Children's Hospital Boston
Informatics Program

Harvard-MIT Division of
Health Sciences and Technology

HST

Using p2p systems to translate routine tissue collection and clinical encounters into robust research discovery

PSL

Pathology Specimen Locator

Andrew_McMurry@hms.harvard.edu

Primary support: NCI, NLM, CDC, DF/HCC, Autism Consortium, Harvard CTSA

Translational Research

Routine care delivery → Robust research discovery

use cases

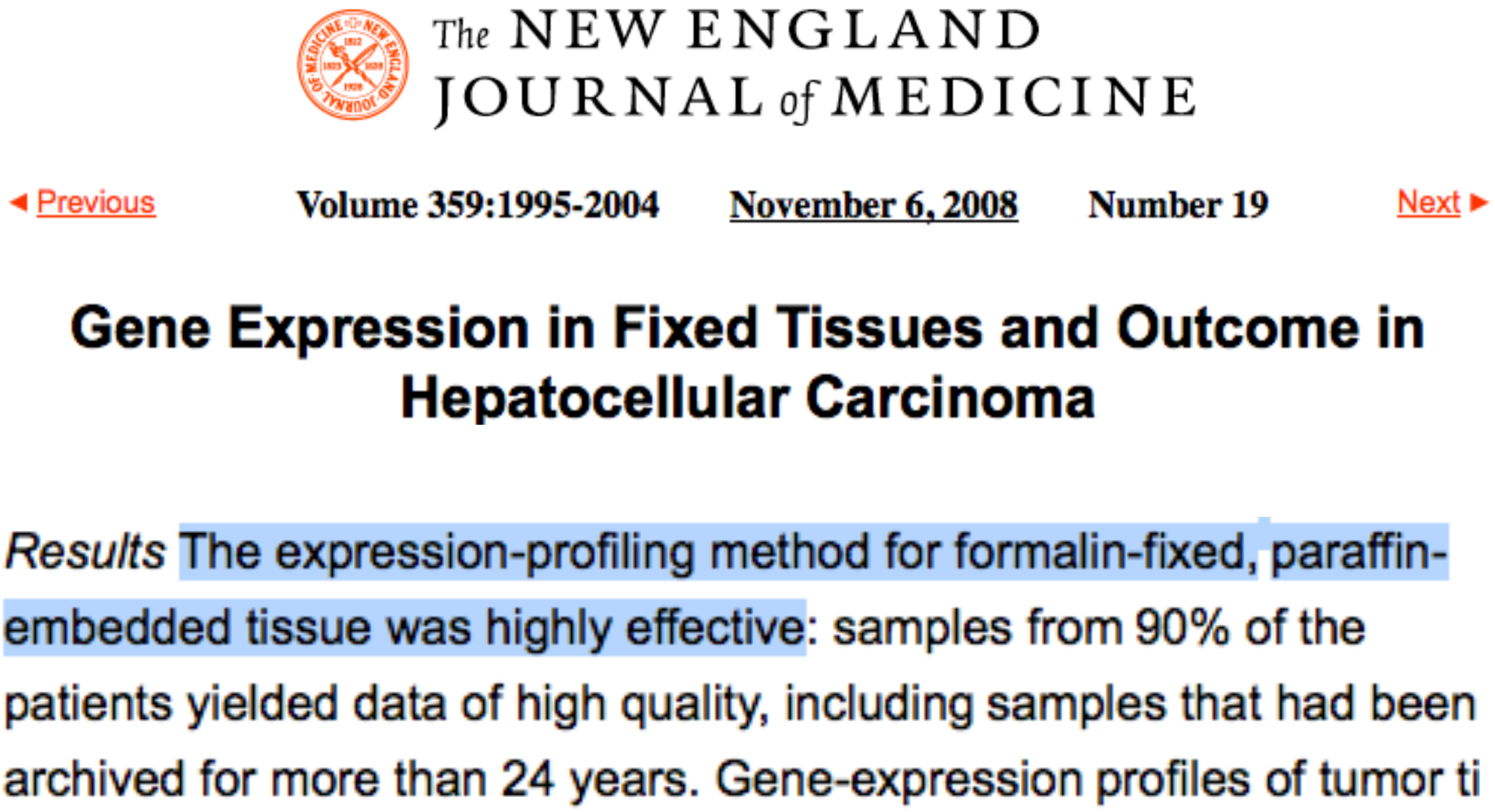
1. Share routinely collected human tissues for biomarker discovery and high-throughput validation

2. Share experimental outcomes derived from tissue processing, with an emphasis on genomic measurements

- **NCI Vision 2001**
  - **Millions of Paraffin Embedded Tissues**
    - Biomarker Discovery / Validation:

      DNA fragments of up to 400 bp and RNA fragments of up to 150 nucleotides can be routinely isolated for mutation detection, SNP analysis, detection of translocation, and microRNA quantification. Pathology services and screens, TMA construction, …
  - **Smaller Collections of Fresh / Frozen Tissues**
    - DNA/RNA Microarrays, chip-chip, chip-seq, etc.

for Translational Research Requiring Human Specimens



The NEW ENGLAND JOURNAL of MEDICINE

◄ Previous Volume 359:1995-2004 November 6, 2008 Number 19 Next ►

Gene Expression in Fixed Tissues and Outcome in Hepatocellular Carcinoma

Results The expression-profiling method for formalin-fixed, paraffin-embedded tissue was highly effective: samples from 90% of the patients yielded data of high quality, including samples that had been archived for more than 24 years. Gene-expression profiles of tumor ti

Sharing Human Tissues for Discovery and Validation

Challenges

- How to link routine pathology databases for research?
  - *Local Control* → each hospital is a "peer" on the network
- How to ensure patient privacy in accordance with HIPAA?
  - *Local Control* → anonymization and statistical aggregates
- How to engender hospital participation?
  - *Local Control* → hospitals remain owners of specimens and stewards of patient data

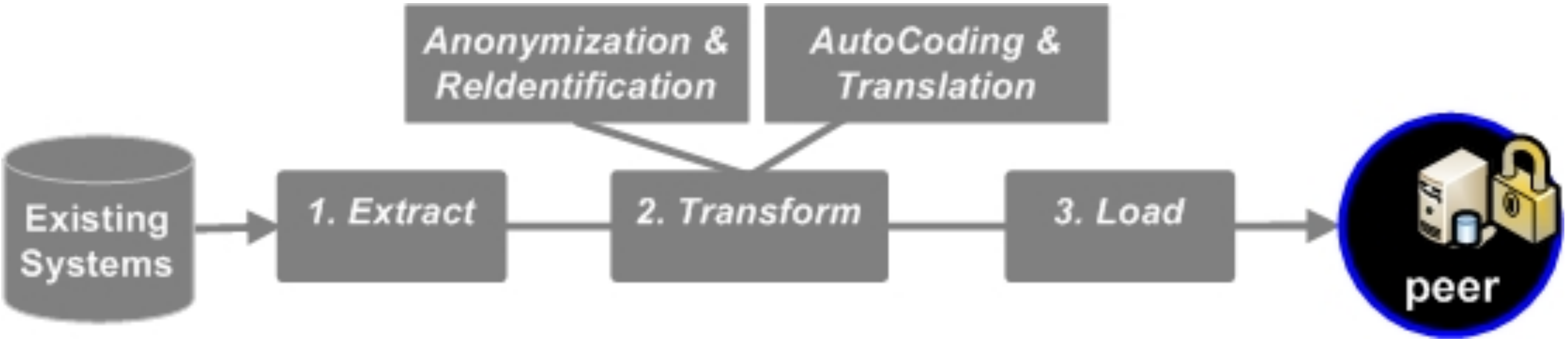
How It Works

1. **Link existing databases**
   - Extract from existing hospital systems
   - Transform the data into common HIPAA-safe vocabulary
   - Load into locally controlled "SPIN peer" with deidentified ID

2. **Protect Patient Privacy per HIPAA**
   - Anonymized: Case Counts / Aggregates
   - Limited: When authorized for individual cases
   - *PHI is rarely used, and only with IRB from each hospital.*

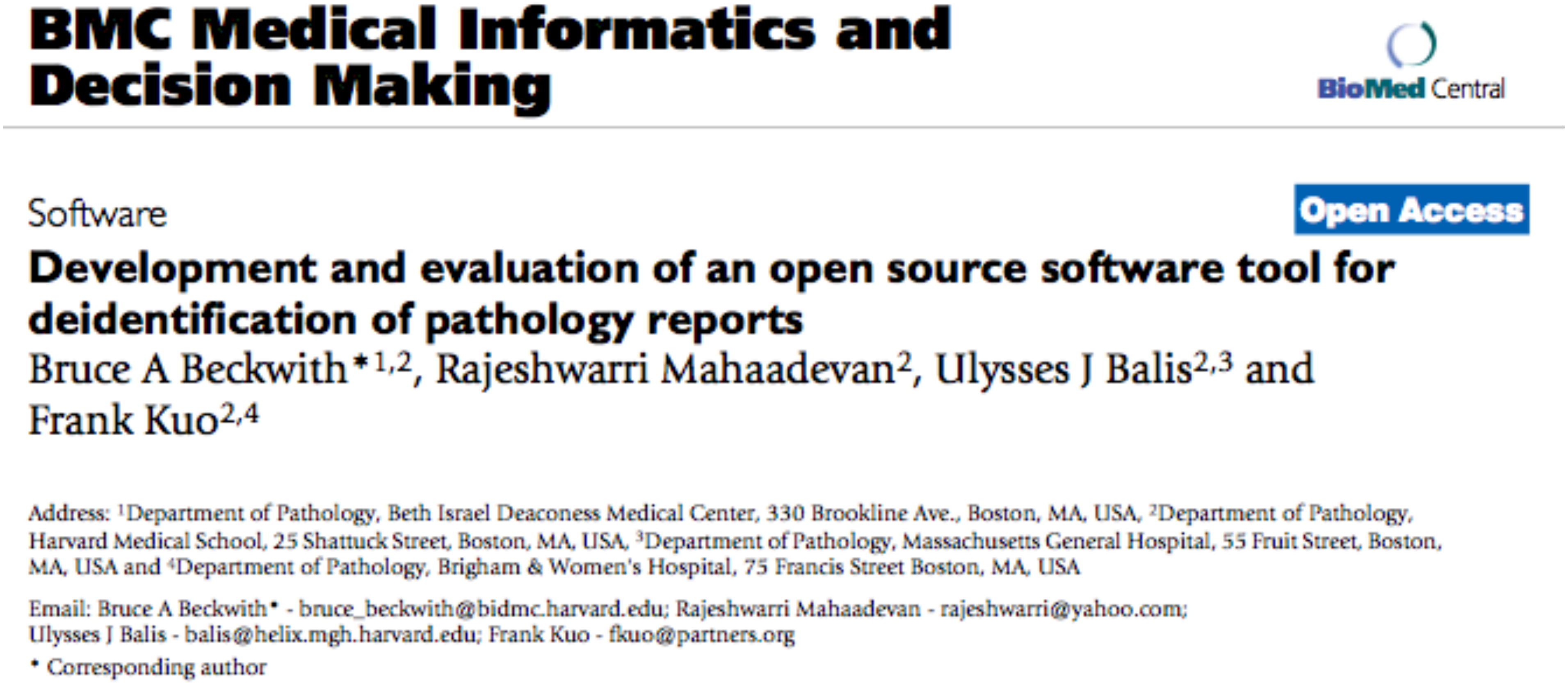
3. **Hospital Control**
   - No central governing body or server
   - Peers (hospital) remains in control over disclosures at all times

