase, creation, and deployment will be essential. Ideally this would be done using an integrated information architecture for the organization. However, systems are often purchased at a departmental level without regard for the larger enterprise. Successful deployment of integrated informatics will require a structured approach, as well as a focus on interoperability. An overarching approach must include 1) inclusion of interoperability requirements, both syntactic and semantic, in system specification, 2) a long-term plan for deployment of systems, since all systems are unlikely to be deployed simultaneously, and 3) an understanding of system dependencies, so that systems are deployed in an appropriate order. Obviously, extensive planning and organizational commitment are therefore necessary for success.

Within the present NIH budget environment, it is unrealistic to expect that the major portion of the costs for creating and deploying systems will come from government resources. Even with the large expenditures within the NCI's caBIG™ initiative, the funds available to a single institution are at best a small percentage of the total cost of deployment. Other large initiatives, such as CTSA grants, have diverse goals and the majority of funding is unlikely to be dedicated to cancer informatics needs. As such, institutions will need to find ways to cover costs, either through philanthropy or as part of ongoing operations. While some institutions will fail to move in this direction, the widespread use of SNP technologies, microarrays, proteomics, and high-throughput screening suggest a future for cancer research with an essential large informatics

component. Centers that do not develop informatics expertise are likely to find many future opportunities closed to them.

SIDEBAR 1

CAFÉ: An Example of Unified Enterprise Research Data Management

In mid 2000, as the result of a comprehensive review of informatics support at the USC/Kenneth Norris Jr. Comprehensive Cancer Center (KNJCCC), it was determined that a more centralized approach to research informatics was needed to both accommodate existing needs and those of future high-throughput technologies. It was determined that a variety of tools (FileMaker Pro, Access, Excel, Oracle, SQL Server, etc) were currently being used in a project-centric manner for research, with no common approach to creating the front-end applications. This resulted in a re-evaluation of how research informatics support was provided and led to the creation of a standardized application development framework that has been used over the past five years to capture research data at KNJCCC. The Common Application Framework (Extensible), CAFÉ, has the design goal of integrating all the research data management applications into a single unified user workspace, while allowing easy extension at any time. Pre-existing research applications have been replaced and integrated by redeploying existing forms and other user interface components as .NET Windows Forms. Existing reports have been integrated by moving them to a .NET version of Crystal Reports and, more recently, to SQL Reporting Services. Furthermore, existing web applications are incorporated easily into a CAFÉ application, since a web browser control is included.

CAFÉ applications rely upon a database containing all user and application specific research data and metadata. CAFÉ provides a dynamically configurable user interface via a menu tree built dynamically from the database and handles user access and role-based security. For data storage/retrieval, CAFÉ utilizes a common data access layer that simplifies development of audit trails and the binding of form controls to database fields. A key support feature utilizes built-in .NET functionality that provides the deployment of a fully featured Windows application from a web server, so that a user can download a Windows executable simply by accessing a URL. When updates are needed, the existing components on the web server are updated, and these are transparently moved to the client desktops by .NET, so that users always run the current version of the application without the need for intervention from IT staff.

Although initially designed for CTMS, CAFÉ is very flexible and has been extended to support a variety of cancer research data management needs. While it uses a central framework to do common underlying operations like reading and writing to the database, managing role based security, audit tracking, etc., the front-end is completely configurable. This makes it possible to build a variety of research applications with CAFÉ and will support rapid deployment of ontologies linked to data in the future. In addition to supporting clinical trials at KNJCCC, CAFÉ is used to maintain a patient registry for an affiliate hospital including trials for a pediatric neuroblastoma consortium, to capture the clinical/surgical experience for several urologic cancers, to capture research data for prostate core biopsies, to automate two prevention trials, and to support several population-based etiologic investigations. Most recently, in support of these etiologic investigations, CAFÉ has been used to capture tissue microarray (TMA) information. A key factor leading to the success of the CAFÉ framework in transforming research informatics has been the close collaboration and interaction between the center's biostatistics, scientific, and informatics personnel when implementing research solutions using CAFÉ. This has resulted in re-use of components and the recognition that a centralized integrated approach to managing research data can be a successful strategy.

SIDEBAR 2

flowLIMS: An Example of Integration of a LIMS in the Basic Science Workflow

An example of a LIMS integrating into natural workflow is the Flow Cytometry LIMS (flowLIMS) created at the Fox Chase Cancer Center and deployed there and at the University of British Columbia. Using virtual pipetting through a web interface, a researcher creates a model of the experiment. The protocols are recorded using controlled vocabularies for cell types and stains (antibody-fluorochrome combinations); the experimental protocol is automatically transferred to the cytometer; and the raw data is automatically captured back into the system. The flowLIMS both provides secure storage and reduces the time spent on the expensive cytometer, since protocols can be defined from the user's desktop computer.

