Clinical Data Mining

"If we knew what it was we were doing, it wouldn’t be called research."
- Albert Einstein

Dr. Keith J. Dreyer
Partners HealthCare System ~ Massachusetts General Hospital
Harvard Medical School

Clinical Data Mining
Goals

- Query for Clinical Conditions
- Extract Related Information
- Knowledge Discovery

Clinical Data Mining
Challenges

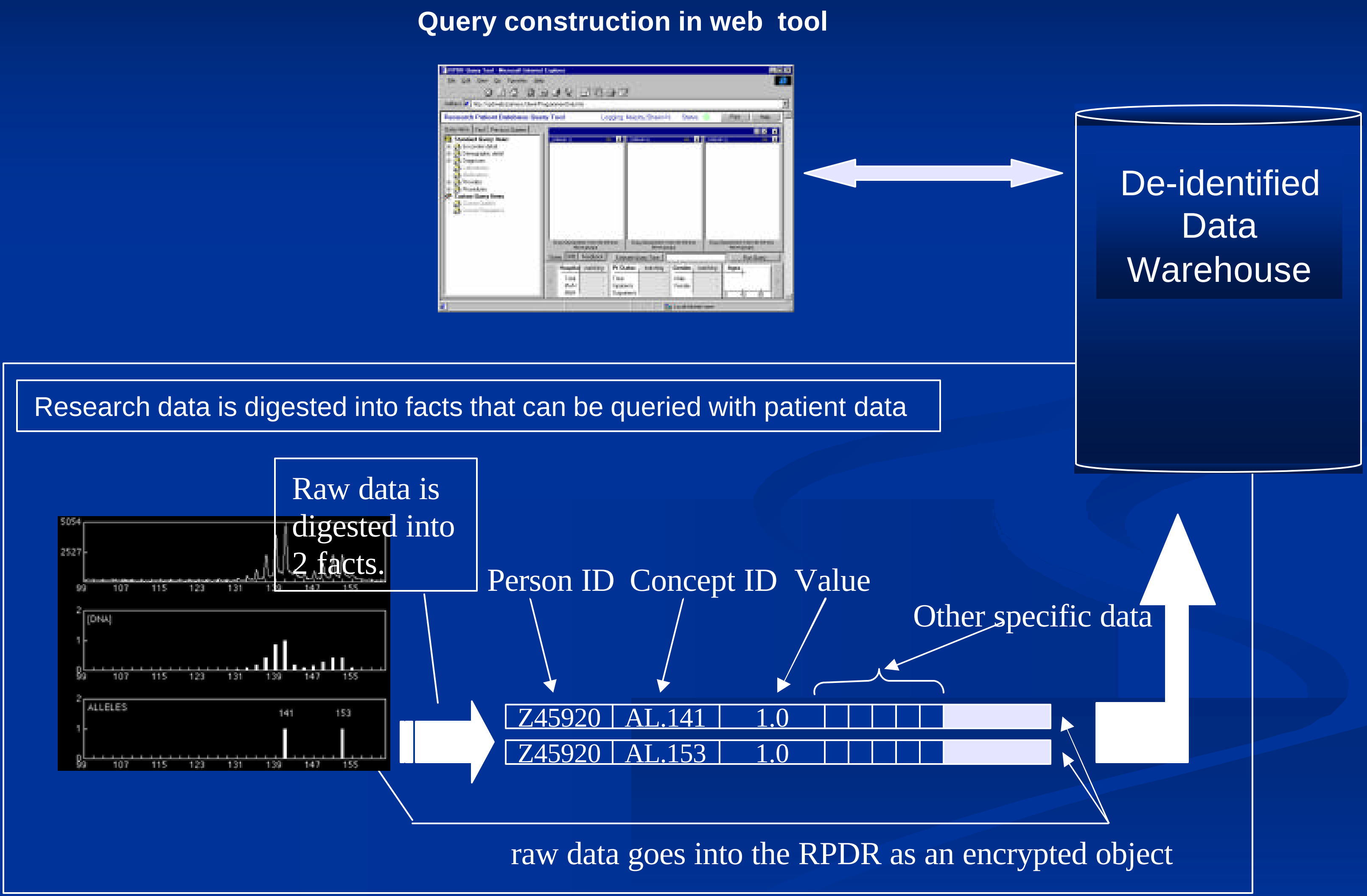
- Various data sources
- Limited Structure
- Imbedded Ontology

Clinical Data Mining Processes

- Digital format for all pertinent data
- Create structure
  - Obtain coded information
  - Natural Language Understanding
- Create a widely accessible repository

Query for Clinical Conditions

Research Patient Data Registry



Query construction in web tool

De-identified Data Warehouse

Research data is digested into facts that can be queried with patient data

Raw data is digested into 2 facts.

Other specific data

Person ID	Concept ID	Value
Z45920	AL.141	1.0
Z45920	AL.153	1.0

raw data goes into the RPDR as an encrypted object

Research Patient Data Repository

- 1.8 million MGH & BWH patients
- 480 million diagnoses, medications, procedures, laboratories, with patient demographic & visit info
- Authorized use by 750 faculty status users.
- Researcher can construct complex independent queries
- Pilot of radiology image display
- Currently no radiology query capability

Integrating Radiology Queries

- Radiology Reports
- Radiology Images

- Challenge ?   No Structure!