(1) Linking routine care systems



- Extract from routine care delivery systems
  - Databases *or* XML
- Transform free text reports
  - "Scrub" patient identifiers (per HIPAA)
  - NLP (autocode) into controlled vocabularies such as UMLS
- Load into the hospital controlled PEER database
  - Assign a randomly generated ID to each case

Transforming Free Text: "Scrubber"



BMC Medical Informatics and Decision Making

BioMed Central

Software

Open Access

Development and evaluation of an open source software tool for deidentification of pathology reports

Bruce A Beckwith*[1,2], Rajeshwarri Mahaadevan[2], Ulysses J Balis[2,3] and Frank Kuo[2,4]

Address: [1]Department of Pathology, Beth Israel Deaconess Medical Center, 330 Brookline Ave., Boston, MA, USA, [2]Department of Pathology, Harvard Medical School, 25 Shattuck Street, Boston, MA, USA, [3]Department of Pathology, Massachusetts General Hospital, 55 Fruit Street, Boston, MA, USA and [4]Department of Pathology, Brigham & Women's Hospital, 75 Francis Street Boston, MA, USA

Email: Bruce A Beckwith* - bruce_beckwith@bidmc.harvard.edu; Rajeshwarri Mahaadevan - rajeshwarri@yahoo.com; Ulysses J Balis - balis@helix.mgh.harvard.edu; Frank Kuo - fkuo@partners.org

* Corresponding author

http://spin.chip.org/software.html

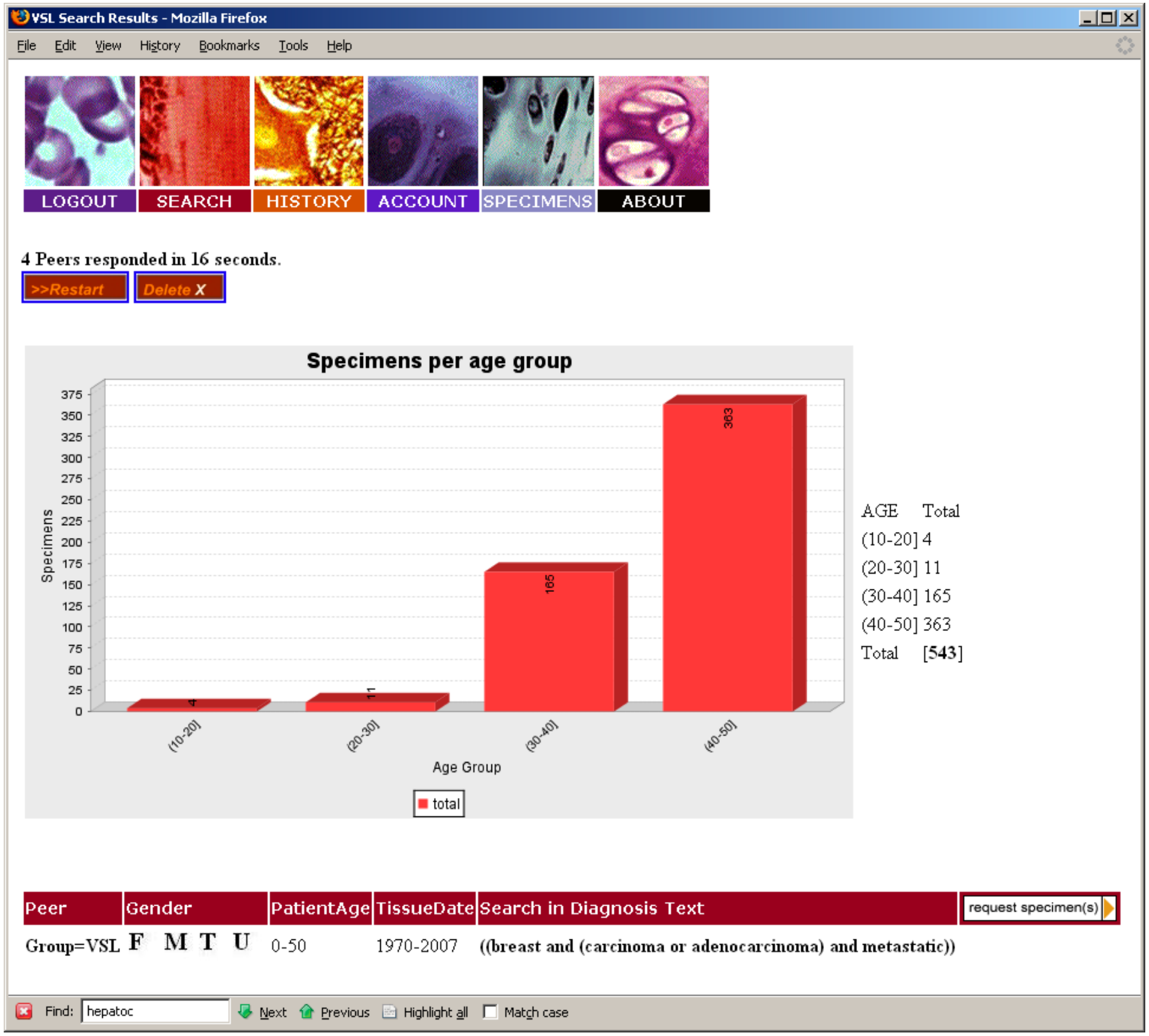
(2) Protecting Patient Privacy

Increasing levels of investigator access commensurate with authorization by the hospital & investigator demonstrated need.

Use Case	*Statistical Queries* *> 90%*	*Non Identifying* *< 10%*	*PHI* *< 1%*
Tissue Sharing	**Feasibility Studies**	**Case Selection**	**Clinical Data**
Public Health	**Automated Analysis**	**Investigation**	**Emergencies**
Genomic Studies	**Significant Markers**	**Case Selection**	**Genotypes**

Feasibility Study:

ascertain if there are enough samples available

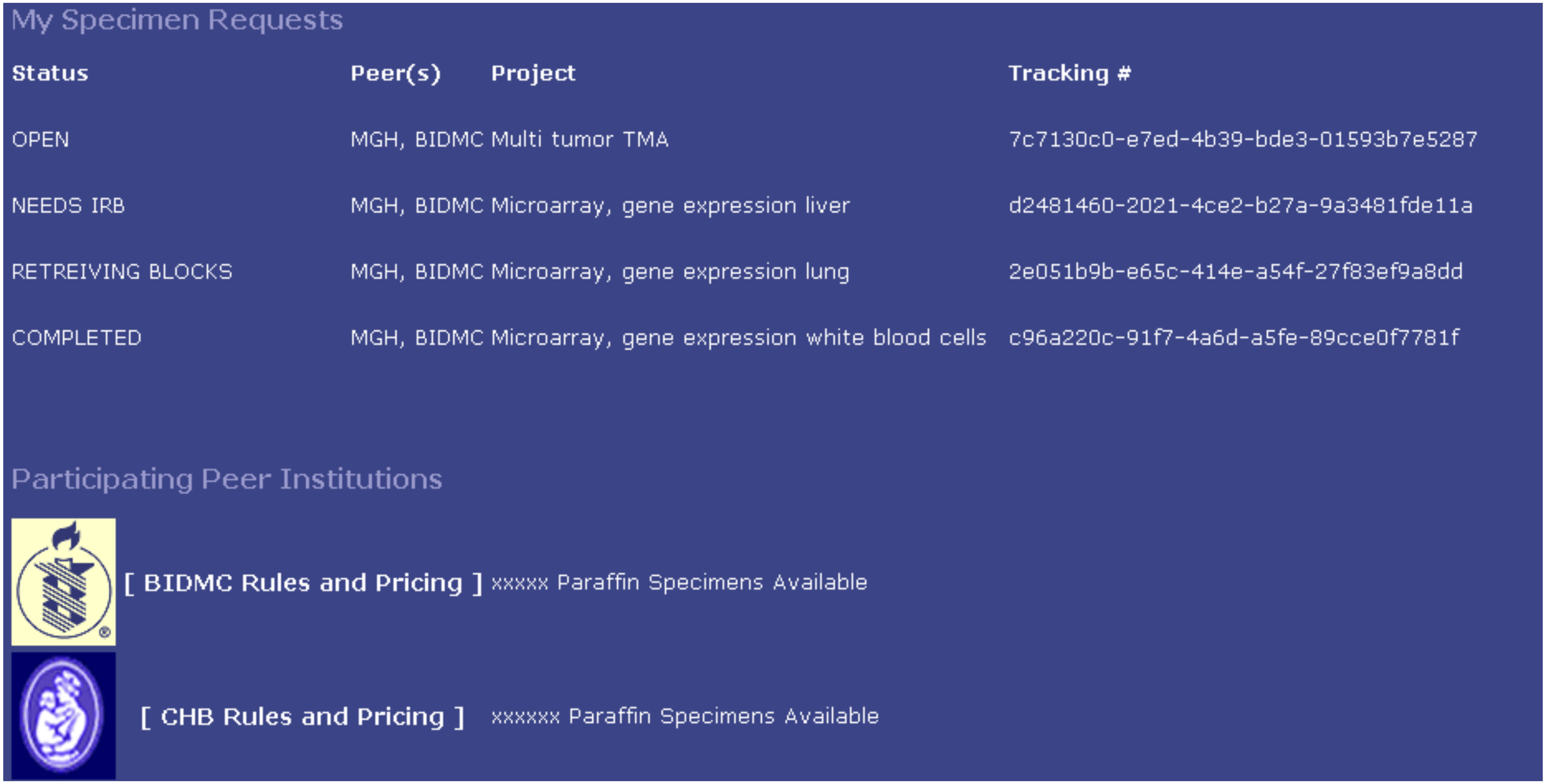


4 Peers responded in 16 seconds.

AGE	Total
(10-20]	4
(20-30]	11
(30-40]	165
(40-50]	363
Total	[543]