The flowLIMS also highlights the cost savings possible in leveraging available institutional infrastructure. The system as deployed at Fox Chase relies on an enterprise tiered storage architecture that handles data aging and backup automatically. Such systems are expensive, but when present for the enterprise, provide highly reliable storage for the relatively minor cost of extending the system. When such systems are not available, implementation planning must include methods to handle storage and backup of terabytes of data, transaction processing, data retrieval, and a strategy for handling increasing amounts of data as other technologies come into use (e.g., high-throughput sequencing, SNP arrays). The flowLIMS also provides an object lesson in the importance of appropriate IT staffing for enterprise systems. Two institutions failed to successfully deploy the system, and in both cases the deployment was attempted by laboratory "IT" personnel and not by dedicated IT professionals trained for enterprise-scale systems. Enterprise systems are more complex and require greater effort and training for successful construction and deployment, however only enterprise-scale systems can handle the large data flows and analysis required.

References

1. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100(1):57. [PubMed: 10647931]
2. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996;14(13):1675. [PubMed: 9634850]
3. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270(5235):467. [PubMed: 7569999]
4. Baggerly KA, Morris JS, Edmonson SR, Coombes KR. Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. Journal of the National Cancer Institute 2005;97(4):307. [PubMed: 15713966]
5. Coombes KR, Wang J, Baggerly KA. Microarrays: retracing steps. Nat Med 2007;13(11):1276. [PubMed: 17987014]
6. Garfinkel LS. Large-scale data management for high content screening. Methods in molecular biology, Clifton, NJ 2007;356:281.
7. Shortliffe EH, Sondik EJ. The public health informatics infrastructure: anticipating its role in cancer. Cancer Causes Control 2006;17(7):861. [PubMed: 16841254]
8. Naeve, C. St Jude Children's Research Hospital. personal communication
9. Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. American journal of clinical pathology 2004;121(2):176. [PubMed: 14983930]
10. Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC medical informatics and decision making [electronic resource] 2006;6:12.
11. Humphreys BL, Lindberg DA. The UMLS project: making the conceptual connection between users and the information they need. Bull Med Libr Assoc 1993;81(2):170. [PubMed: 8472002]

12. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Inform Assoc 1994;1(1):35. [PubMed: 7719786]

13. Stoeckert C, Ball C, Brazma A, Brinkman R, Causton H, Fan L, Fostel J, Fragoso G, Heiskanen M, Holstege F, Morrison N, Parkinson H, Quackenbush J, Rocca-Serra P, Sansone SA, Sarkans U, Sherlock G, Stevens R, Taylor C, Taylor R, Whetzel P, White J. Wrestling with SUMO and bio-ontologies. Nat Biotechnol 2006;24(1):21. [PubMed: 16404382]

14. Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, Shabo Shvo A. HL7 Clinical Document Architecture, Release 2. J Am Med Inform Assoc 2006;13(1):30. [PubMed: 16221939]

15. The Cancer Biomedical Informatics Grid (caBIG): infrastructure and applications for a worldwide research community. Medinfo 2007;12(Pt 1):330.

16. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5(10):R80. [PubMed: 15461798]

17. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13(11):2498. [PubMed: 14597658]

18. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 2005;33(Web Server issue):W741. [PubMed: 15980575]

19. http://uscnorriscancer.usc.edu/Publicationviewer/pubsstart.aspx

20. Stead WW, Miller RA, Musen MA, Hersh WR. Integration and beyond: linking information from disparate sources and into workflow. J Am Med Inform Assoc 2000;7(2):135. [PubMed: 10730596]