Creating Structure

- 2 million patient demographics
- 6 million radiology interpretations
- 150 million radiology images

Current Data Load: 100 TB

LEXIMER

Lexicon Mediated Entropy Reduction

- Radiology Language Understanding
- Lexicon Based Hierarchical Decision Trees
  - Trained to remove report noise, thus retain signal
    - Where Signal = F_T and R_T (defined through trained lexicon)
  - Lexicon trained by heuristics for CT and MRI
  - Expanded to all of radiology by modality training sets
    - Training stop point - 95% accuracy
  - Currently the HDTs contain 4,132 nodes

LEXIMER

Create an automated method to detect:

F_T - The observation of a positive finding

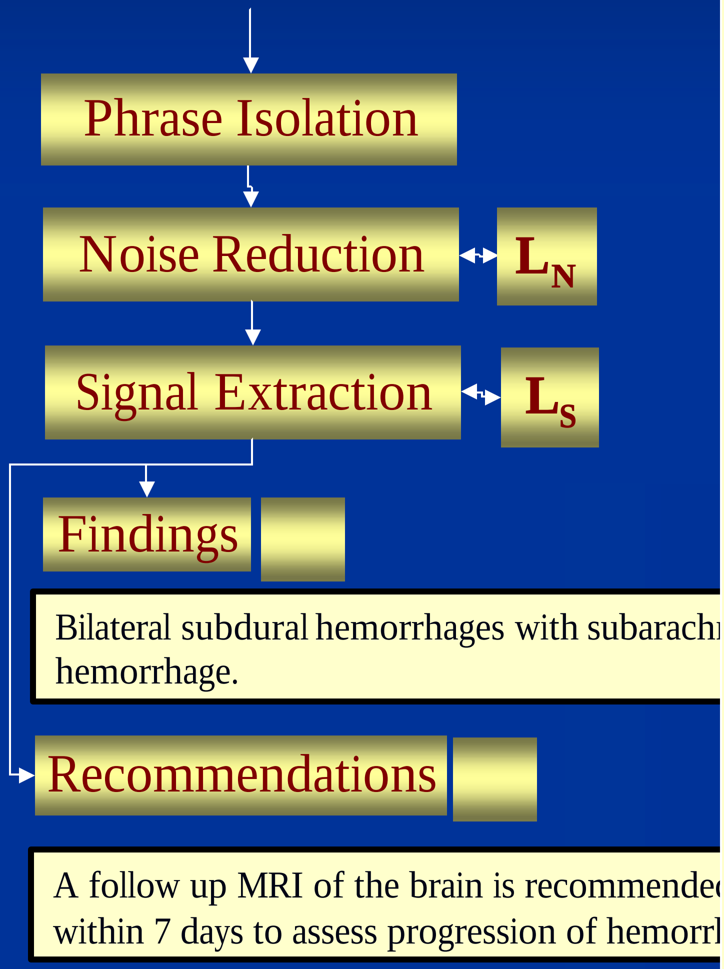
R_T - A recommendation that further imaging be performed

Yield = F_T Exams / Total Exams

Recommendation Rate = R_T Exams / Total Exams

Radiology Report Understanding

~ LEXIMER ~



Bilateral subdural hemorrhages with subarach hemorrhage.

A follow up MRI of the brain is recommende within 7 days to assess progression of hemorr

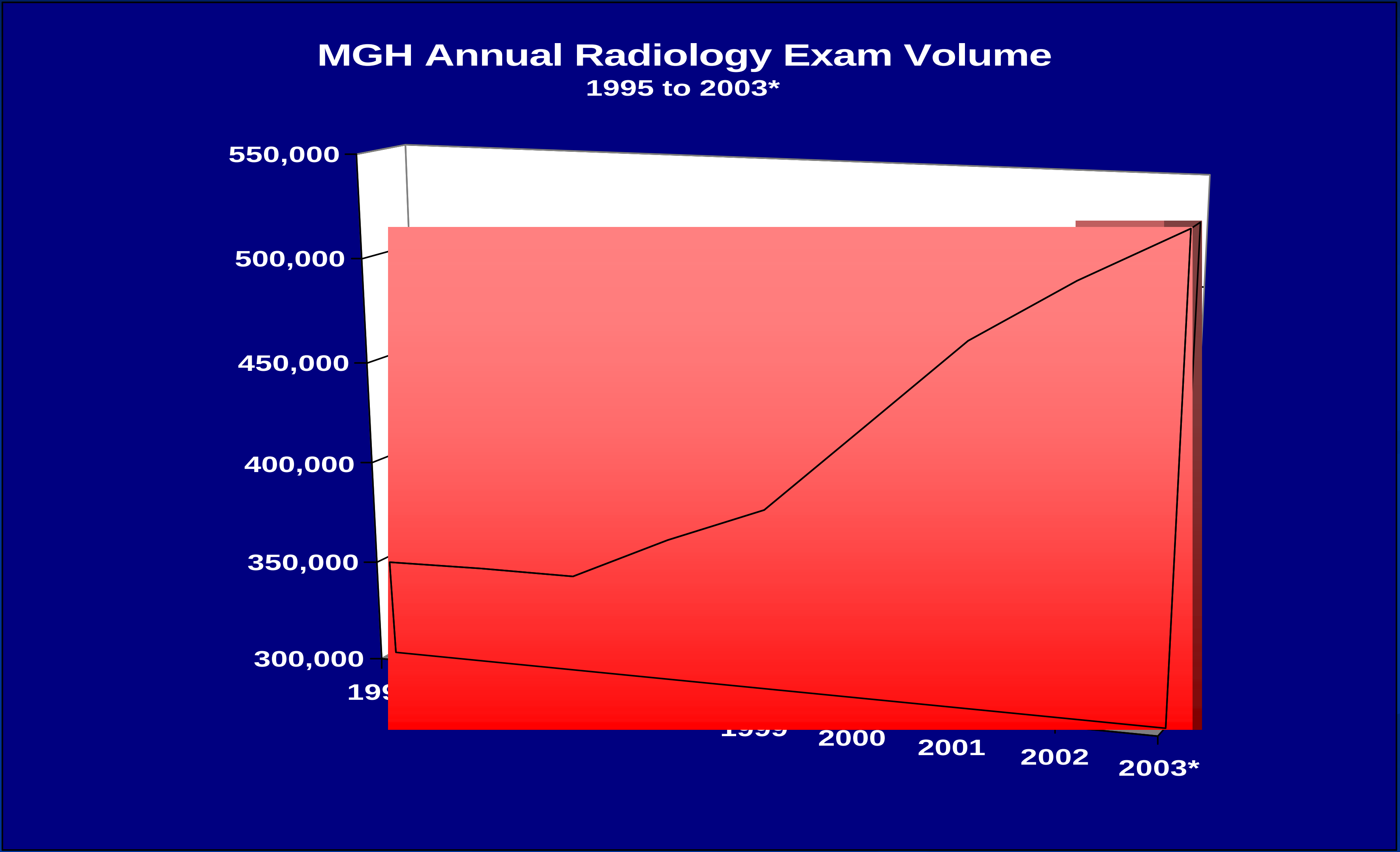
This study is reviewed with Dr Smith. Standard protocol was used to obtain an MRI of the brain with MRA of the circle of Willis and DWI imaging.