Peer	Gender	PatientAge	TissueDate	Search in Diagnosis Text
Group=VSL	F M T U	0-50	1970-2007	((breast and (carcinoma or adenocarcinoma) and metastatic))

Case Selection and Retrieval



My Specimen Requests

Status	Peer(s)	Project	Tracking #
OPEN	MGH, BIDMC	Multi tumor TMA	7c7130c0-e7ed-4b39-bde3-01593b7e5287
NEEDS IRB	MGH, BIDMC	Microarray, gene expression liver	d2481460-2021-4ce2-b27a-9a3481fde11a
RETREIVING BLOCKS	MGH, BIDMC	Microarray, gene expression lung	2e051b9b-e65c-414e-a54f-27f83ef9a8dd
COMPLETED	MGH, BIDMC	Microarray, gene expression white blood cells	c96a220c-91f7-4a6d-a5fe-89cce0f7781f

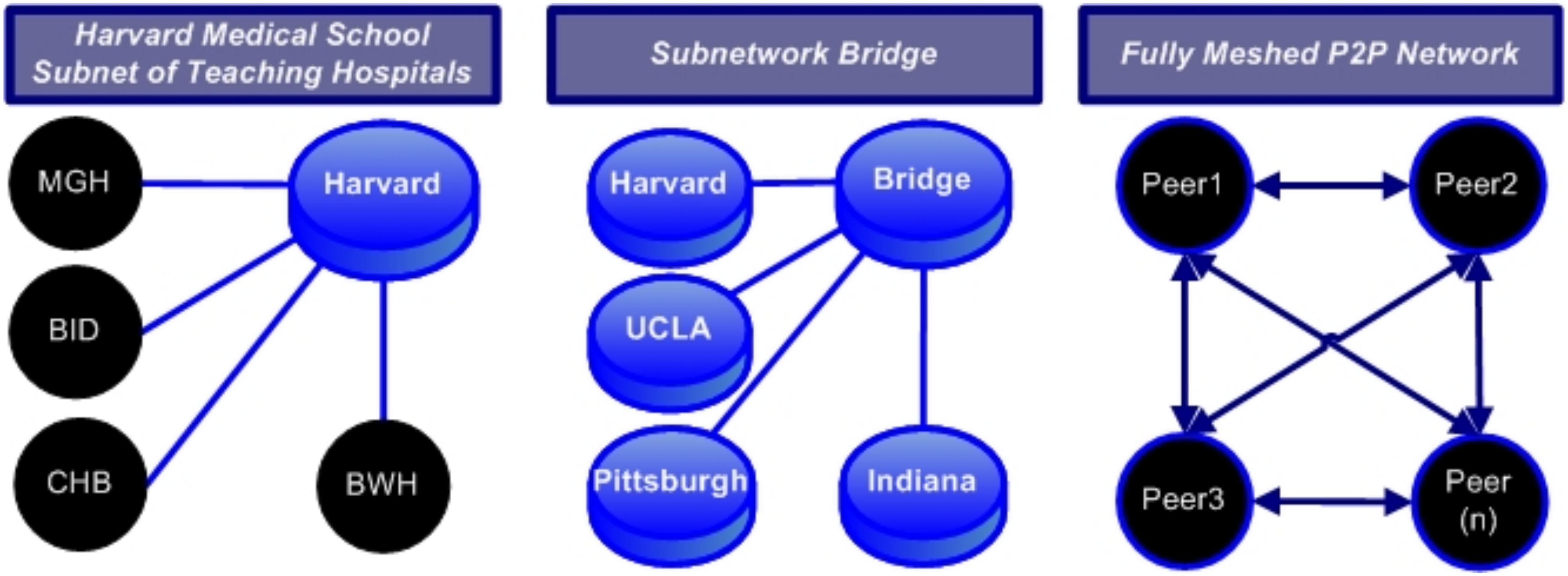
Participating Peer Institutions

[BIDMC Rules and Pricing] xxxxx Paraffin Specimens Available

[CHB Rules and Pricing] xxxxxx Paraffin Specimens Available

(3) Hospitals remain in control

- Each hospital (Peer) chooses **who** to share with



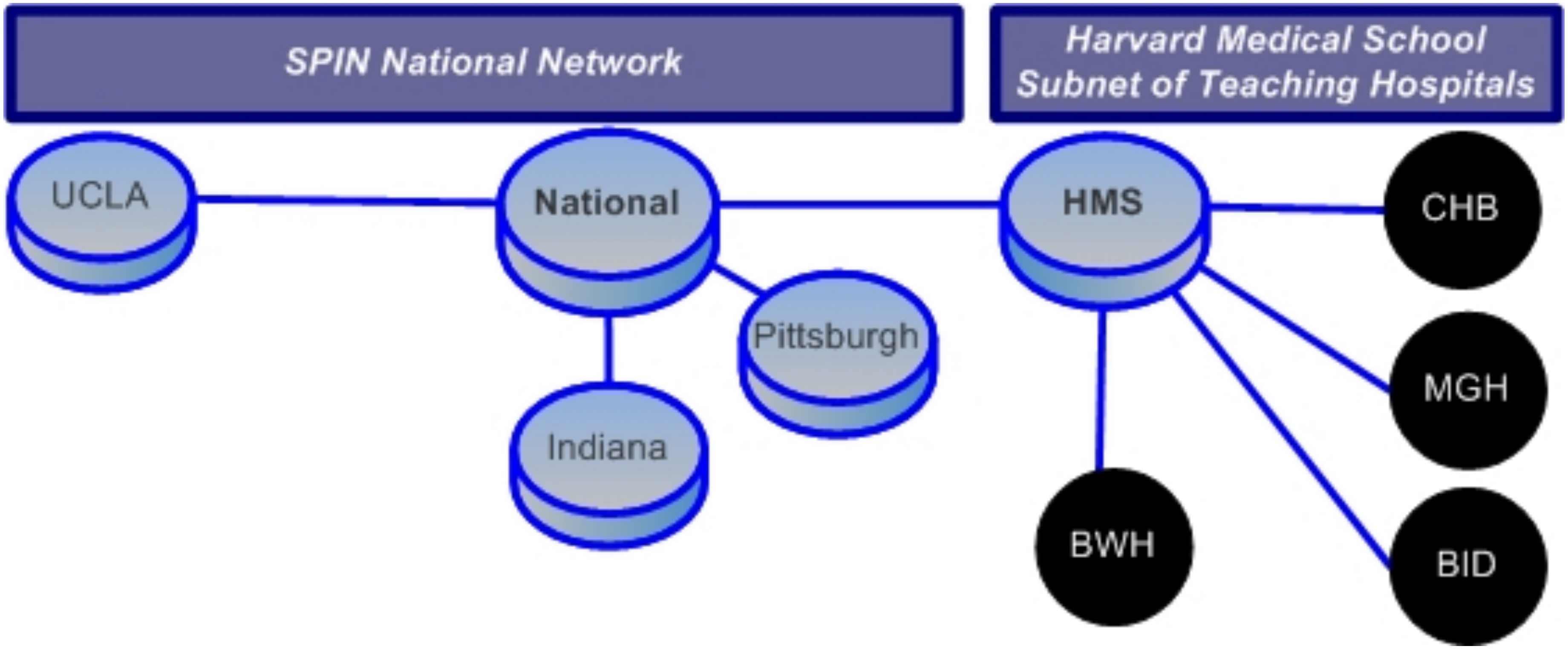
- And **what** to share (Path Reports, ED Visits, …)

Sites Participating in National Demonstration

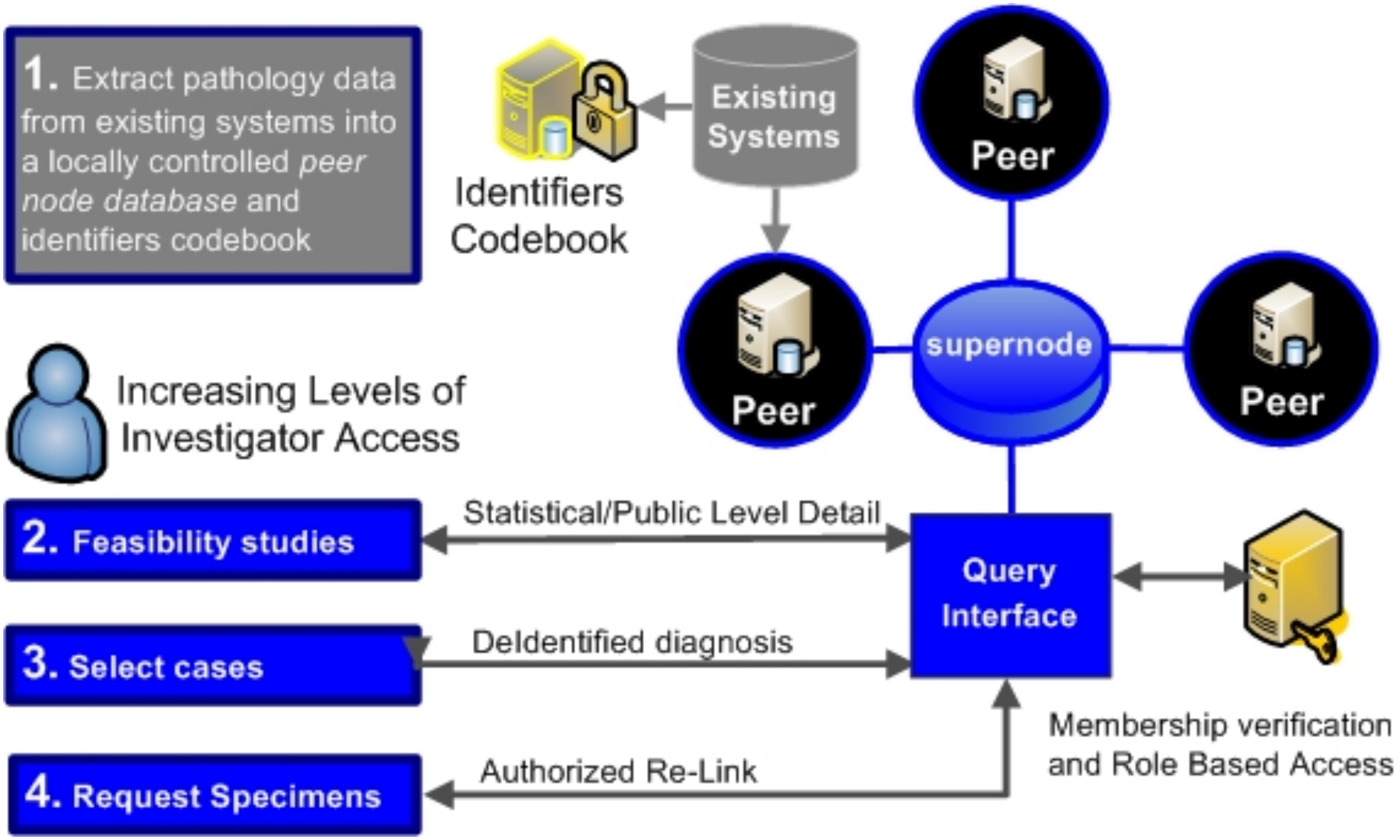
1. **Brigham & Women's Hospital***
2. **Beth Israel Deaconess Medical Center***
3. Cedars-Sinai Medical Center
4. **Dana-Farber Cancer Institute***
5. **Children's Hospital Boston***
6. **Harvard Medical School***
7. **Massachusetts General Hospital***
8. National Institutes of Health
9. National Cancer Institute
10. Olive View Medical Center
11. Regenstrief Institute
12. University of California at Los Angeles Medical Center
13. University of Pittsburgh Medical Center
14. VA Greater LA Healthcare System