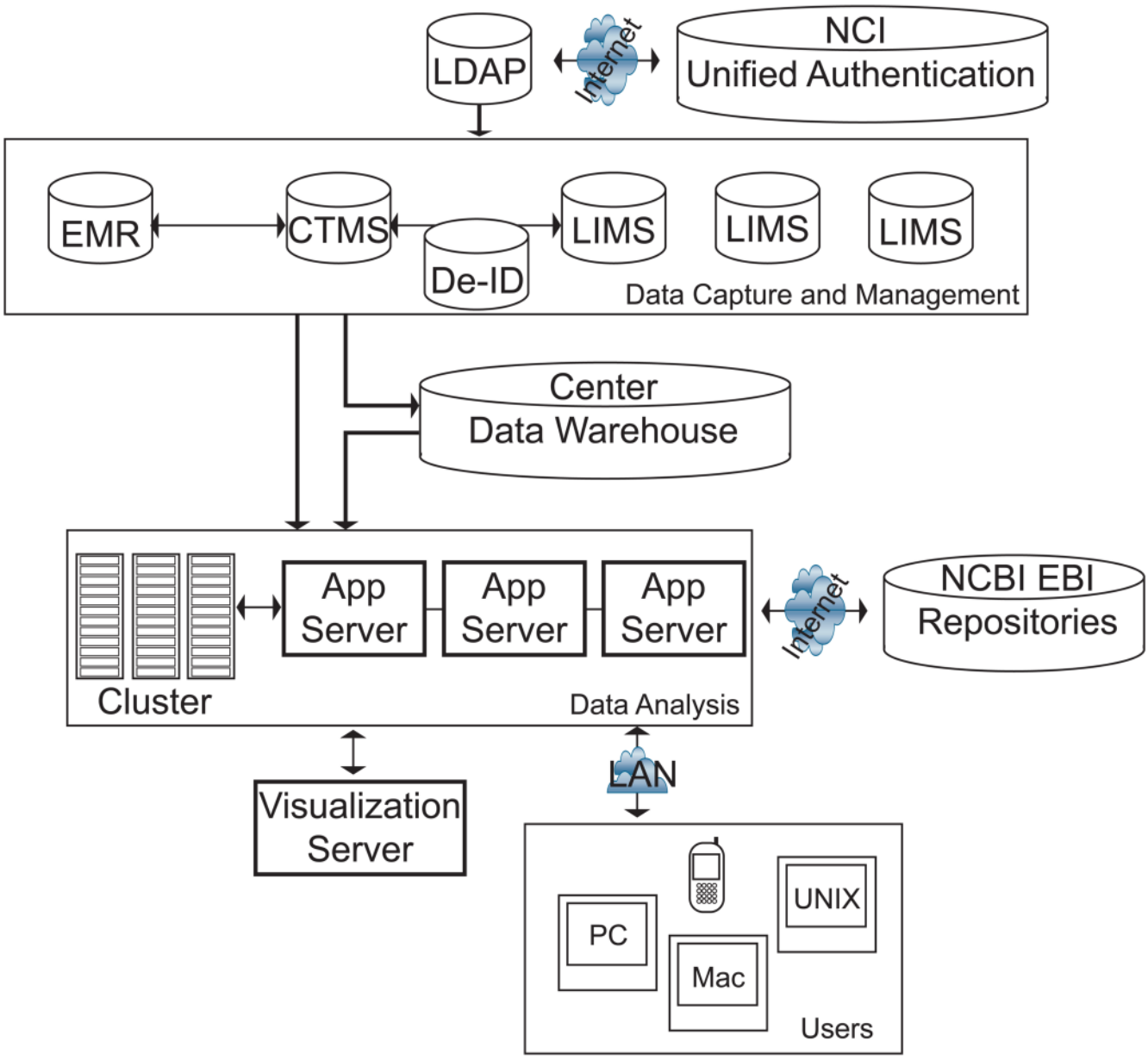

Figure 1.

The systems needed for cancer research informatics. This schematic shows the systems necessary for handling data management and analysis. The LDAP or other authentication/ authorization server handles centralized login and is expected to communicate with NCI caBIGTM servers in the future to provide cross-center access. CTMS and LIMS systems capture research data, while an EMR provides data from clinical systems including pharmacy and laboratory. To reduce system requirements, data de-identification is done to minimize HIPAA issues, since LIMS store model organism data and cell line data as well as human subject data. These systems can all feed a data warehouse if desired. Data analysis is performed by linked application servers providing scalable tools accessible by users through web interfaces, potentially including PDAs and cellphones, and a cluster is provided for high-throughput data analysis. These systems would also be linked to enter prise systems such as grant management, HIPAA/IRB, and billing, which are not shown.