Dizziness and recurrent syncope. Please evaluate the posterior circulation. Comparison is to a CT of the head performed 3 September 99. Comparison is also to a CT performed the day after the MRI on 5 September 1909. Bilateral subdural hemorrhages are present. The right sided subdural hemorrhage appears improved when compared to the prior CT. It has a component extending further posteriorly than appreciated on the CT, appearing to involve the occipital lobe on the right side. The left subdural hemorrhage is worse than it appeared on the initial CT. There is extensive subarachnoid hemorrhage better appreciated on MRI than on CT.

There is no evidence of tentorial subdural hematoma. The subsequent CT did show such a bleed, this must have occurred in the interval between studies. DWI imaging of the brain parenchyma is normal in appearance. There is no evidence of acute infarction. The circle of Willis was imaged with particular attention to the posterior circulation. The right vertebral artery appears prominent. The procedure circulation appears entirely normal. Because imaging was centered on the procedure circulation, the MCA's are not completely evaluated. The ventricular system and CSF spaces do not show evidence of abnormal dilation. The visualized extracranial structures are normal in appearance.

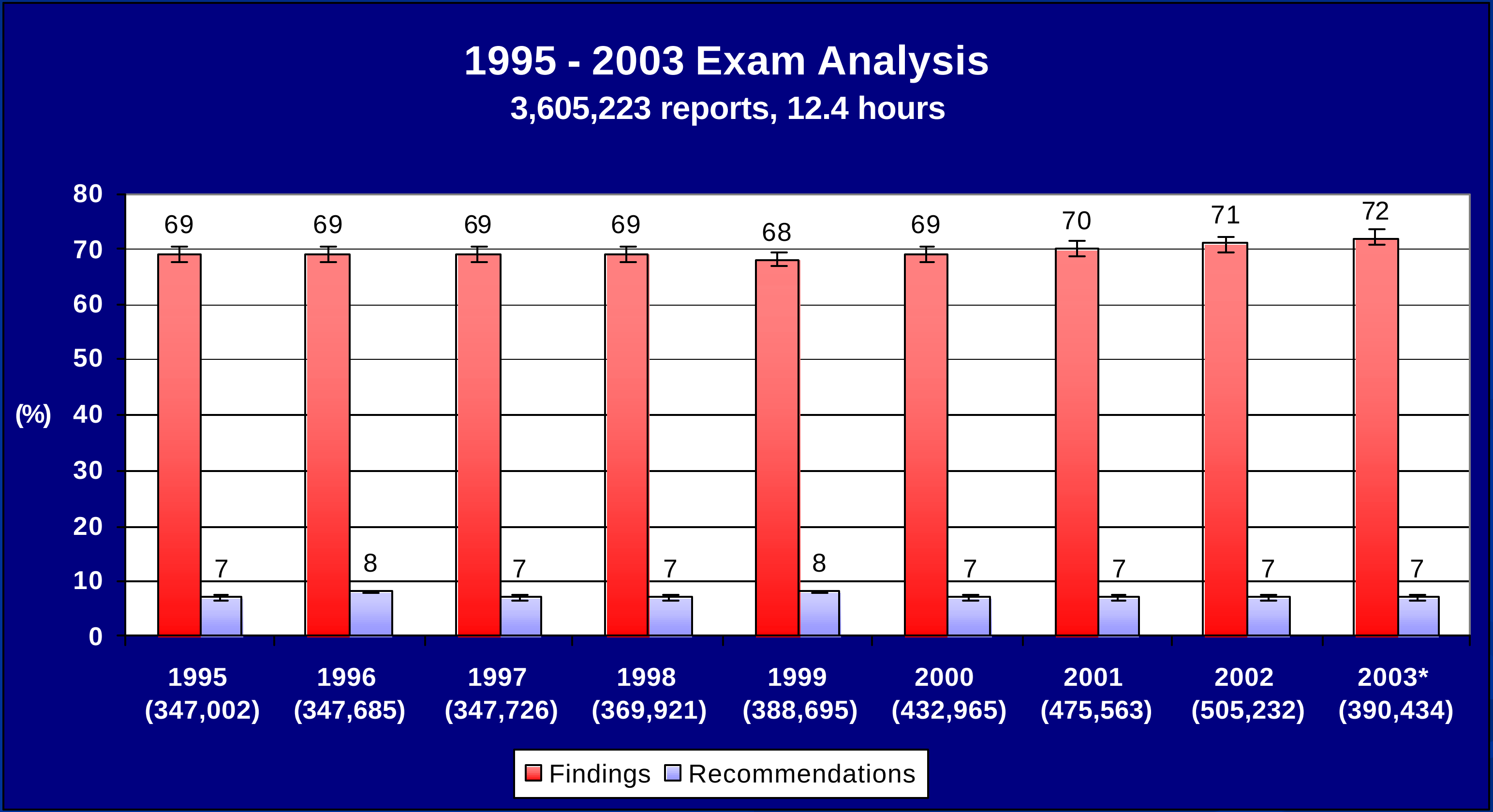
Impression. No evidence of acute infarction on diffusion weighted imaging. Bilateral subdural hemorrhages with subarachnoid hemorrhage. The posterior circulation appears entirely normal. A follow up MRI of the brain is recommended within 7 days to assess progression of hemorrhage.