*** Participate in ongoing "Virtual Specimen Locator" collaboration**

Sites Participating in National Demonstration



Overview



1. Extract pathology data from existing systems into a locally controlled *peer node database* and identifiers codebook

2. Feasibility studies

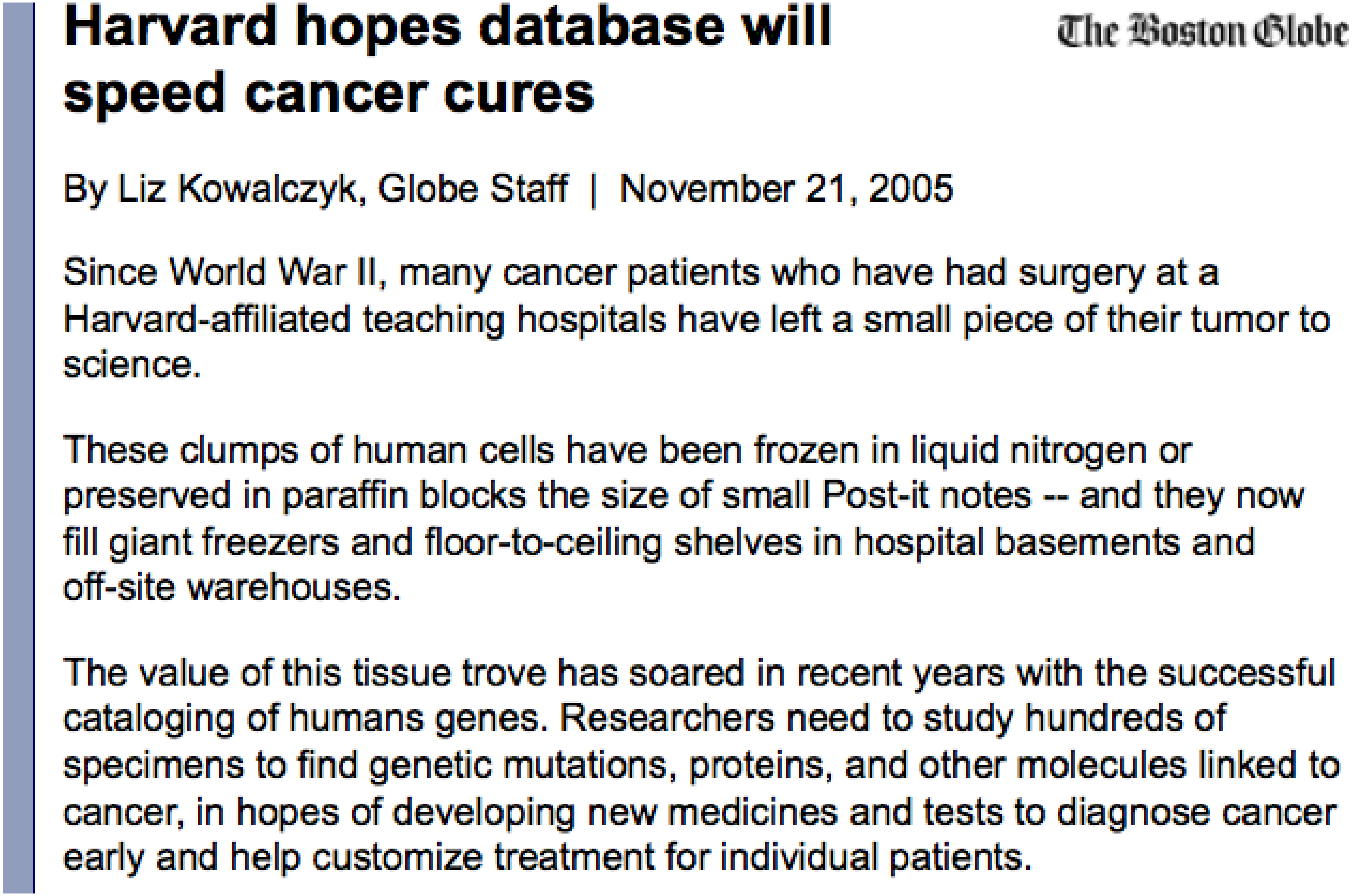
3. Select cases

4. Request Specimens

Sharing Human Tissues for Discovery and Validation

- **Results**
  - National prototype including HMS, UCLA, Indiana, UPMC, …
  - Live Production instance at HMS including 4 hospitals
  - Developed Open Source Tools
  - caBIG adopted caTIES from SPIN
  - Influenced Markle's Common Framework federated query
  - TMA construction using specimens from four sites

SPIN: Sharing Human Tissues for Discovery and Validation



Harvard hopes database will speed cancer cures

The Boston Globe

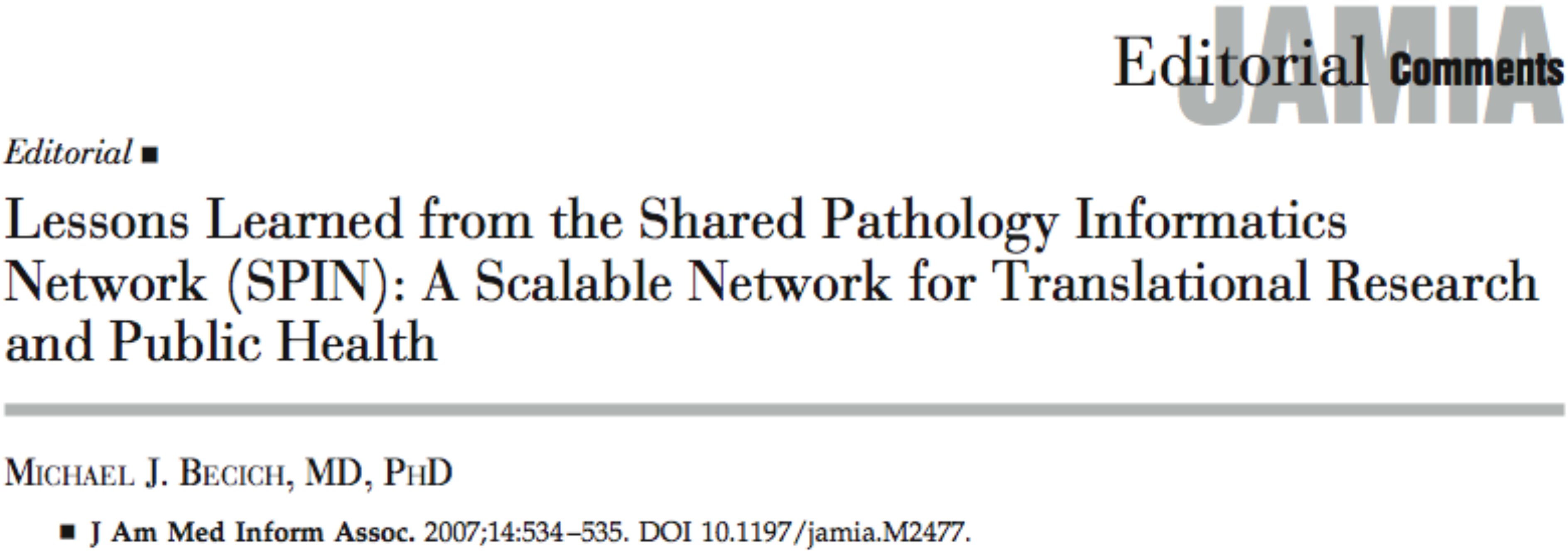
By Liz Kowalczyk, Globe Staff | November 21, 2005

Since World War II, many cancer patients who have had surgery at a Harvard-affiliated teaching hospitals have left a small piece of their tumor to science.

These clumps of human cells have been frozen in liquid nitrogen or preserved in paraffin blocks the size of small Post-it notes -- and they now fill giant freezers and floor-to-ceiling shelves in hospital basements and off-site warehouses.

The value of this tissue trove has soared in recent years with the successful cataloging of humans genes. Researchers need to study hundreds of specimens to find genetic mutations, proteins, and other molecules linked to cancer, in hopes of developing new medicines and tests to diagnose cancer early and help customize treatment for individual patients.

Sharing Human Tissues for Discovery and Validation



JAMIA Editorial Comments

Editorial ■

Lessons Learned from the Shared Pathology Informatics Network (SPIN): A Scalable Network for Translational Research and Public Health

MICHAEL J. BECICH, MD, PHD

■ **J Am Med Inform Assoc.** 2007;14:534–535. DOI 10.1197/jamia.M2477.