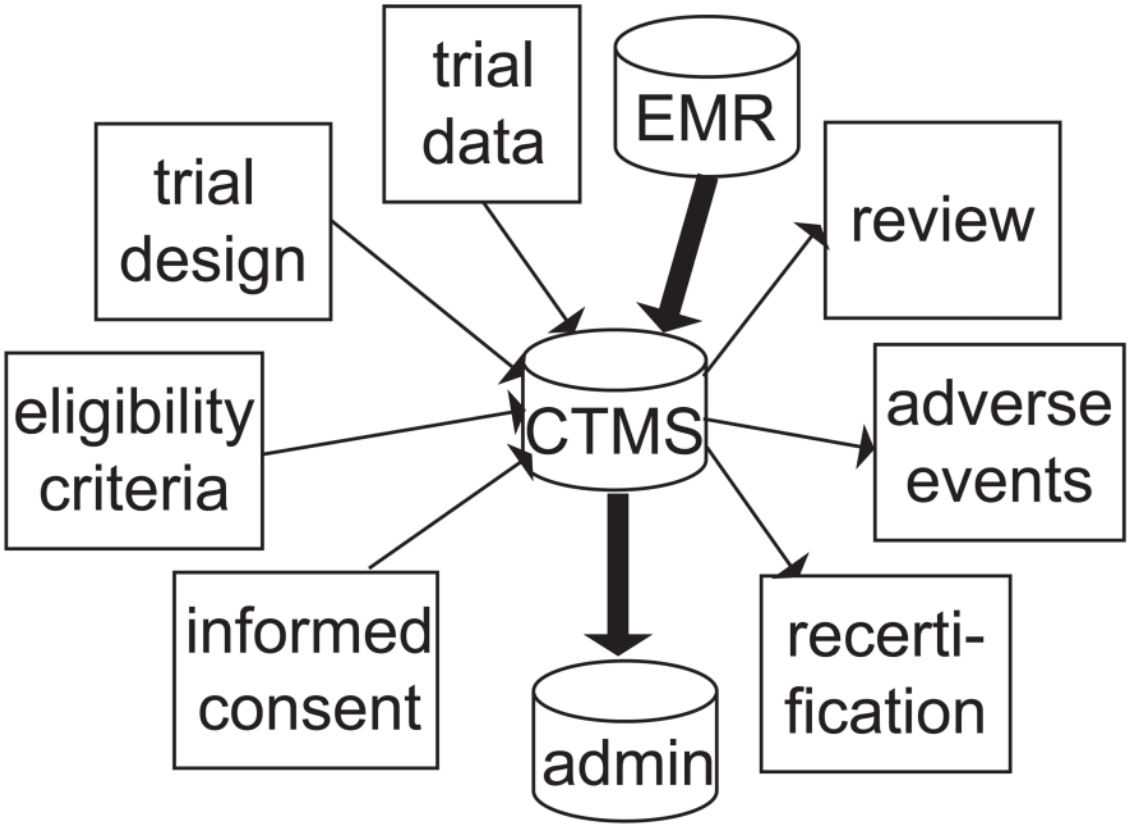

Figure 2.
The ideal CTMS system. The CTMS will communicate with other systems, such as an EMR for capturing patient data, as in Figure 1, and administrative systems for billing and reporting. It will provide tools for clinical trial design, for eligibility and informed consent tracking, and for producing reports for trial reviews, adverse events, and recerti cation. It must also capture data on clinical trials from vendor systems and paper reports. In addition, the system would link to IRB and data monitoring systems where available.

Table 1

Selected terms and abbreviations

caBIG	Cancer Biomedical Informatics Grid, NCI sponsored informatics effort
CTMS	Clinical Trial Management System, a system for capturing clinical trial information
LIMS	Laboratory Information Management System, a system for capturing research laboratory data
EMR	Electronic Medical Record, an electronic version of the medical record
IT	Information Technology
metadata	"data about data", information for placing data in context
SNP chip	"snip-chip", a microarray that measures specific single nucleotide polymorphisms
TMA	tissue microarray array, a slide with spots from multiple tumors
NCBI	National Center for Biotechnology Information, home to most US genomics databases
EBI	European Bioinformatics Institute, home to most European genomics databases
PIR	Protein Information Resource, annotated protein sequence database
PDB	Protein Data Bank, repository for protein structures
CGAP	Cancer Genome Anatomy Project, expression profiles in cancer
CTSA	Clinical and Translational Science Award, large NIH grant to individual institutions to support translational research
TCGA	The Cancer Genome Atlas, repository for data from tumor profiling
HIPAA	Health Insurance Portability and Accountability Act, rules for protection of health information
IRB	Institutional Review Board, institutional committee overseeing human subjects research
UMLS	Unified Medical Language System, a system semantically integrating multiple medical vocabularies
HL7	Health Level 7, a standard for syntactic communication between medical systems
EVS	Enterprise Vocabulary Server, a server of all caBIG data elements
caDSR	Cancer Data Standards Repository, a wrapper on EVS providing access to data elements
BRIDG	Biomedical Research Integrated Domain Group, a project to link standards in clinical trials research
LAN	Local Area Network, the institutional computer network
PDF	Portable Document Format, a standard for exchanging human readable documents
XML	Extensible Markup Language, a standard for exchanging computer readable documents
XMI	XML Metadata Interchange, a standard for defining metadata in XML