Early Detection Research Network

Statistical and Informatics Infrastructure: EDRN DMCC and JPL

Ziding Feng, DMCC
Mark Thornquist, DMCC
Dan Crichton, JPL

Early Detection Research Network

Statistical Tools and Research Management

Ziding Feng, Ph.D.
Mark Thornquist, Ph.D.
Margaret Pepe, Ph.D.
EDRN Data Management and Coordinating Center

Early Detection Research Network

Areas of Emphasis

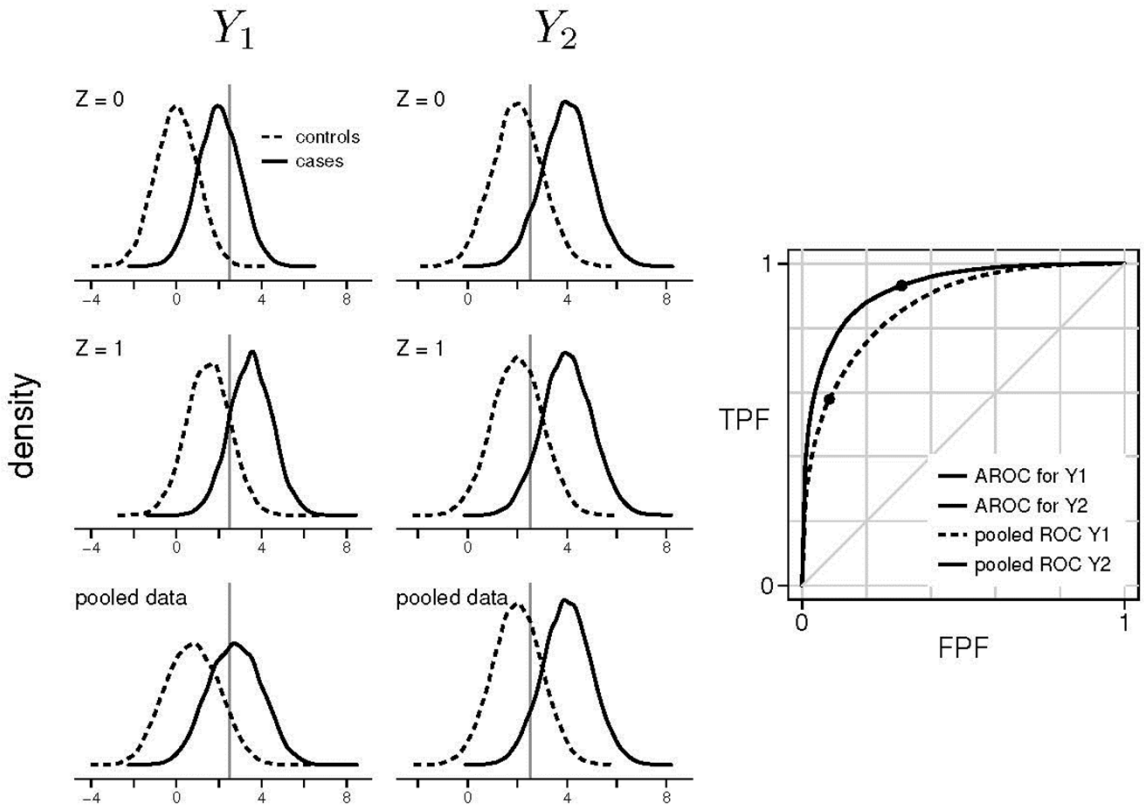
1. Design of validation studies
2. Analysis of validation studies
3. Censored event time outcomes
4. Issues in risk prediction
5. Quantitative proteomics methods
6. Biomarker discovery

Dissemination

Example: Analysis of Validation Studies

Early Detection Research Network

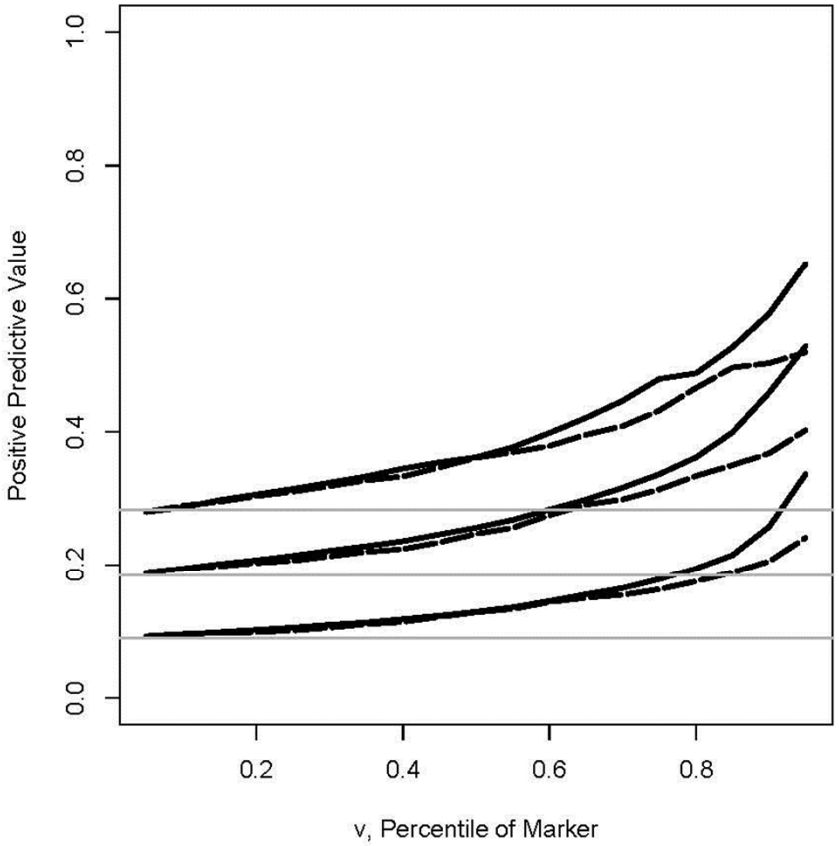
Y_1 and Y_2 have the same performance, Y_1 varies with Z, but Y_2 does not. The unadjusted ROC curve for Y_1 is attenuated while that for Y_2 is not.