Motivation

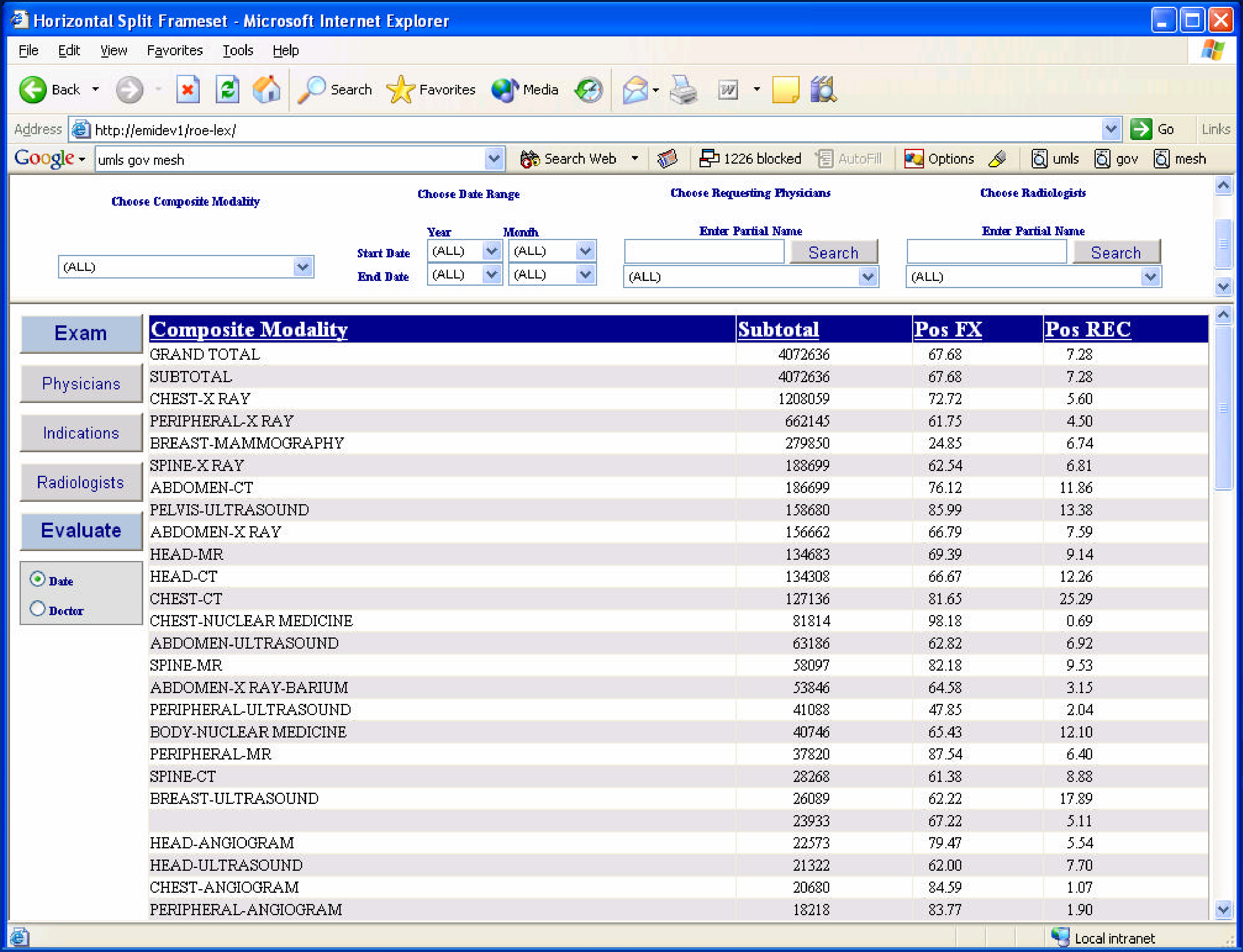

MGH Annual Radiology Exam Volume
1995 to 2003*
550,000
500,000
450,000
400,000
350,000
300,000
199
1999
2000
2001
2002
2003*

Applications

Annual Trends



Modality Analysis



Address: http://emidev1/roe-lex/

Google: umls gov mesh

Choose Composite Modality: (ALL)

Choose Date Range

	Year	Month
Start Date	(ALL)	(ALL)
End Date	(ALL)	(ALL)

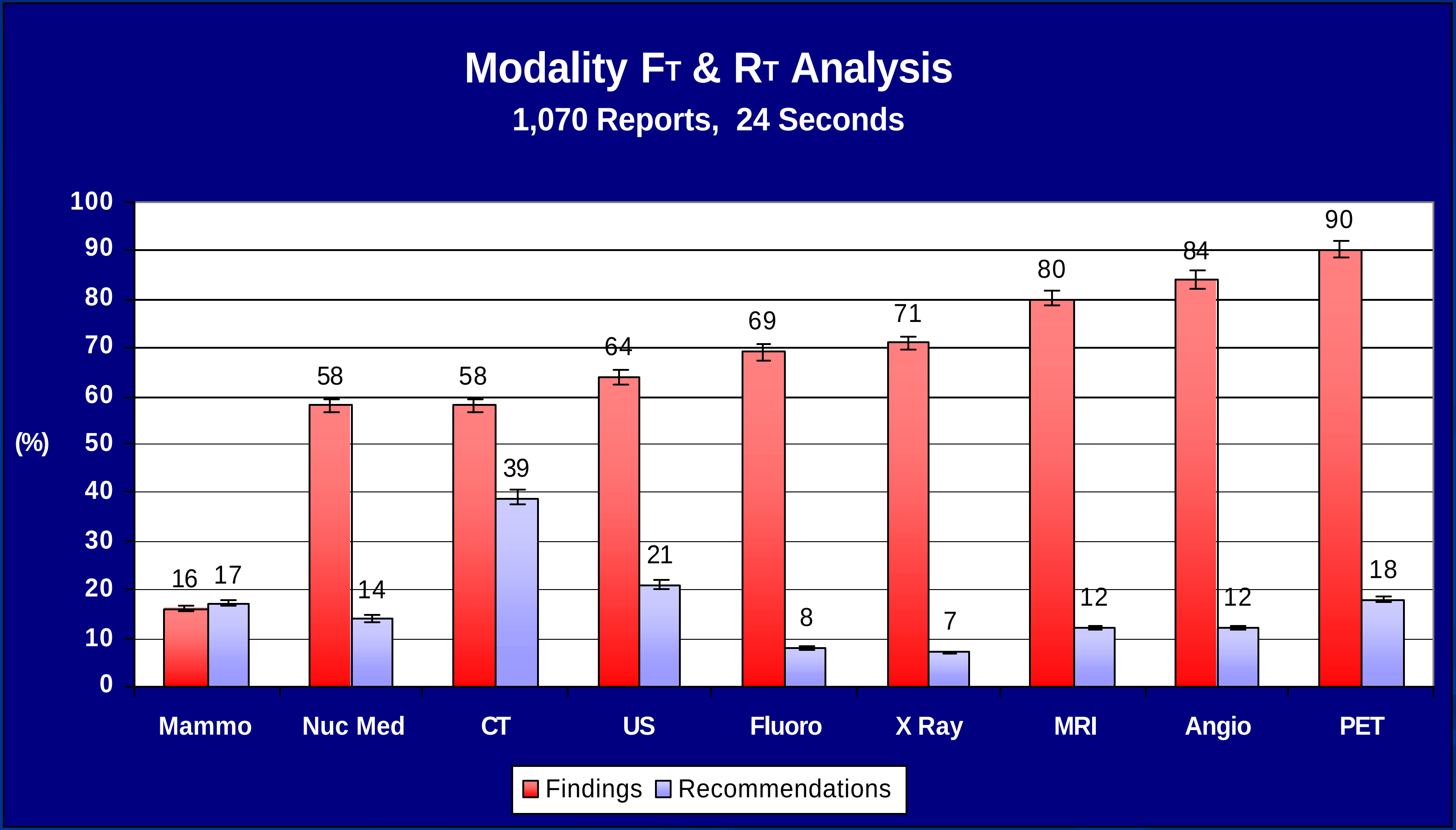
Choose Requesting Physicians — Enter Partial Name: (ALL)