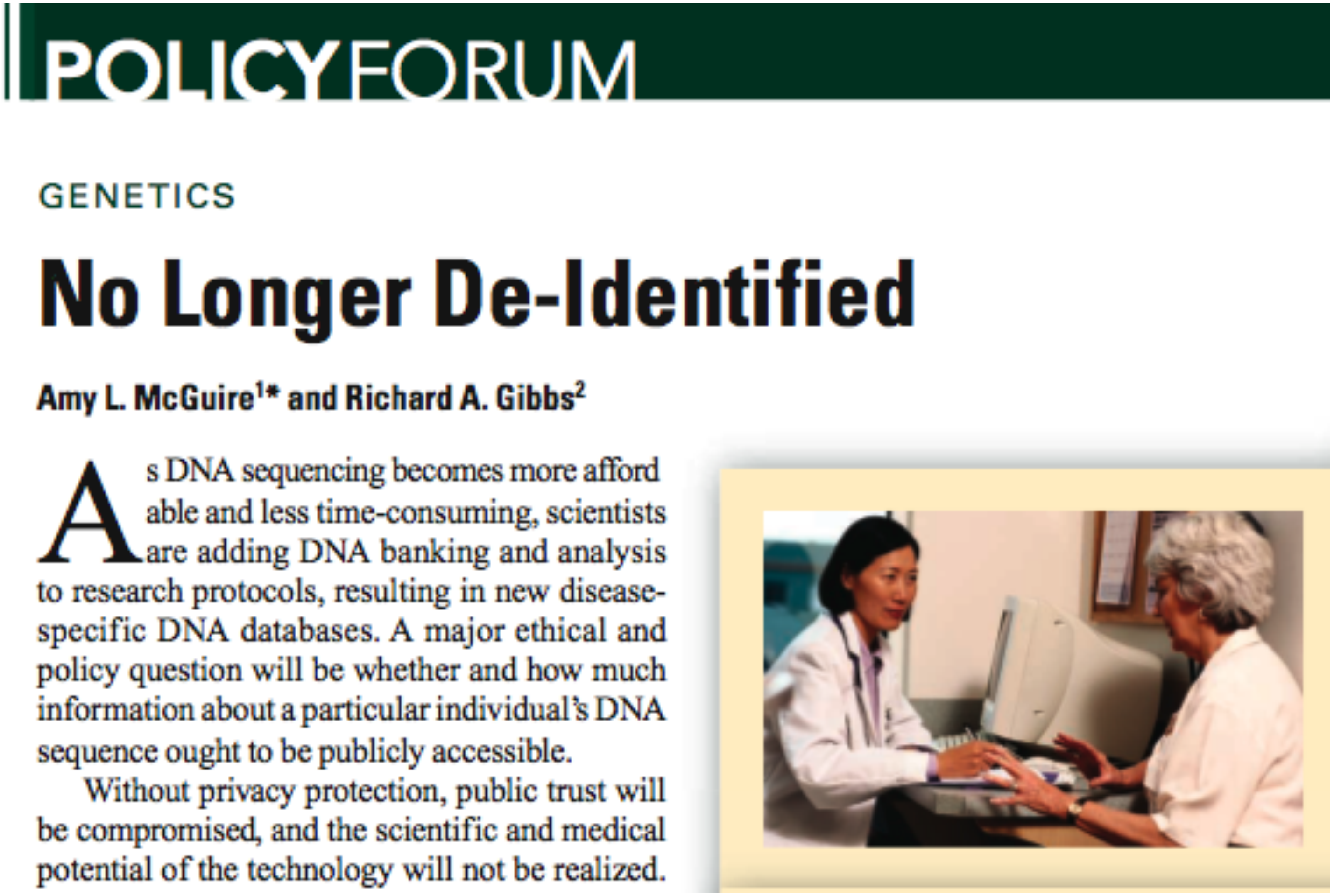
Sharing Genomic Results for Association Studies?

- **Motivation:**
  - Enable Phenotype – Genotype association studies for Autism Spectrum Disorders
  - Integrative genomics across multiple measurement modalities such as DNA->RNA (EQTL)

- **New Challenges:**
  - Privacy Policy: genotypes are clearly identifiable
  - Resources: storage, processing, network load for SNP data
  - Multiple Testing and False Discovery

Sharing Genomic Results for Association Studies?

- **Policy Challenges**



POLICYFORUM

GENETICS

No Longer De-Identified

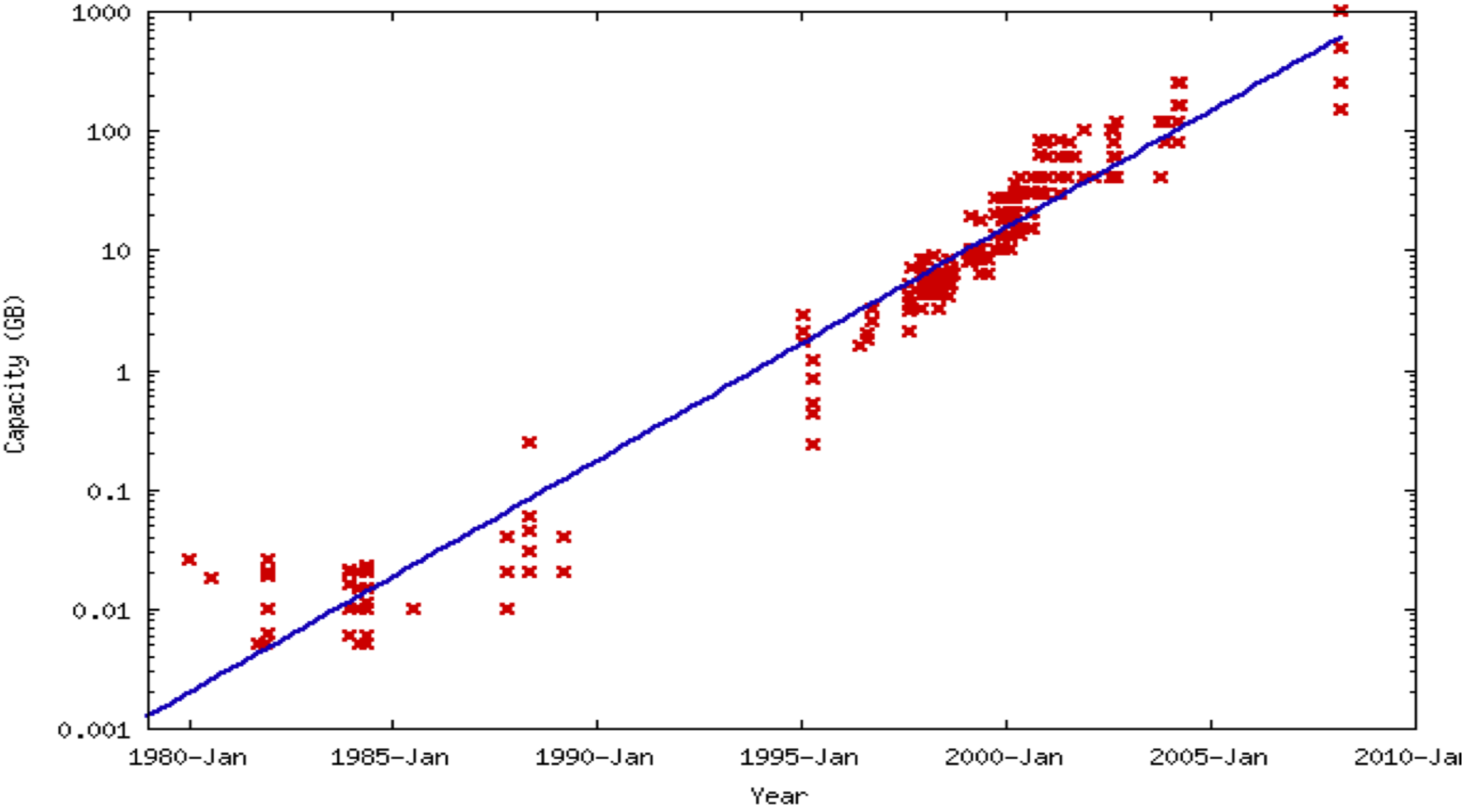
Amy L. McGuire[1]* and Richard A. Gibbs[2]

As DNA sequencing becomes more affordable and less time-consuming, scientists are adding DNA banking and analysis to research protocols, resulting in new disease-specific DNA databases. A major ethical and policy question will be whether and how much information about a particular individual's DNA sequence ought to be publicly accessible.

Without privacy protection, public trust will be compromised, and the scientific and medical potential of the technology will not be realized.

Sharing Genomic Results for Association Studies?

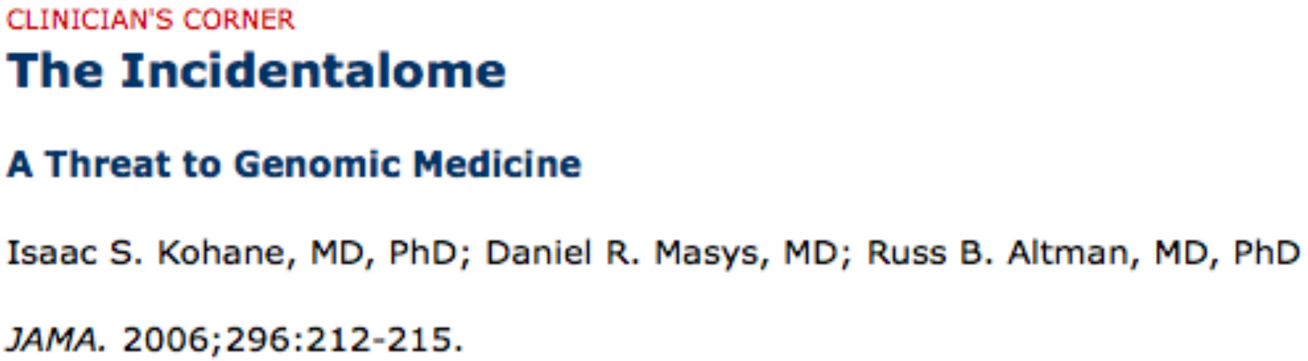
- **Technical Challenges**



http://en.wikipedia.org/wiki/Hard_drive

Sharing Genomic Results for Association Studies?