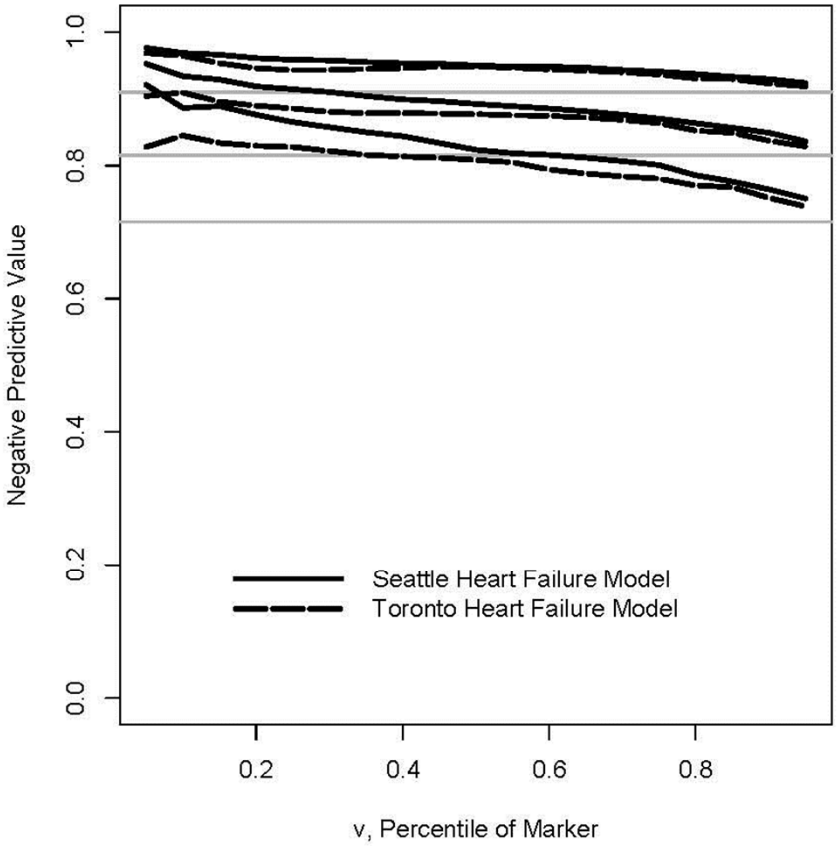


Example: Analysis of Markers' Prognostic Value

Early Detection Research Network

PPV (left panel) and NPV (right panel) curves at $t = 1$, $t = 2$ and $t = 3$ years (from top to bottom) after enrollment. Solid lines, estimates from the Seattle heart failure model; dashed lines, estimates from the Toronto heart failure model; Gray horizontal lines are for $P(T < 1year)$, $P(T < 2years)$ and $P(T < 3years)$ in the PPV plot and $P(T \geq 1year)$, $P(T \geq 2years)$ and $P(T \geq 3years)$ in the NPV plot.





Dissemination

Early Detection Research Network

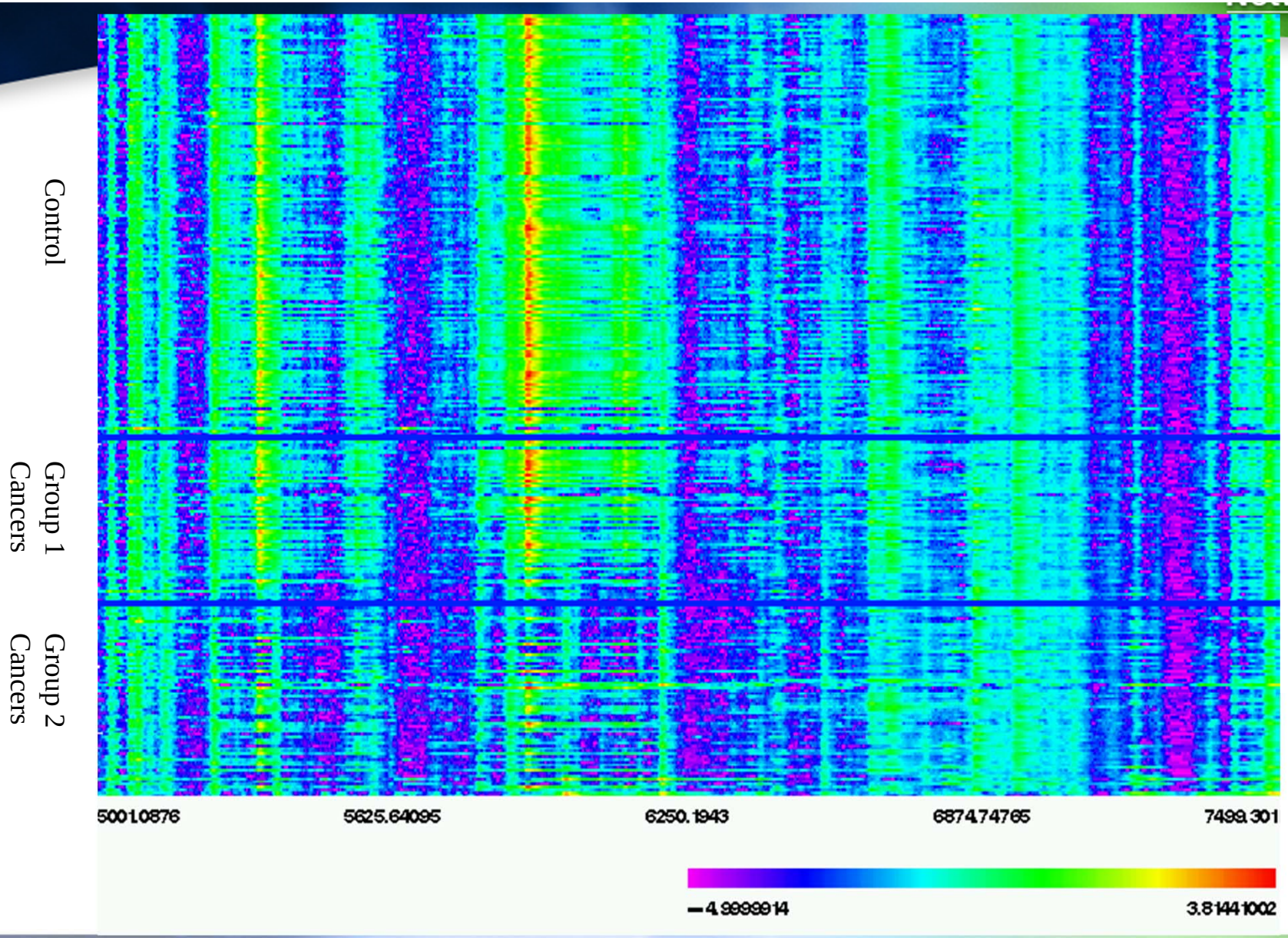
Short Courses

- For statistical audiences
- 2007: FDA (September), JSM (August), SISG (June)
- Plans to expand to other audiences

Software

- DABS website accessible from EDRN website
- Stata commands for ROC analysis to do estimation, comparisons and regression with and without covariate adjustment
- Plans to translate to R, which is free
- Stata commands for power calculations also
- Proteomics data and code on EDRN website. Code incorporated into the Insightful **Splus Proteome** tool

Cases and Controls Ordered by Blood Draw Date



Pepe et al.: Design of Phase II Biomarker Validation Studies Check Lists

Early Detection Research Network

Clinical Context

- Define the population; Enrollment criteria and process; Setting for specimen collection
- Is generality adequate? Consider multiple institutions, heterogeneous population, simple protocol
- Specify procedures for outcome ascertainment; Prospective specimen collection prior to outcome ascertainment?
- Define case (seek biomarker positive in cases) and subsets of interest; Define control (seek biomarker negative in controls) and subsets of interest; Do all subjects in the population fit into a case or control category?
- Random selection of cases and controls?
- Matching of controls to cases on factors related to the biomarker? Consider limitations on questions that can be addressed by the study

Marker Performance

- Define TPR= Proportion of cases positive; FPR= Proportion of controls positive
- Does 'time between obtaining specimen and occurrence of outcome' enter into definition of (TPR,FPR)?
- Are there subgroups of cases and controls for whom (TPR,FPR) will be calculated separately?
- What are minimally acceptable values (or ranges) for (TPR,FPR) in the clinical application?
- For a prediction marker: 'positive' is defined as the prediction probability exceeding a threshold. Provide the threshold and rationale for its choice. Similarly justify a threshold if identification of a low risk subgroup is the study objective.
- For a prediction marker: define adequate calibration of the risk model
- Does a classification method currently exist? What is its performance? How will comparative performance be quantified? What are target levels?