Choose Radiologists — Enter Partial Name: (ALL)

● Date
○ Doctor

Composite Modality	Subtotal	Pos FX	Pos REC
GRAND TOTAL	4072636	67.68	7.28
SUBTOTAL	4072636	67.68	7.28
CHEST-X RAY	1208059	72.72	5.60
PERIPHERAL-X RAY	662145	61.75	4.50
BREAST-MAMMOGRAPHY	279850	24.85	6.74
SPINE-X RAY	188699	62.54	6.81
ABDOMEN-CT	186699	76.12	11.86
PELVIS-ULTRASOUND	158680	85.99	13.38
ABDOMEN-X RAY	156662	66.79	7.59
HEAD-MR	134683	69.39	9.14
HEAD-CT	134308	66.67	12.26
CHEST-CT	127136	81.65	25.29
CHEST-NUCLEAR MEDICINE	81814	98.18	0.69
ABDOMEN-ULTRASOUND	63186	62.82	6.92
SPINE-MR	58097	82.18	9.53
ABDOMEN-X RAY-BARIUM	53846	64.58	3.15
PERIPHERAL-ULTRASOUND	41088	47.85	2.04
BODY-NUCLEAR MEDICINE	40746	65.43	12.10
PERIPHERAL-MR	37820	87.54	6.40
SPINE-CT	28268	61.38	8.88
BREAST-ULTRASOUND	26089	62.22	17.89
	23933	67.22	5.11
HEAD-ANGIOGRAM	22573	79.47	5.54
HEAD-ULTRASOUND	21322	62.00	7.70
CHEST-ANGIOGRAM	20680	84.59	1.07
PERIPHERAL-ANGIOGRAM	18218	83.77	1.90