- **Multiple Testing Challenges**



CLINICIAN'S CORNER

The Incidentalome

A Threat to Genomic Medicine

Isaac S. Kohane, MD, PhD; Daniel R. Masys, MD; Russ B. Altman, MD, PhD

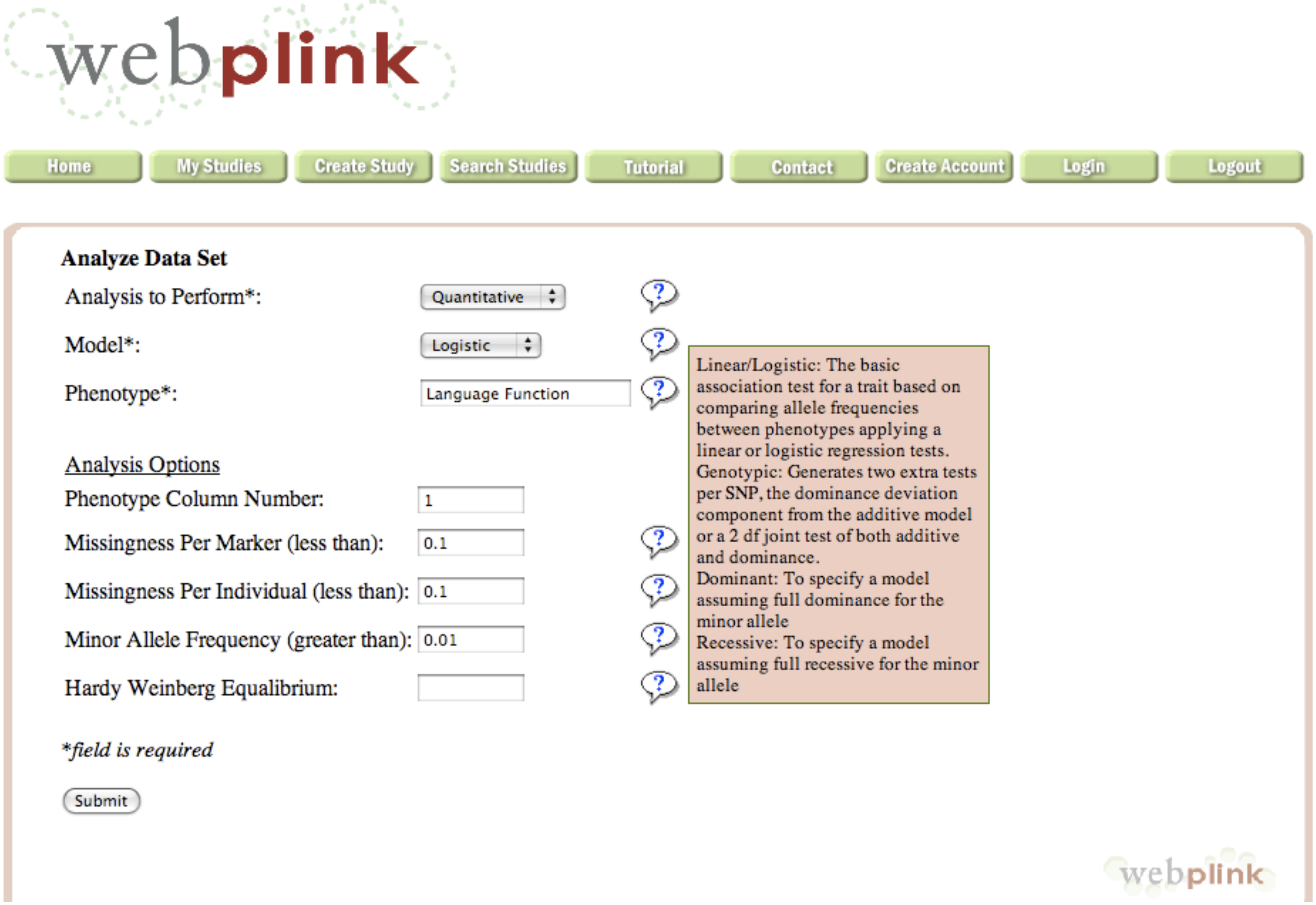
JAMA. 2006;296:212-215.

"In the genomic era, **the lack of prior probabilities** regarding the clinical import of each genetic variant **creates the likelihood** of a large proportion of **false positives**, if genetic testing is not placed on a systematic quantitative basis."

Sharing Genomic Results for Association Studies?

- **Solution: what worked before?**
  - Link genomic test results to the clinical data in a spin peer
  - Protect patient privacy with anonymization and statistical aggregation techniques
  - Engender participation by reasserting local ownership of microarray data and stewardship of patient privacy

Sharing Genomic Results for Association Studies?



Analyze Data Set

Analysis to Perform*:	Quantitative
Model*:	Logistic
Phenotype*:	Language Function

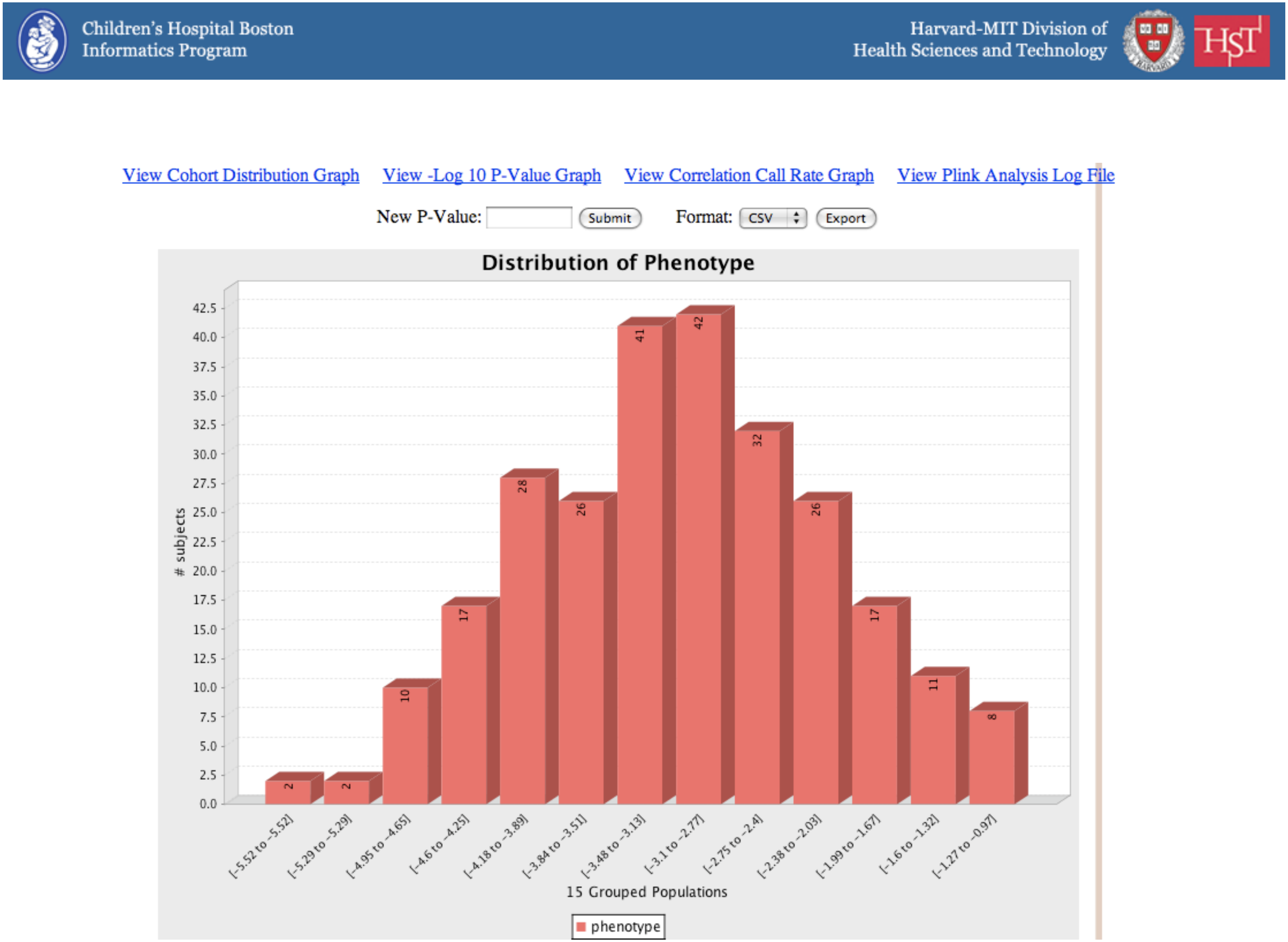
<u>Analysis Options</u>

Phenotype Column Number:	1
Missingness Per Marker (less than):	0.1
Missingness Per Individual (less than):	0.1
Minor Allele Frequency (greater than):	0.01
Hardy Weinberg Equalibrium:	

field is required

Submit

Linear/Logistic: The basic association test for a trait based on comparing allele frequencies between phenotypes applying a linear or logistic regression tests.
Genotypic: Generates two extra tests per SNP, the dominance deviation component from the additive model or a 2 df joint test of both additive and dominance.
Dominant: To specify a model assuming full dominance for the minor allele
Recessive: To specify a model assuming full recessive for the minor allele