Marker Measurement

•Specify procedures for specimen collection, processing, storage and retrieval
•Specify assay procedures and how results are reported
•Are mechanisms in place to blind specimen handling, assay and reporting of results to outcome status?
•Are the biomarker data to be combined with other information on the patient in the intended clinical application? This could include other clinical information, other markers, previous measurements of the biomarker in the patient. Specific algorithm for calculating the combination must be defined (it cannot be developed within the study).
•If other biomarkers or predictors will be combined or compared: Describe in detail protocols and procedures for obtaining these data; Provide assurance that procurement of these items is blinded to patient outcomes.

Sample Size and Power

•Recall minimally acceptable performance criteria
•Define and provide rationale (data evidence) for anticipated performance levels.
•Calculate case and control sample sizes
•Plan for prospective collection until sufficient numbers of cases and controls are enrolled.
•Plan for early termination of the study if appropriate.

Study Coordination

Early Detection Research Network

Administration
- Conduct planning meeting
- Hold regular conference calls
- Create mailing lists
- Develop timeline
- Track IRB approvals and MTAs
- Develop Manual of Operations
- Resolve intellectual property issues
- Resolve commercial/industrial issues
- Accommodate FDA requirements

Development
- Calculate sample sizes
- Determine matching criteria
- Blind samples
- Perform inter-lab comparisons
- Determine data elements
- Standardize specimen collection, processing and shipping guidelines

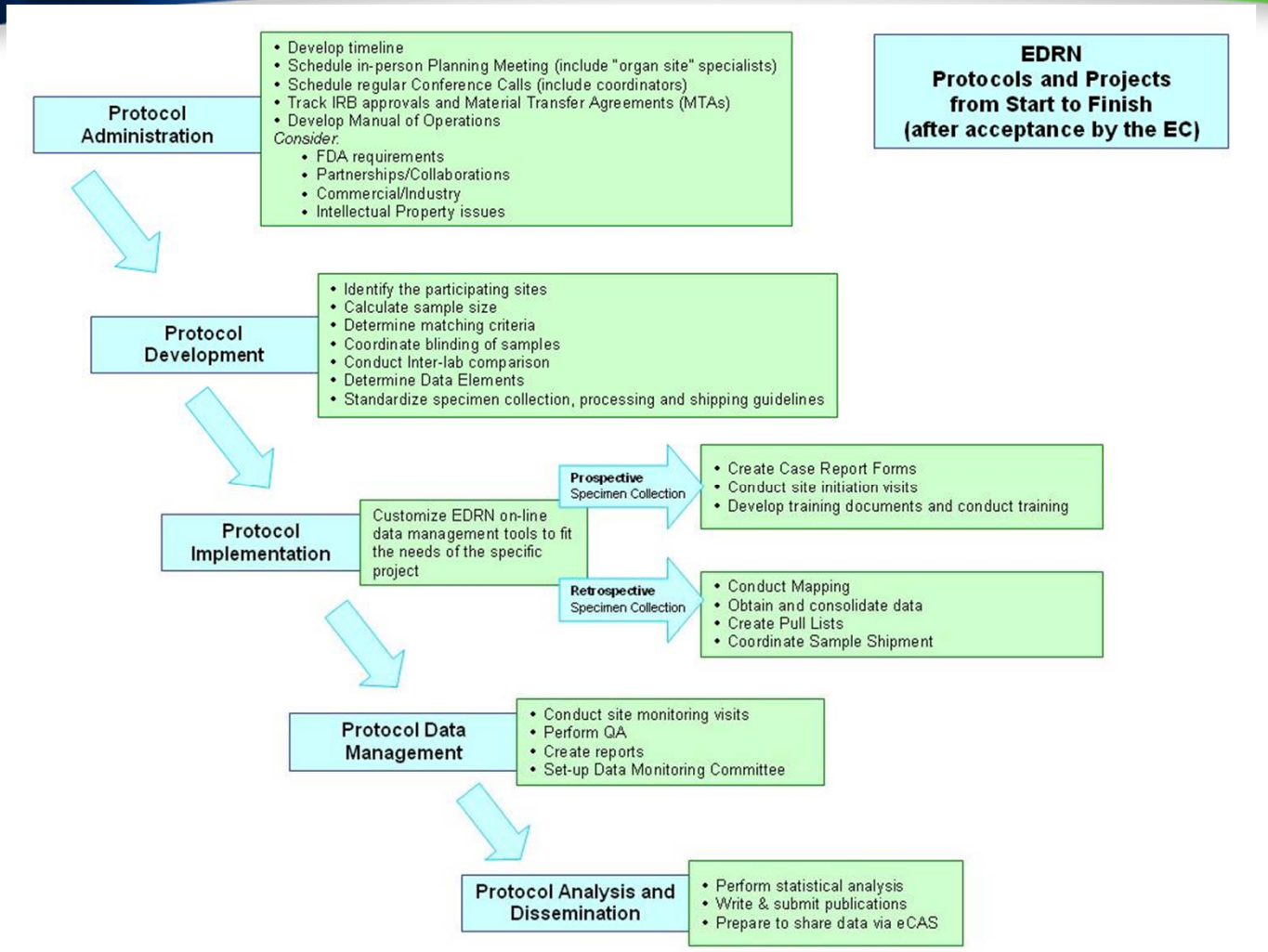
Analysis and Dissemination
- Statistical Analysis
- Publication
- Dissemination of Data

Implementation
- Create case report forms
- Customize data management tools
- Conduct site initiation visits & training
- Obtain data
- Select matched cases and controls
- Perform sample assay assignment

Data Management
- Conduct site monitoring visits
- Prepare reports
- Perform quality assurance
- Hold regular data monitoring committee meetings
- Revise protocol as needed
- Update Manual of Operations as needed
- Distribute revisions and updates to performance sites

Flow Chart of DMCC Study Coordination

Early Detection Research Network



Validation Study Information Management System (VSIMS)

Early Detection Research Network

EDRN's web-based data management system to promote consistent study execution and high-quality study data collected using Common Data Elements (CDEs)

Data Entry

QA

Specimen Tracking

Administrative

Flexible
Consistent
Easy to Implement
Easy to Maintain
Traceable

VSIMS Functionality

Early Detection Research Network

Data Entry
- Audit trail
- Built-in edit checks
- Easy access to instructions
- Print CRFs for site's records
- Participant privacy

Confirm Eligibility
- Consent form status tracking
- Required/optional data elements
- Assigned study group
- Visit schedules

Issue Tracking
- Data Clarification Form
- Inquiry to Coordinating Center
- Request from Coordinating Center
- Status
- Full History

Specimen Tracking System
- Tracks Original and Child Specimens
- 2-D barcodes & heat thermal printed labels
- Tracks specimen-related CDEs
- Assay laboratories are blinded to participant information
- All specimens collected in standardized fashion
- Cases and Controls are paired according to matching criteria defined by the protocol
- Pull Lists are created and Algorithm Batch Lists are designed for processing lab