Applications
Modality Analysis



Modality FT & RT Analysis

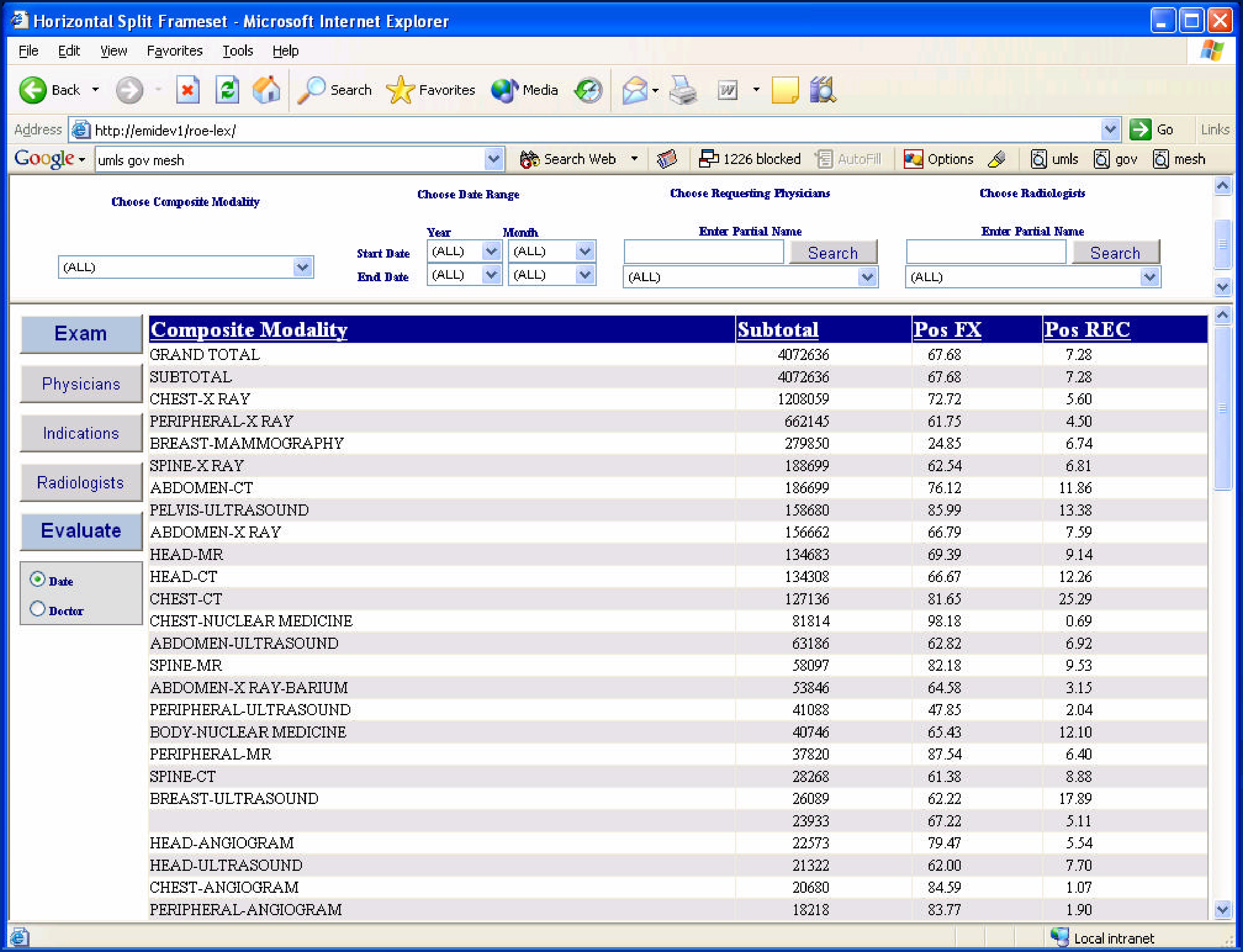
1,070 Reports, 24 Seconds

Extraction of Related Information

Extract Practice Pattern Inconsistencies

- Ordering practices and yield of all referring physicians
- Recommendation practices of all interpreting radiologists
- Provide online access to individual and group level statistics

Practice Pattern Analysis



Address: http://emidev1/roe-lex/

Google: umls gov mesh

Choose Composite Modality: (ALL)

Choose Date Range

	Year	Month
Start Date	(ALL)	(ALL)
End Date	(ALL)	(ALL)

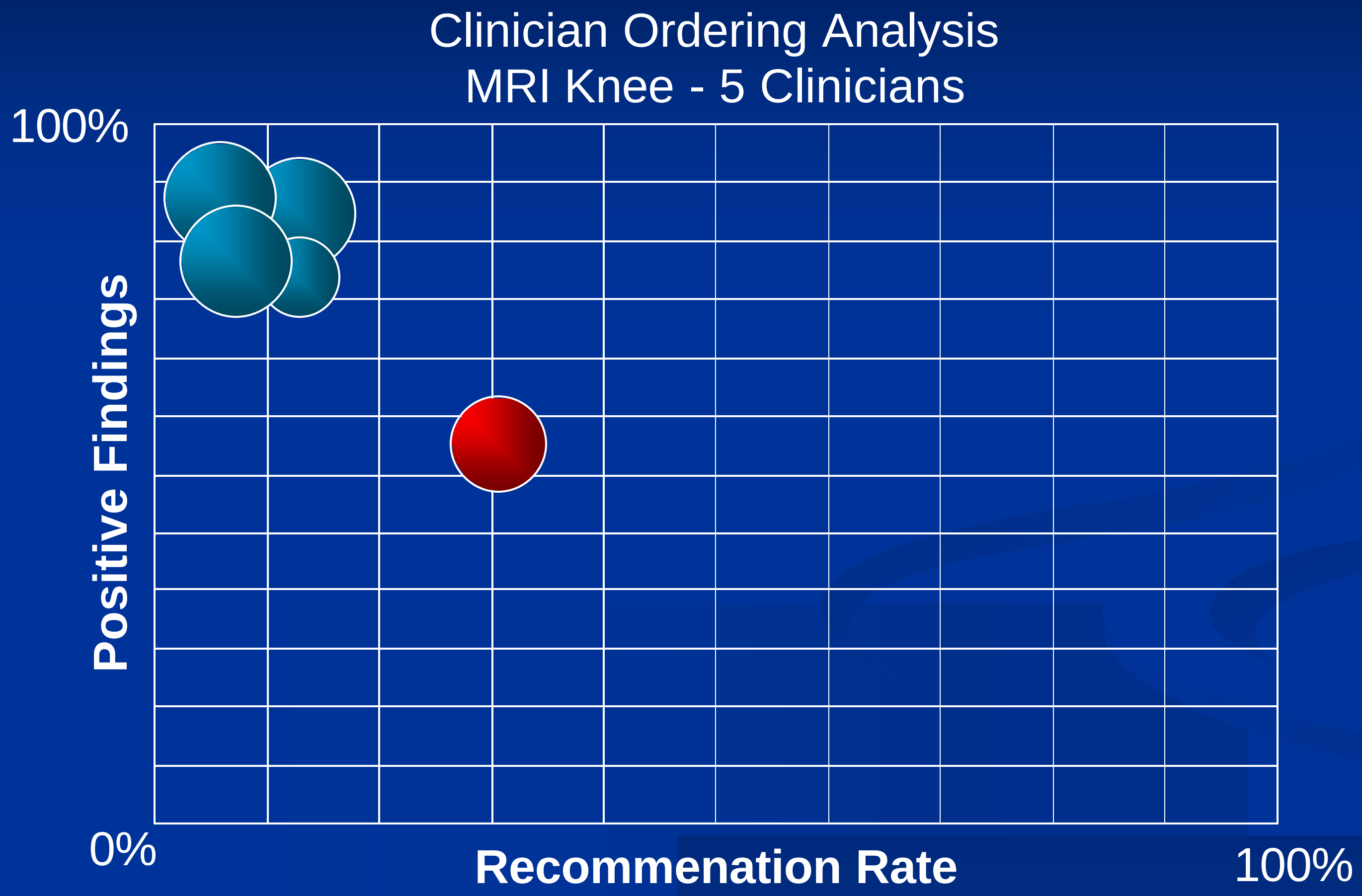
Choose Requesting Physicians
Enter Partial Name: (ALL)

Choose Radiologists
Enter Partial Name: (ALL)