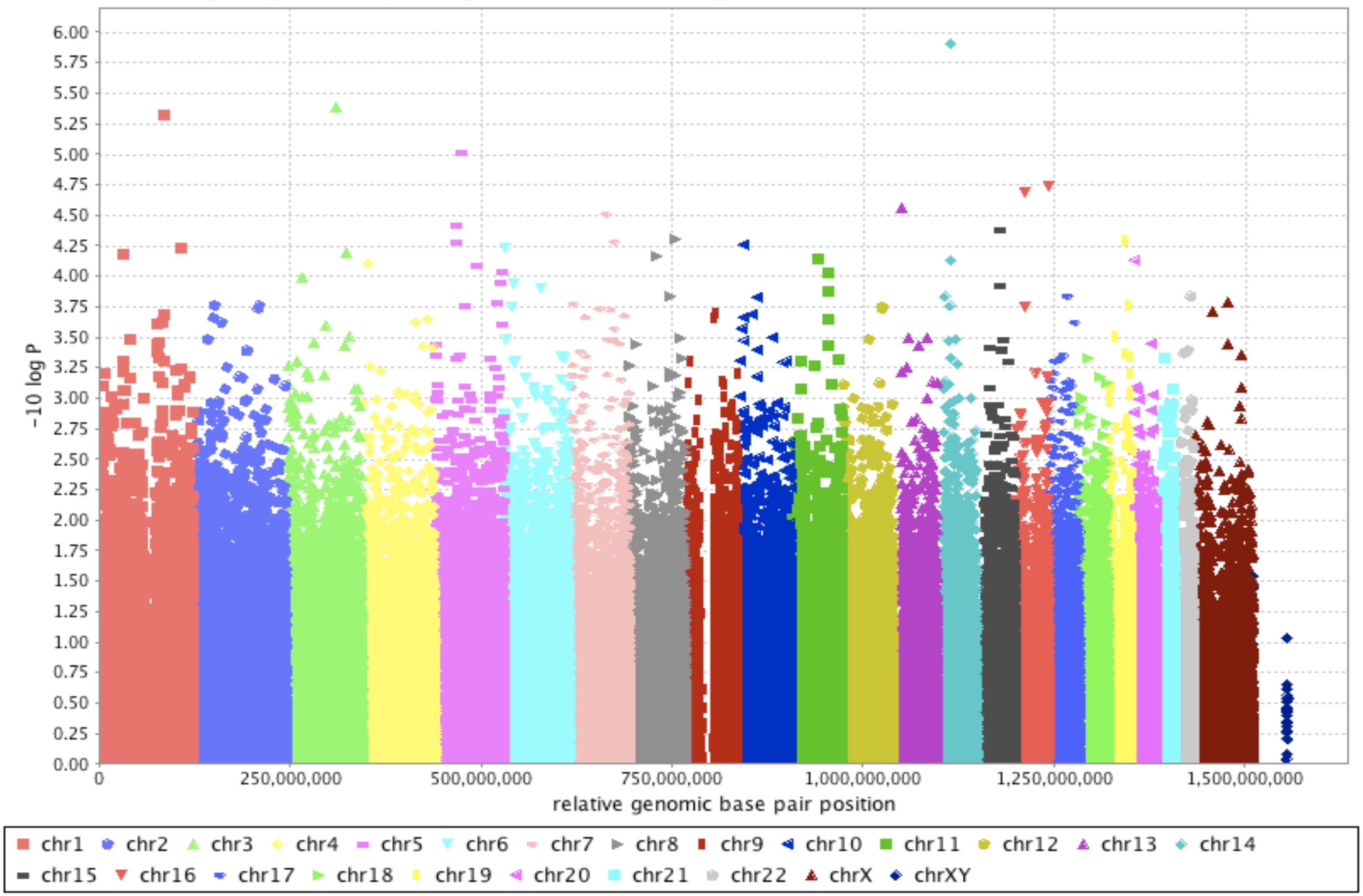


Children's Hospital Boston
Informatics Program

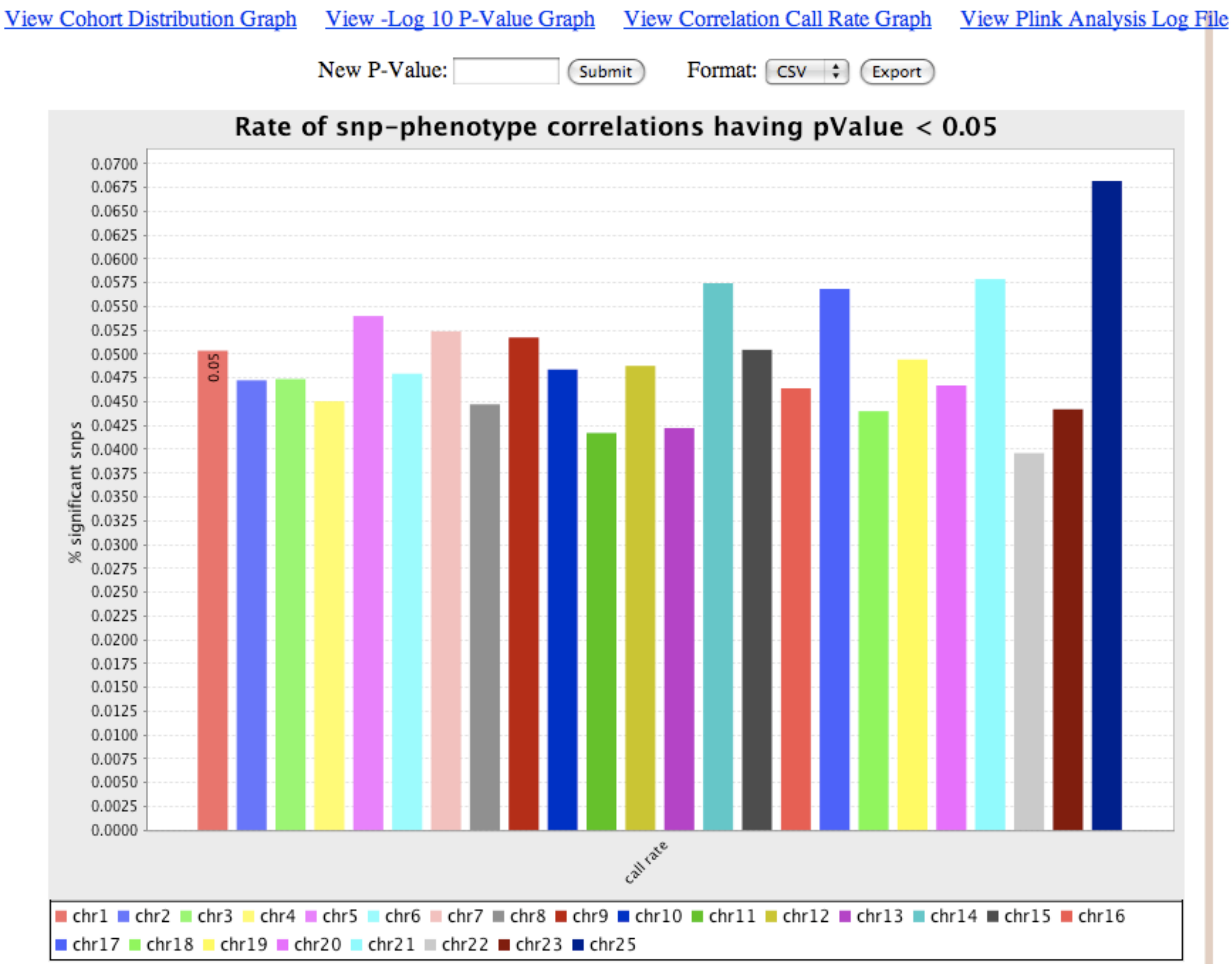
Harvard-MIT Division of
Health Sciences and Technology

View Cohort Distribution Graph
View -Log 10 P-Value Graph
View Correlation Call Rate Graph
View Plink Analysis Log File
New P-Value:
Submit
Format:
CSV
Export

Distribution of Phenotype



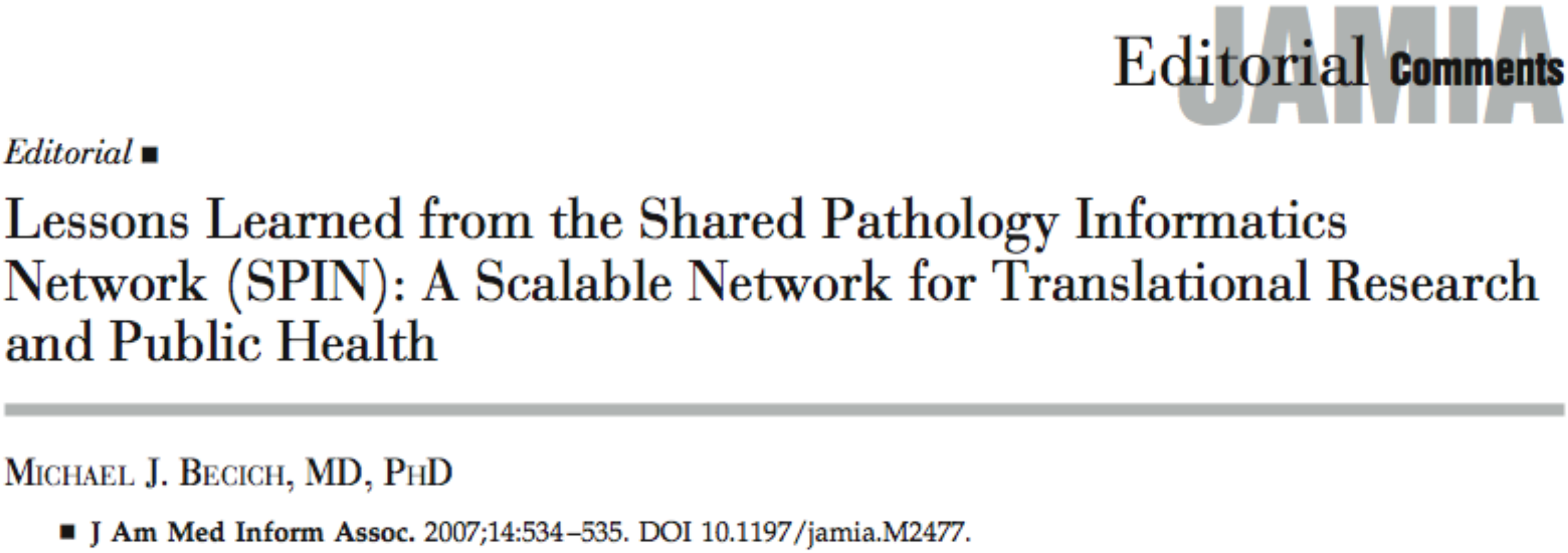
[7]Language Delay : logistic recessive,geno<.1,mind<.1,maf<.01,hwe<.05



View Cohort Distribution Graph View -Log 10 P-Value Graph View Correlation Call Rate Graph View Plink Analysis Log File

New P-Value: Submit Format: CSV Export

Applying lessons learned:
Common Architecture



JAMIA Editorial Comments

Editorial ■

Lessons Learned from the Shared Pathology Informatics Network (SPIN): A Scalable Network for Translational Research and Public Health