Study Reports
- Participant Reports
  - Enrollment
  - Adherence
  - Completeness
- Specimen Reports
  - Defective List
  - Specimens Collected
  - Status of Assay Results
  - Pull Lists

Study Information
- Protocol
- Manual of Operations
- Bulletins
- Mailing List Subscriptions
- Directory
- Study-specific newsletters
- IRB

Satisfies 21 CFR Part 11 requirements for data systems and so can be used to support studies intended for submission to FDA

EDRN'S CONTRIBUTIONS TO BIOMARKER RESEARCH

Early Detection Research Network

- **Phase structure for biomarker discovery and validation studies is being used extensively by the extramural community**
- **Recognition of chance, bias, and other confounders in biomarker research through EDRN experience; Corrective measures and methodologies developed by EDRN are acknowledged by the statistical community**
- **Dissemination of methodologies, study design issues and validation concepts to the research and regulatory communities, e.g., EDRN-FDA meeting, statistical workshops**
- **Study designs and methodologies developed for proteomic-based diagnostics are proving useful to other researchers**

Early Detection Research Network

Connecting Biomarker Research Across the EDRN

Dan Crichton

Program Manager, Planetary Data System Engineering

PI, EDRN Informatics

Principal Computer Scientist

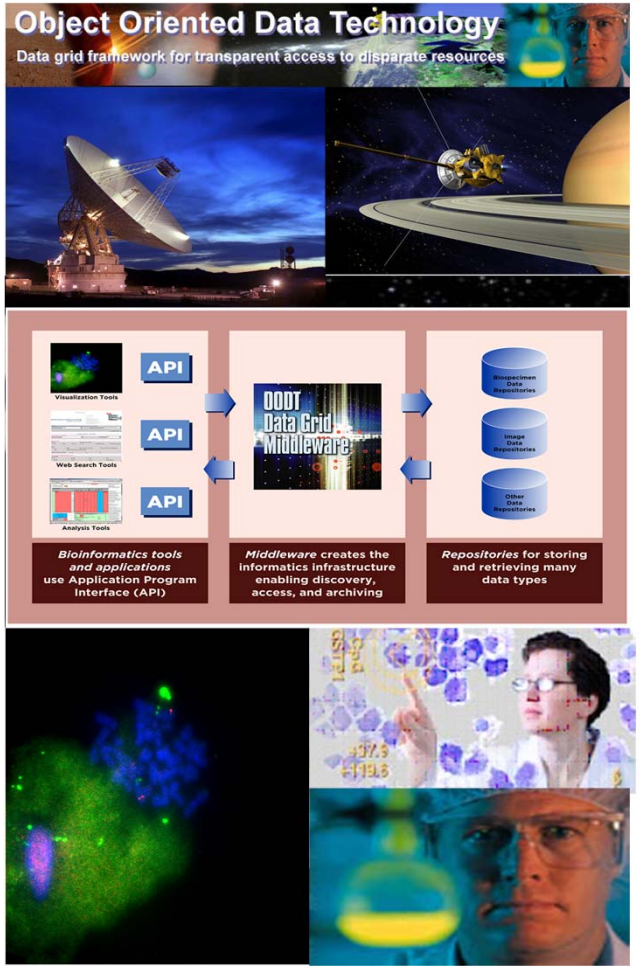
Jet Propulsion Laboratory, Caltech

Motivation and Challenges for Informatics in Science

Early Detection Research Network

- **Distributed informatics environments are being developed to support the capture and dissemination of scientific results across physical and life sciences**
  - Good success cases exist (e.g., Planetary Data System)
  - Provides a data management research platform for distributed communities
  - Technology is not the limiting factor; technologies and standards exist to connect enterprises together

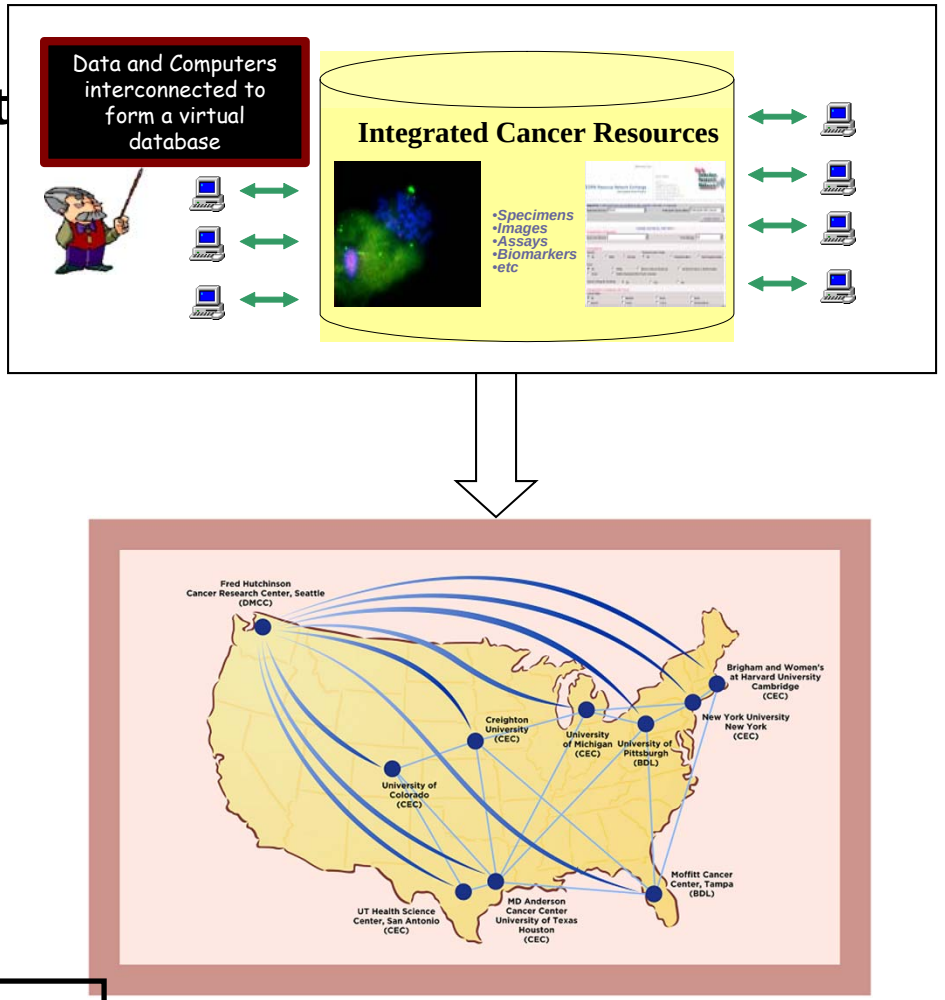
- **The challenge is informatics infusion into the scientific discovery process**
  - Requires an “informatics architecture” to support the scientific discovery process
  - Operational model needs to include cross-disciplinary team including working scientists; teams are often organized around sub-disciplines