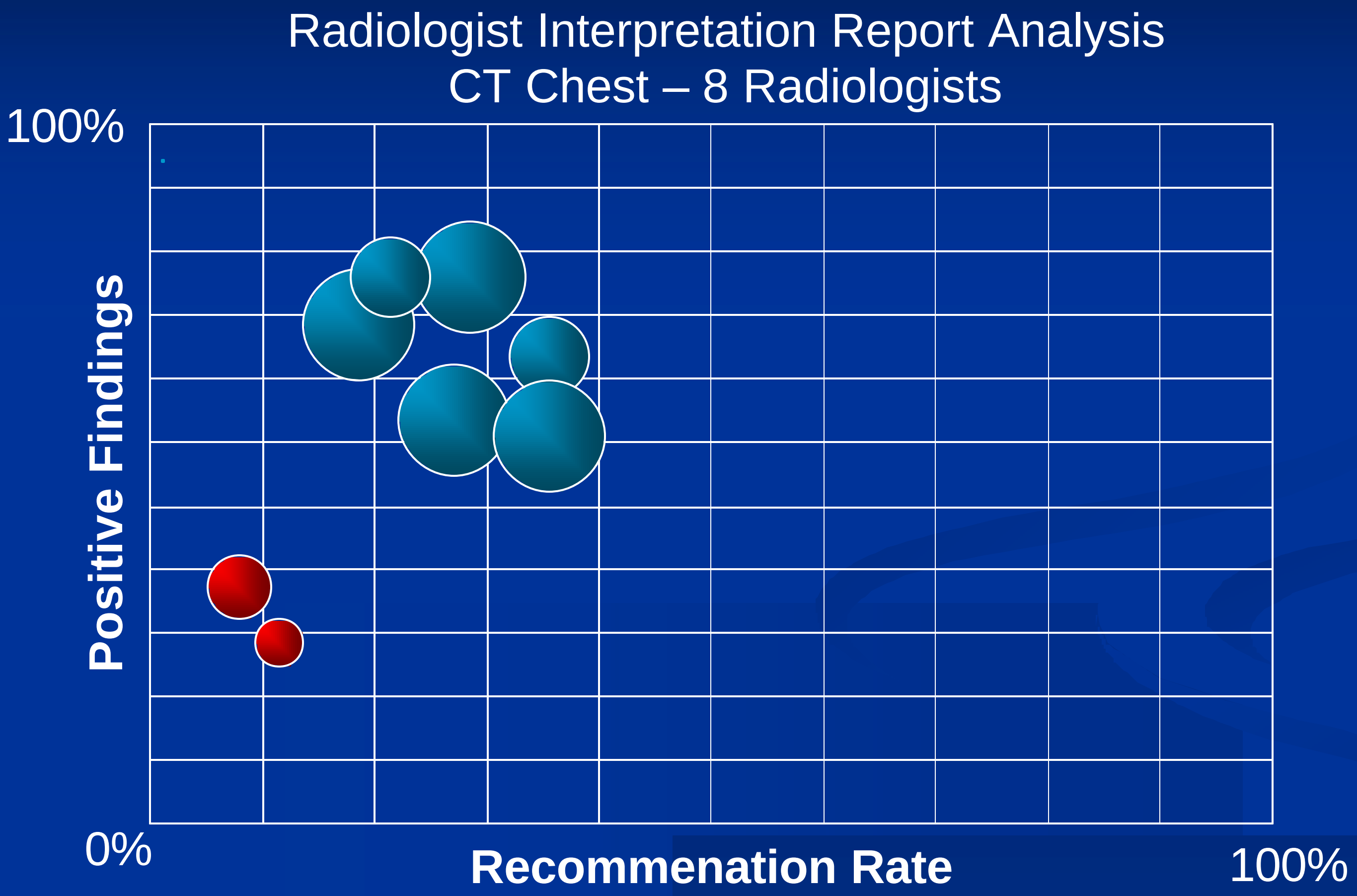
● Date
○ Doctor

Composite Modality	Subtotal	Pos FX	Pos REC
GRAND TOTAL	4072636	67.68	7.28
SUBTOTAL	4072636	67.68	7.28
CHEST-X RAY	1208059	72.72	5.60
PERIPHERAL-X RAY	662145	61.75	4.50
BREAST-MAMMOGRAPHY	279850	24.85	6.74
SPINE-X RAY	188699	62.54	6.81
ABDOMEN-CT	186699	76.12	11.86
PELVIS-ULTRASOUND	158680	85.99	13.38
ABDOMEN-X RAY	156662	66.79	7.59
HEAD-MR	134683	69.39	9.14
HEAD-CT	134308	66.67	12.26
CHEST-CT	127136	81.65	25.29
CHEST-NUCLEAR MEDICINE	81814	98.18	0.69
ABDOMEN-ULTRASOUND	63186	62.82	6.92
SPINE-MR	58097	82.18	9.53
ABDOMEN-X RAY-BARIUM	53846	64.58	3.15
PERIPHERAL-ULTRASOUND	41088	47.85	2.04
BODY-NUCLEAR MEDICINE	40746	65.43	12.10
PERIPHERAL-MR	37820	87.54	6.40
SPINE-CT	28268	61.38	8.88
BREAST-ULTRASOUND	26089	62.22	17.89
	23933	67.22	5.11
HEAD-ANGIOGRAM	22573	79.47	5.54
HEAD-ULTRASOUND	21322	62.00	7.70
CHEST-ANGIOGRAM	20680	84.59	1.07
PERIPHERAL-ANGIOGRAM	18218	83.77	1.90