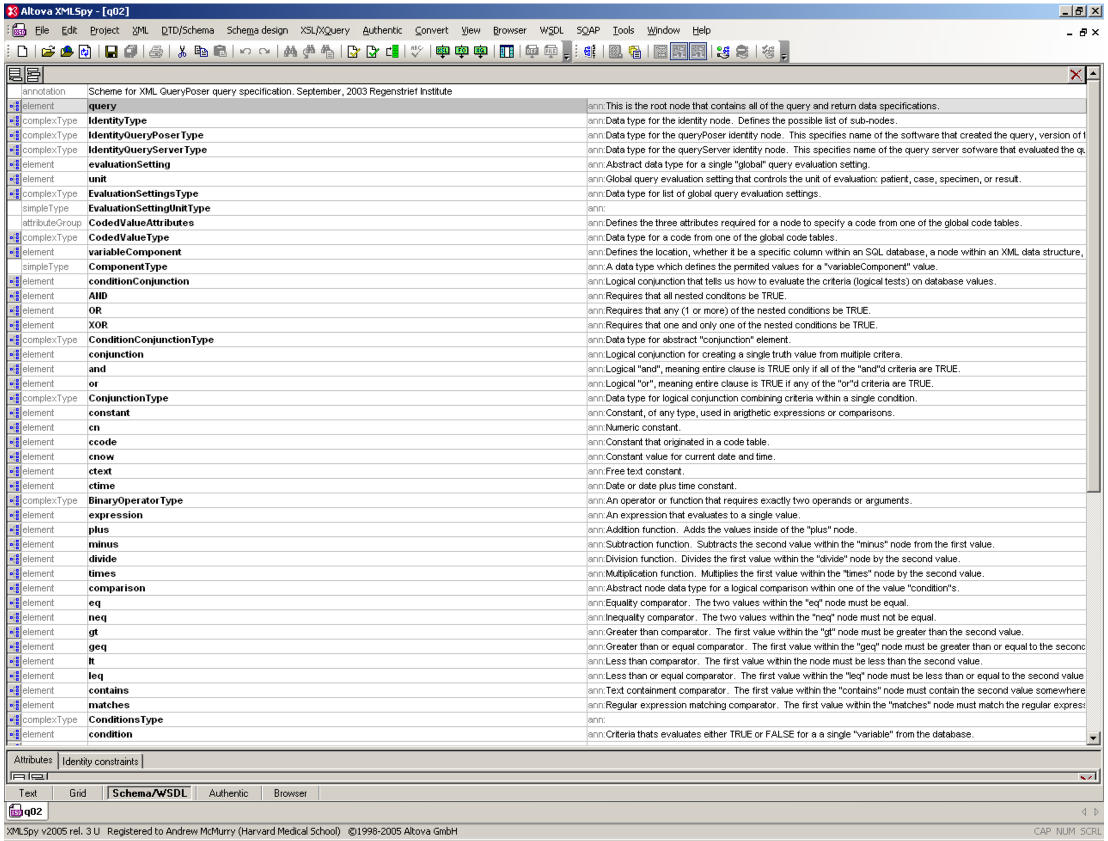
Michael J. Becich, MD, PhD

■ **J Am Med Inform Assoc.** 2007;14:534–535. DOI 10.1197/jamia.M2477.

Lessons Learned: IRBs and political will

- Statistical level queries easy are OK by IRBs
- Difficulty arises going to the next step
  - HIPAA limited data set
  - PHI
- ANY use of patient data for research imposes SOME risk
- Minimize risk, show that research benefit is overwhelmingly in the best interest of patients

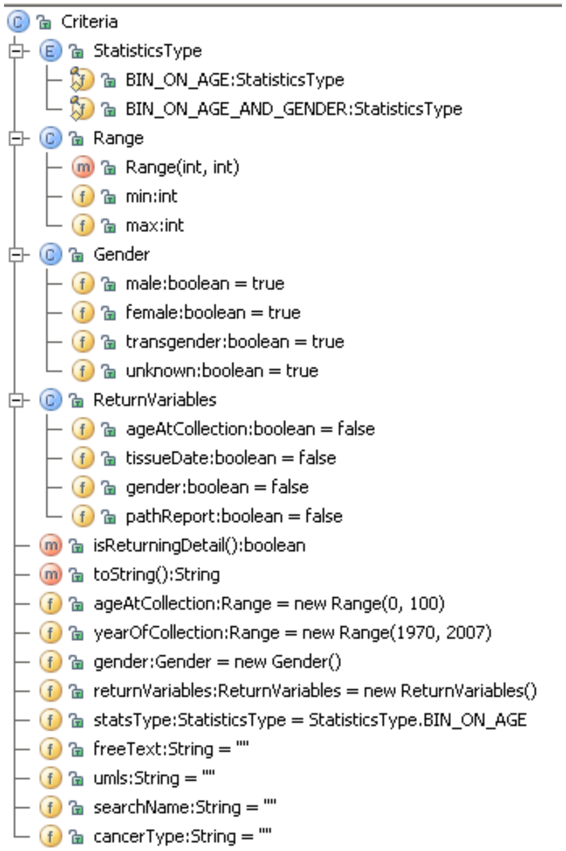
Lessons Learned: mapping heterogeneous DBs


Altova XMLSpy - [q02]
File Edit Project XML DTD/Schema Schema design XSL/XQuery Authentic Convert View Browser WSDL SOAP Tools Window Help
Attributes
Identity constraints
Text
Grid
Schema/WSDL
Authentic
Browser
q02
CAP NUM SCRL

annotation	Scheme for XML QueryPoser query specification. September, 2003 Regenstrief Institute	
element	**query**	ann:This is the root node that contains all of the query and return data specifications.
complexType	**IdentityType**	ann:Data type for the identity node. Defines the possible list of sub-nodes.
complexType	**IdentityQueryPoserType**	ann:Data type for the queryPoser identity node. This specifies name of the software that created the query, version of t
complexType	**IdentityQueryServerType**	ann:Data type for the queryServer identity node. This specifies name of the query server sofware that evaluated the qu
element	**evaluationSetting**	ann:Abstract data type for a single "global" query evaluation setting.
element	**unit**	ann:Global query evaluation setting that controls the unit of evaluation: patient, case, specimen, or result.
complexType	**EvaluationSettingsType**	ann:Data type for list of global query evaluation settings.
simpleType	**EvaluationSettingUnitType**	ann:
attributeGroup	**CodedValueAttributes**	ann:Defines the three attributes required for a node to specify a code from one of the global code tables.
complexType	**CodedValueType**	ann:Data type for a code from one of the global code tables.
element	**variableComponent**	ann:Defines the location, whether it be a specific column within an SQL database, a node within an XML data structure,
simpleType	**ComponentType**	ann:A data type which defines the permited values for a "variableComponent" value.
element	**conditionConjunction**	ann:Logical conjunction that tells us how to evaluate the criteria (logical tests) on database values.
element	**AND**	ann:Requires that all nested conditons be TRUE.
element	**OR**	ann:Requires that any (1 or more) of the nested conditions be TRUE.
element	**XOR**	ann:Requires that one and only one of the nested conditions be TRUE.
complexType	**ConditionConjunctionType**	ann:Data type for abstract "conjunction" element.
element	**conjunction**	ann:Logical conjunction for creating a single truth value from muitiple critera.
element	**and**	ann:Logical "and", meaning entire clause is TRUE only if all of the "and"d criteria are TRUE.
element	**or**	ann:Logical "or", meaning entire clause is TRUE if any of the "or"d criteria are TRUE.
complexType	**ConjunctionType**	ann:Data type for logical conjunction combining criteria within a single condition.
element	**constant**	ann:Constant, of any type, used in arigthetic expressions or comparisons.
element	**cn**	ann:Numeric constant.
element	**ccode**	ann:Constant that originated in a code table.
element	**cnow**	ann:Constant value for current date and time.
element	**ctext**	ann:Free text constant.
element	**ctime**	ann:Date or date plus time constant.
complexType	**BinaryOperatorType**	ann:An operator or function that requires exactly two operands or arguments.
element	**expression**	ann:An expression that evaluates to a single value.
element	**plus**	ann:Addition function. Adds the values inside of the "plus" node.
element	**minus**	ann:Subtraction function. Subtracts the second value within the "minus" node from the first value.
element	**divide**	ann:Division function. Divides the first value within the "divide" node by the second value.
element	**times**	ann:Multiplication function. Multiplies the first value within the "times" node by the second value.
element	**comparison**	ann:Abstract node data type for a logical comparison within one of the value "condition"s.
element	**eq**	ann:Equality comparator. The two values within the "eq" node must be equal.
element	**neq**	ann:Inequality comparator. The two values within the "neq" node must not be equal.
element	**gt**	ann:Greater than comparator. The first value within the "gt" node must be greater than the second value.
element	**geq**	ann:Greater than or equal comparator. The first value within the "geq" node must be greater than or equal to the seconc
element	**lt**	ann:Less than comparator. The first value within the node must be less than the second value.
element	**leq**	ann:Less than or equal comparator. The first value within the "leq" node must be less than or equal to the second value
element	**contains**	ann:Text containment comparator. The first value within the "contains" node must contain the second value somewhere
element	**matches**	ann:Regular expression matching comparator. The first value within the "matches" node must match the regular express
complexType	**ConditionsType**	ann:
element	**condition**	ann:Criteria thats evaluates either TRUE or FALSE for a a single "variable" from the database.

XMLSpy v2005 rel. 3 U Registered to Andrew McMurry (Harvard Medical School) ©1998-2005 Altova GmbH

VS



```
Criteria
  StatisticsType
    BIN_ON_AGE:StatisticsType
    BIN_ON_AGE_AND_GENDER:StatisticsType
  Range
    Range(int, int)
    min:int
    max:int
  Gender
    male:boolean = true
    female:boolean = true
    transgender:boolean = true
    unknown:boolean = true
  ReturnVariables
    ageAtCollection:boolean = false
    tissueDate:boolean = false
    gender:boolean = false
    pathReport:boolean = false
  isReturningDetail():boolean
  toString():String
  ageAtCollection:Range = new Range(0, 100)
  yearOfCollection:Range = new Range(1970, 2007)
  gender:Gender = new Gender()
  returnVariables:ReturnVariables = new ReturnVariables()
  statsType:StatisticsType = StatisticsType.BIN_ON_AGE
  freeText:String = ""
  umls:String = ""
  searchName:String = ""
  cancerType:String = ""
```

Start SMALL : Grow the number of common terms!

Lessons Learned: mapping heterogeneous DBs

1. Request for Capabilities & Statistics (What is available?)

2. Availability limits scope of the vocabulary

3. Which BIG questions can be asked with only a few identifiers?

Pathology:	age, gender, collection,	free text "diagnosis"
Public Health:	age, gender, location,	free text "complaint"
CTSA:	age, gender, ...,	free text mining

4. Parallel tracks: autocoding and standard vocabulary approach
   - Different low hanging fruit: diagnosis *vs* MRN

5. Quick End-To-End lifecyles
   - Question, development, research, new question

Summary

Addressed 3 pervasive issues:

- Linking routine care systems for robust research
- Protecting patient privacy
- Engendering participation among hospitals

Use Cases

- Routinely collected human tissues for biomarker discovery and high-throughput validation

- Genomic measurements derived from tissue sharing

Collaborators & Acknowledgements

- **Biospecimen Sharing Community**
  - Too many to list!
  - http://spin.chip.org/community.html

- **Public Health Surveillance**
  - http://chip.org/ihl

- **ASD Genotype Phenotype Associations**
  - Developers: Mike Banos , Gregory Polumbo
  - Investigators: Alexa McCray , Dennis Wall, Amanda Sedgewick
  - Collaborator: Shaun Purcell (plink author)

- **Special Thanks**
  - Advisors: Zak Kohane & Ken Mandl
  - Investigators: Kamila Naxerova & Alal Eran