Informatics for Biomarker Research

Early Detection Research Network

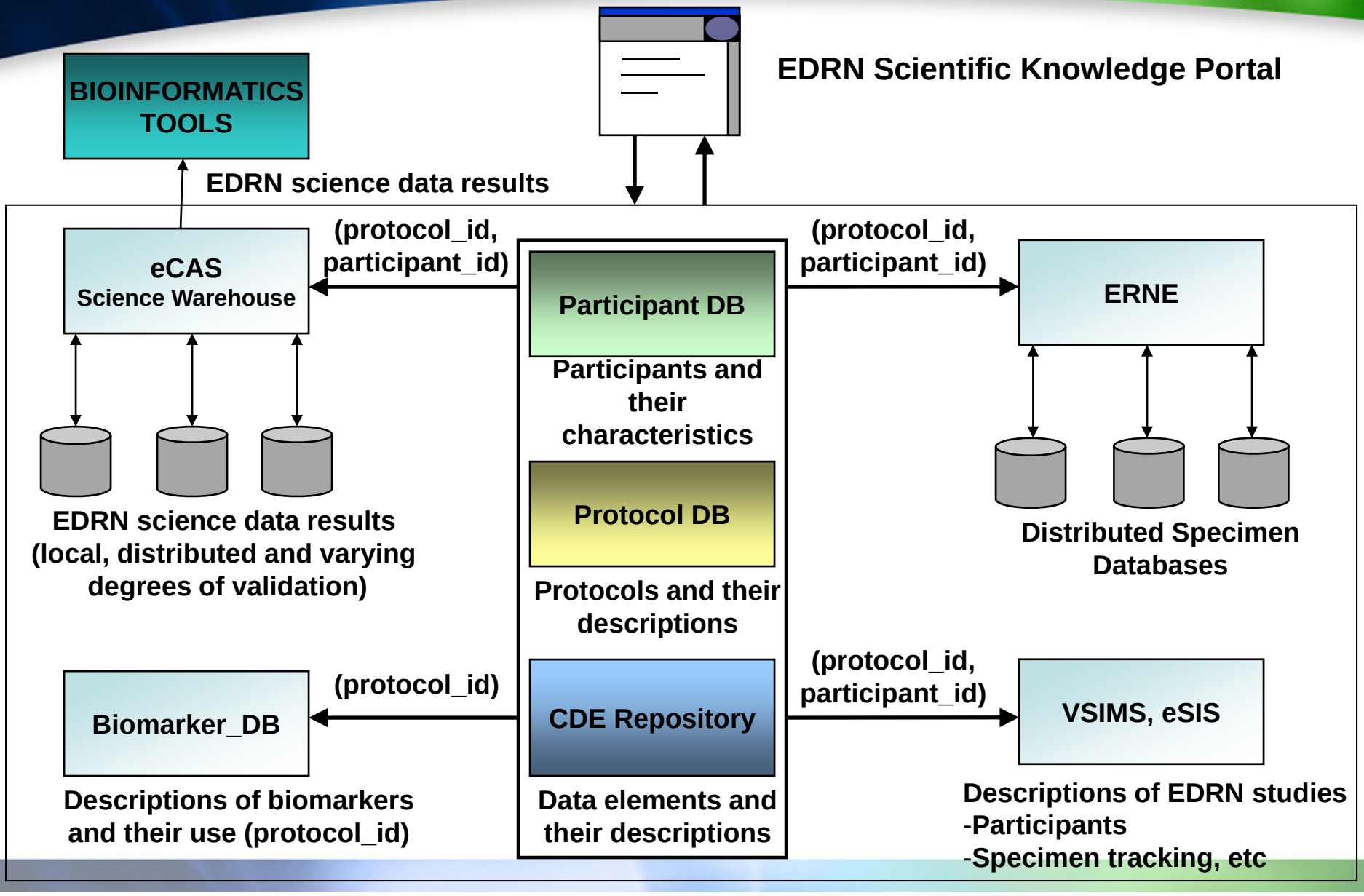
- **EDRN has pioneered in the use of informatics technologies to support biomarker research**
- **EDRN has developed a comprehensive infrastructure to support biomarker data management across EDRN's distributed cancer centers***
- **It supports capture and access to a distributed diverse set of information and results**
  - •Biomarkers
  - •Proteomics
  - •Biospecimens
  - •Various technologies and data products (image, micro-satellite, …)
  - •Study Management



*** EDRN was highlighted for its informatics in a recent NCI-FDA-AACR report on biomarker research**

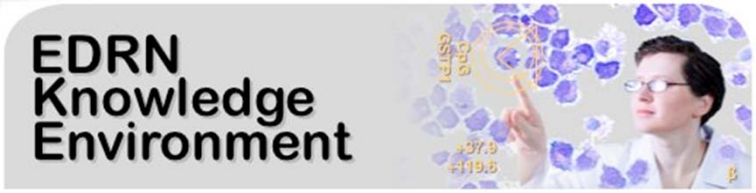
Biomarker Informatics Architecture

Early Detection Research Network

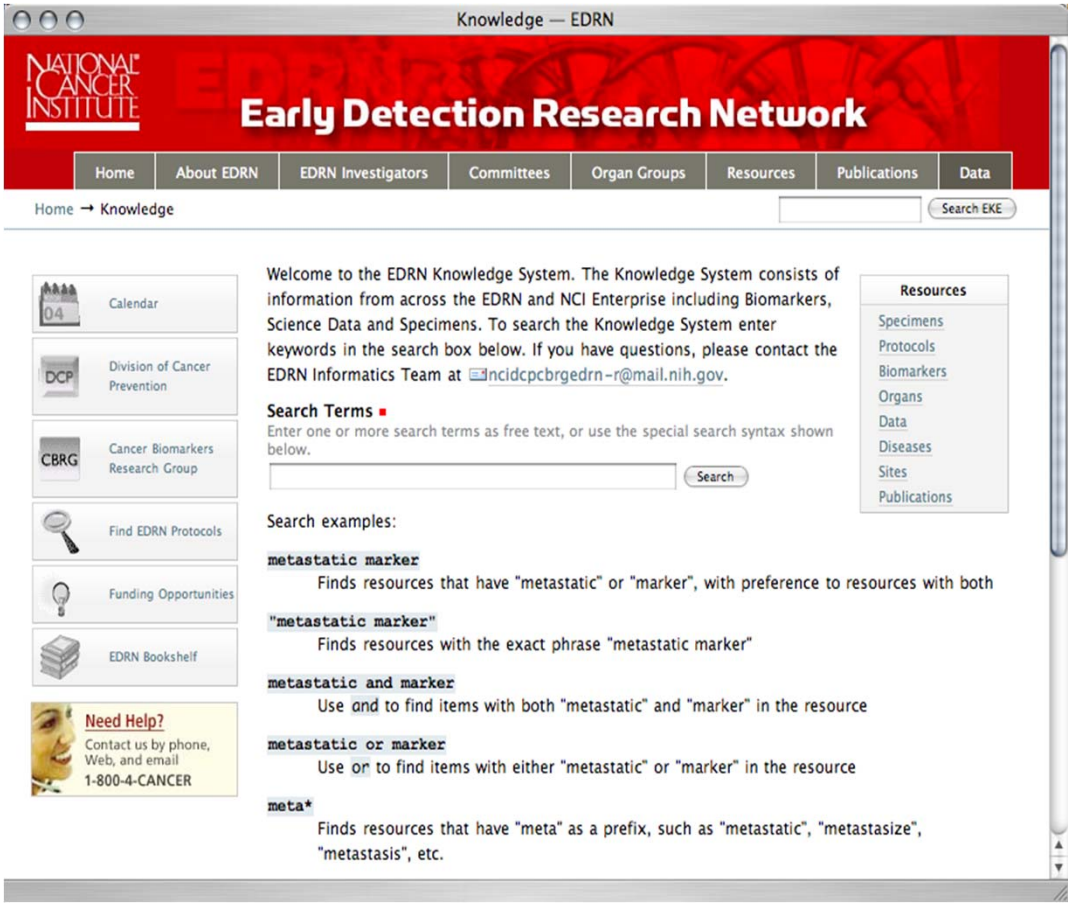


Integrated EDRN Knowledge Portal

Early Detection Research Network



- **Highly scalable scientific information portal for biomarker research**
- **Adapts to a dynamic set of "object types" (mass spec, immunohistochemistry, etc) based on the biomarker information model**
- **Provides access to distributed repositories of information**
- **Provides "google-like" search of the object types**



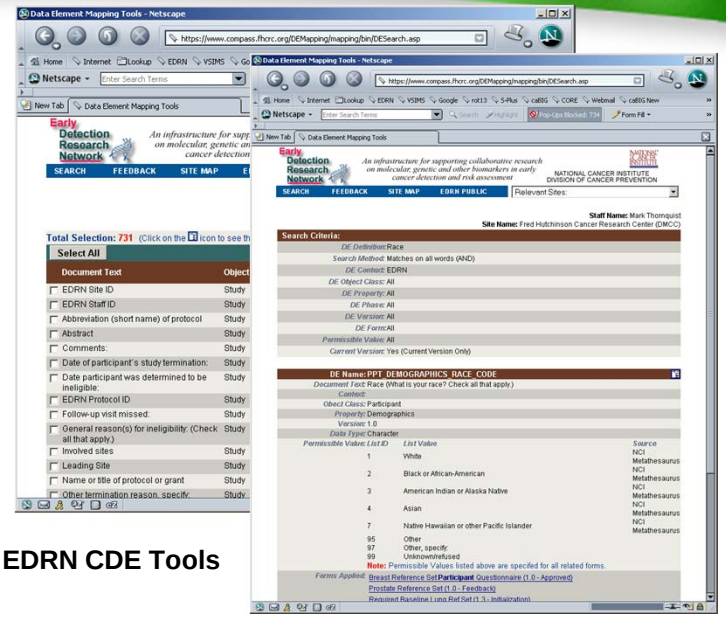
EDRN Public Science Portal

Informatics Standards for Data Capture and Curation

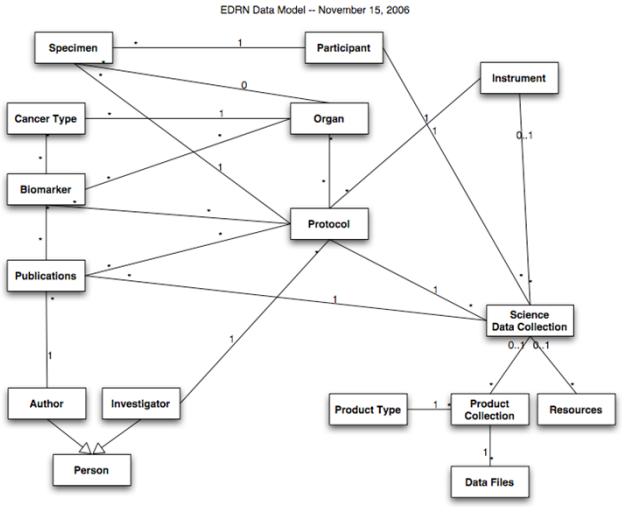
Early Detection Research Network

- **EDRN has developed a High level *ontology* model for biomarker research which provides standards for the capture of biomarker information across the enterprise**
- **Specific models are derived from this high level model**
  - –Model of biospecimens
  - –Model for each class of science data
- **EDRN is specifically focusing on a granular model for annotating biomarkers and their studies**
- **EDRN has a set of EDRN CDEs which is used to provide standard data elements and values for the capture and exchange of data**

EDRN's underlying biomarker information model needs to be flexible in order to manage a variety of different types of science data captured by different experiments at different phases in studying biomarkers…



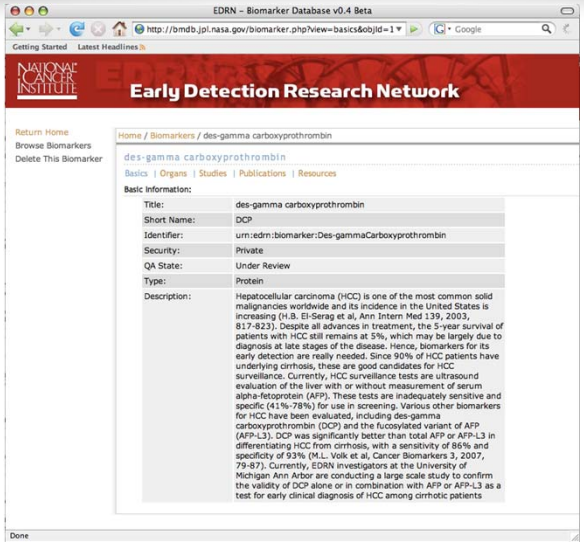
EDRN CDE Tools



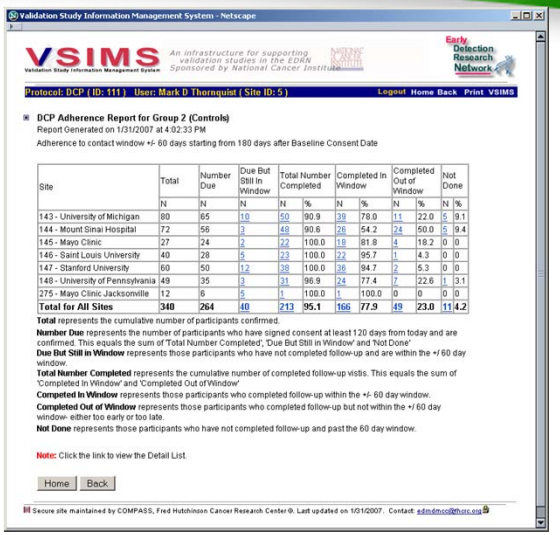
EDRN Biomarker Ontology Model

Biomarker Data Management

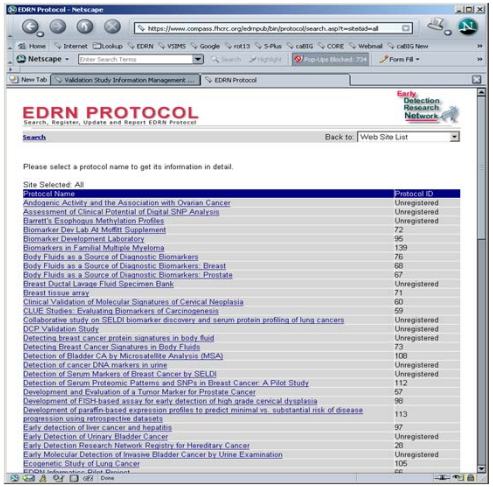
- **EDRN provides an infrastructure to capture, curate and manage biomarker information including:**
  - Biomarker Annotations (Biomarker Database)
  - Study Design (VSIMS)
  - Protocols (eSIS)
  - Study Results (Science Data Warehouse)
- **All databases are designed using EDRN Common Data Elements in order to link EDRN into an integrated enterprise**



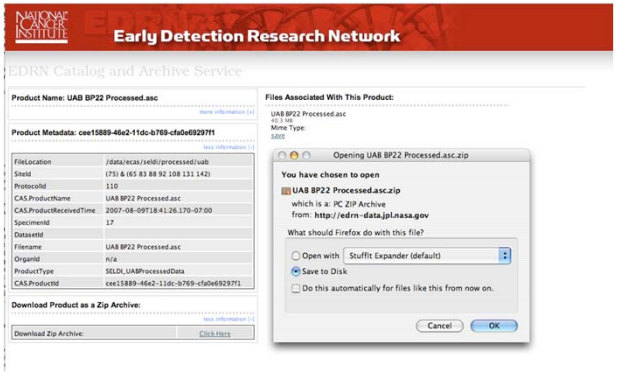
Biomarker Database



Validation Study Database



eSIS Protocol Database

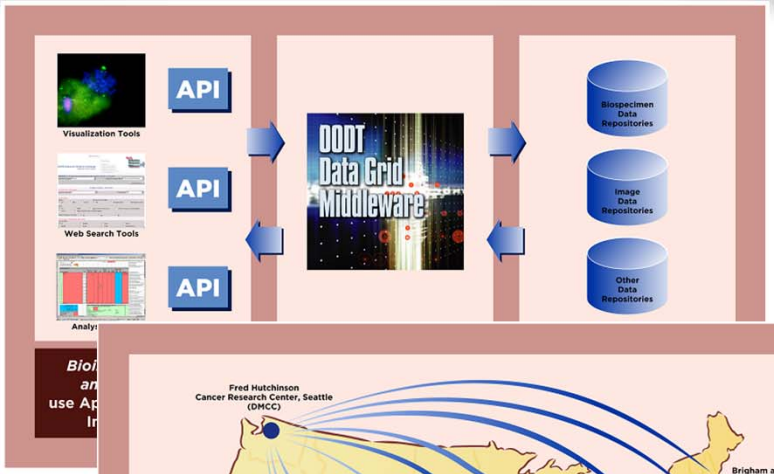


Science Data Warehouse of EDRN raw and processed science data

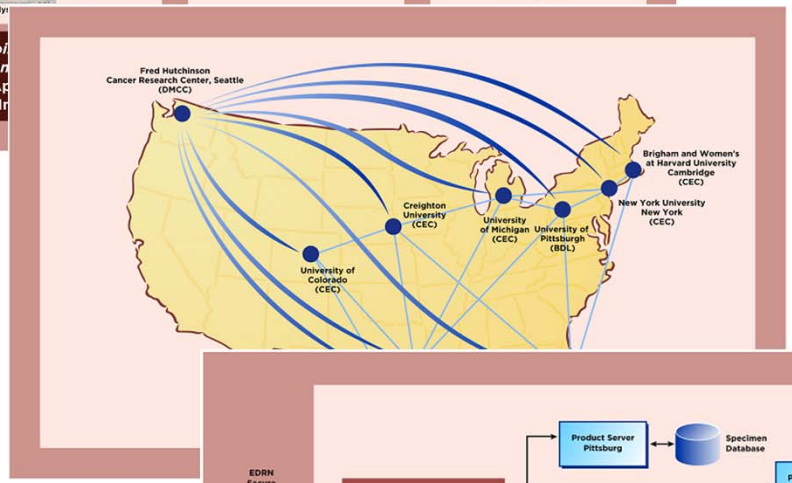
Access to Specimen Information

Early Detection Research Network

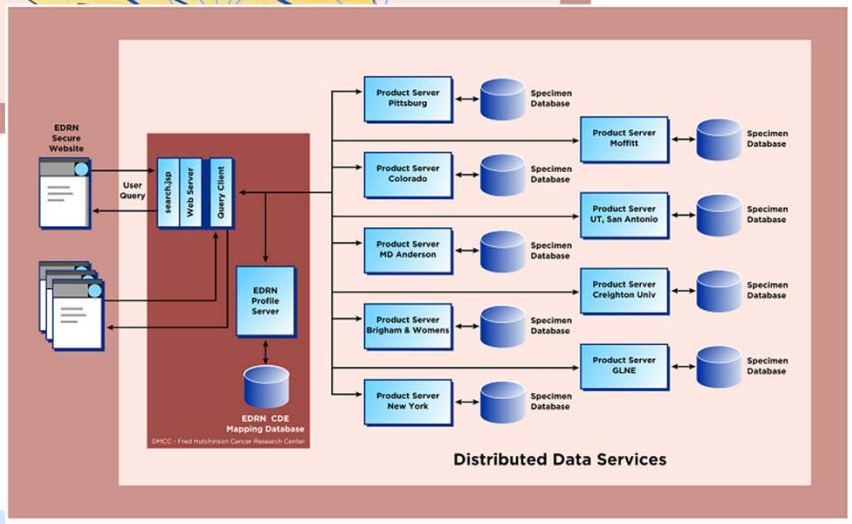
- **EDRN has architected, ERNE, the EDRN Resource Network Exchange which provide access to specimens**
  - Information about available specimens remains managed at cancer centers
  - Access is built using JPL open source software for access to distributed repositories, the same software used to manage access to planetary, earth and astrophysics repositores
- **EDRN has been successful at integrating a diverse community of institutions and technologies**
  - Many, many different technologies have been integrated
  - EDRN has succeeded in working within different local institutional policies (IRB, Security, etc)
  - EDRN and caBIG have worked together to demonstrate how caBIG tools can plug into EDRN (e.g., caTissue)



JPL developed data grid infrastructure



Deployed across several EDRN sites (~12)



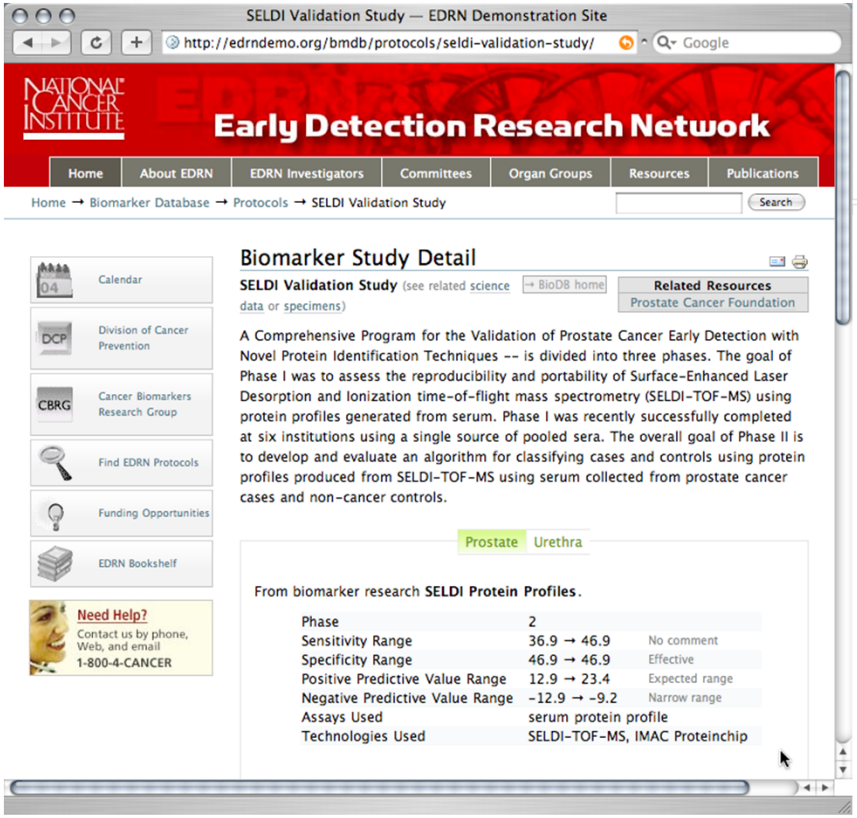
Access to many different repositories, using many different technologies

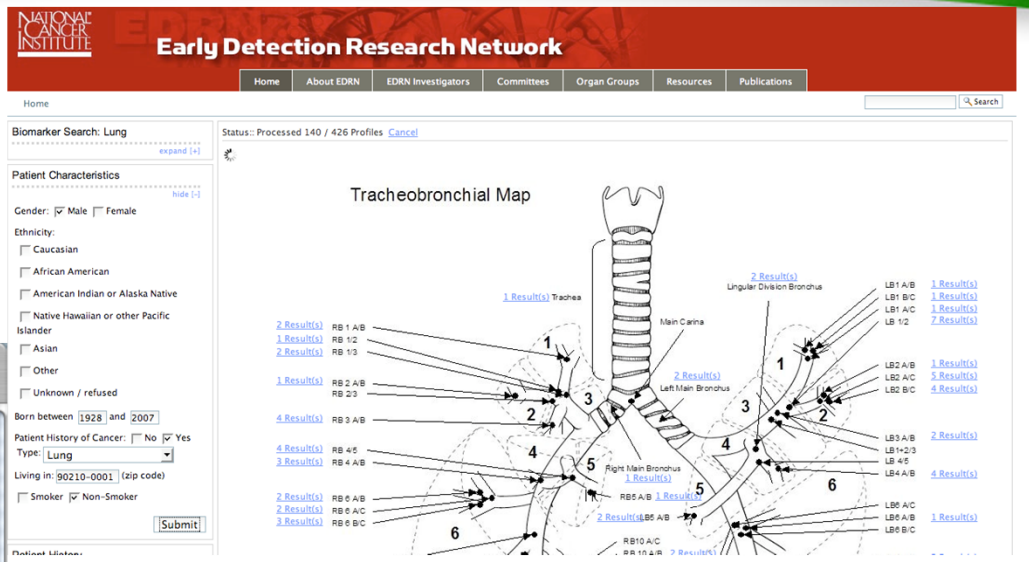
Searching for Biomarker Information

Early Detection Research Network

Focus on access to science data using general and organ-specific tools

- **Google-like Searching**
- **"Map-based" Searching for organ-centric information**





Interactions across NCI

Early Detection Research Network

- **caBIG Interactions**
  - EDRN and caBIG have worked together to integrate ERNE and caTissue
    - A working pilot exists which will plug in any caTissue site into ERNE
  - EDRN's underlying Common Data Elements are built using ISO 11179, a standard that caBIG has embraced for data elements
  - EDRN has submitted its CDEs to the caDSR
  - EDRN has participated and presented in several caBIG meetings

- **Other interactions**
  - EDRN has met with and shared its architecture and success with other groups (e.g., Prostate NBN Pilot)
  - EDRN participated in the NCI-FDA-AACR meetings on an infrastructure to support biomarker research providing significant input related to the EDRN informatics architecture
  - EDRN participated in the studies on the National Biospecimen Network

Challenges and Next Steps

Early Detection Research Network

- **EDRN has developed a comprehensive informatics infrastructure for Biomarkers Research**
  - It's now at an inflection point as it moves towards an operational model to capture and curate the information from EDRN studies
  - Informatics needs to continue to be agile, evolving to support science research
- **New bioinformatics tools can now be developed on top of the informatics infrastructure which has laid a foundation for access, management and sharing of biomarker information**
- **EDRN plans to continue to pioneer its novel biomarker knowledge system to expand the data and information that is available and to ensure that it is leveraging and developing modern technologies to build the "e-science" infrastructure for biomarker research***

* D. Crichton, S. Kelly, C. Mattmann, Q. Xiao, J. S. Hughes, J. Oh, M. Thornquist, D. Johnsey, S. Srivastava, L. Esserman, W. Bigbee. A Distributed Information Services Architecture to Support Biomarker Discovery in Early Detection of Cancer. In Proceedings of the 2nd IEEE International Conference on e-Science and Grid Computing, pp. 44, Amsterdam, the Netherlands, December 4th- 6th, 2006.

Acknowledgements

Mark Thornquist and members of Data Management and Coordinator Center at the Fred Hutchinson Cancer Research Center

Sudhir Srivastava and Don Johnsey, National Cancer Institute

The EDRN informatics advisory group Bill Bigbee, Laura Essermann, Wilbur Franklin, Tony Hollingsworth, Jeffrey Marks

Current sites collaborating on informatics:

- H. Lee Moffitt Cancer Center
- University of Texas, San Antonio
- Creighton University
- University of Colorado
- University of Pittsburgh
- University of Michigan/Dartmouth University (Great Lakes New England Consortium)
- Brigham and Womens
- MD Anderson
- New York University
- Duke
- Johns Hopkins University
- University of San Francisco
- Fox Chase Cancer Center
- Beth Israel
- Center for Disease Control
- Roswell Park

NASA Jet Propulsion Laboratory

Early Detection Research Network

Backup Slides

Phases of Informatics Development

Early Detection Research Network

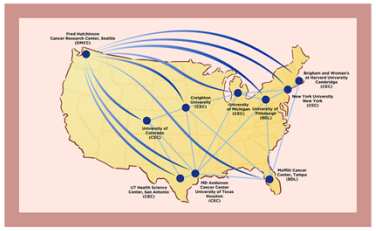
2002

Initiate System Architecture

- EDRN works with NASA-JPL to Develop Architecture
- Basic Information Model Developed
- Initial set of sites connected

Middleware for linking EDRN distributed data repositories

2004



Extend Core Framework

- Extend middleware to link nine EDRN sites together
- Common tools to manage deployed system
- Portal developed to access information
- Validation Study Infrastructure

Major EDRN sites connected

2006

EDRN Portal

- EDRN begins building new tools for study and biomarker management
- ERNE is extended to new sites
- EDRN deploys a public portal

Public Access to EDRN Data

2008

Operational deployment of biomarker databases

- EDRN deploys its knowledge portal for science acc ess
- EDRN deploys its curation infrastructure for biomarker data
- EDRN builds new tools for organ-specific access to data
- EDRN establishes a curator network for data

EDRN begins full operational deployment of its tools, focusing on curation of the data