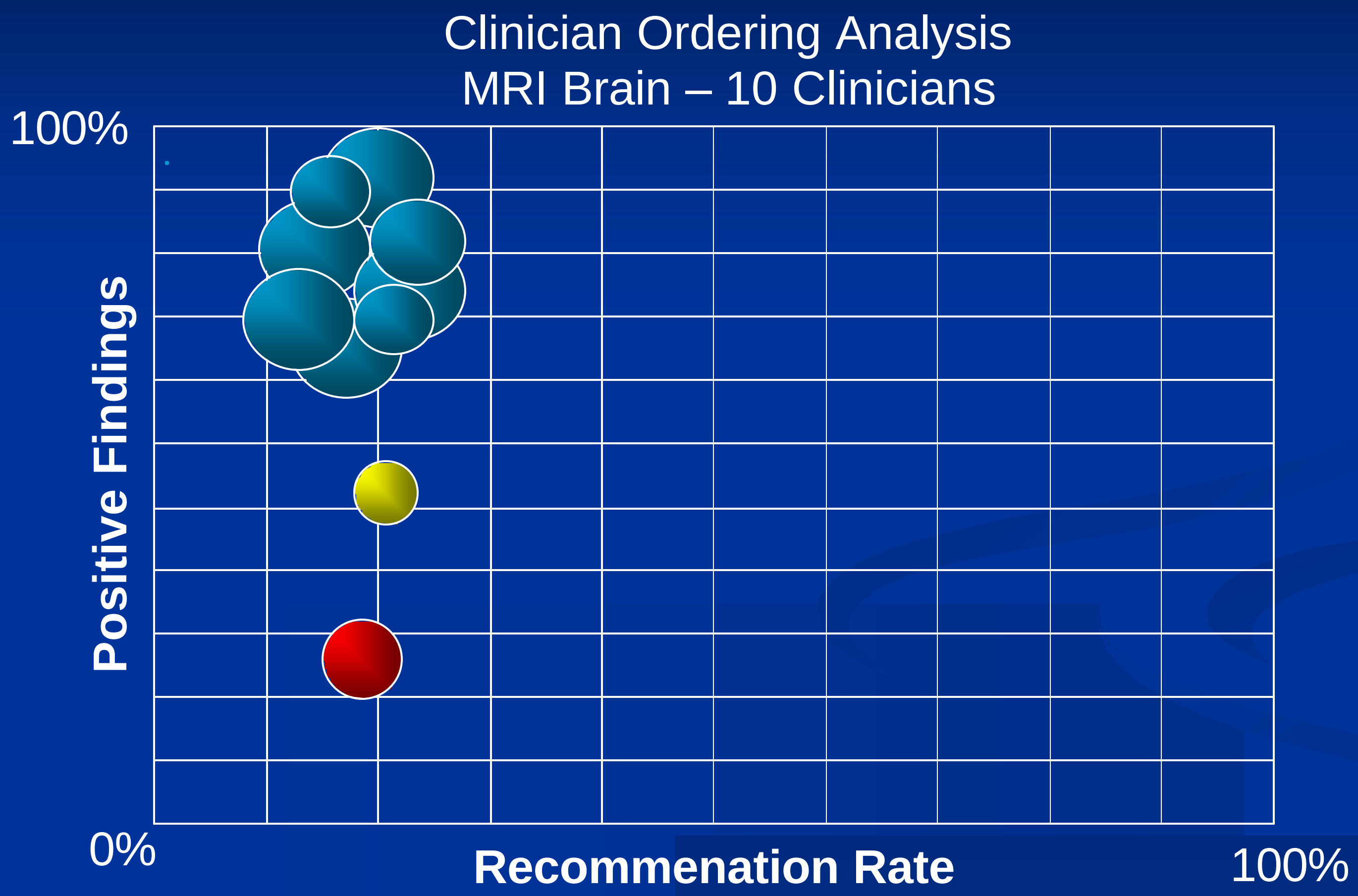
Diagnostic Decisions Evaluation



Diagnostic Decisions Evaluation

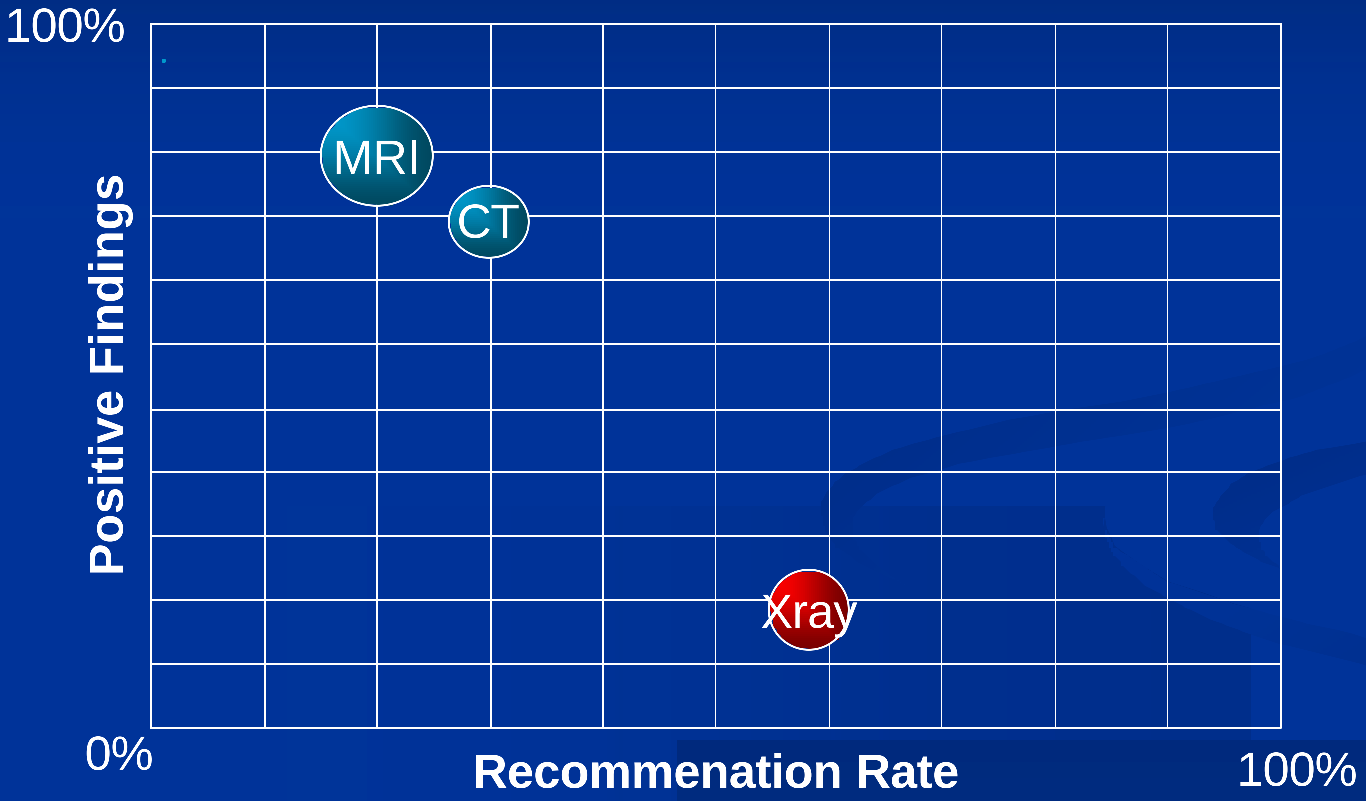


Diagnostic Decisions Evaluation



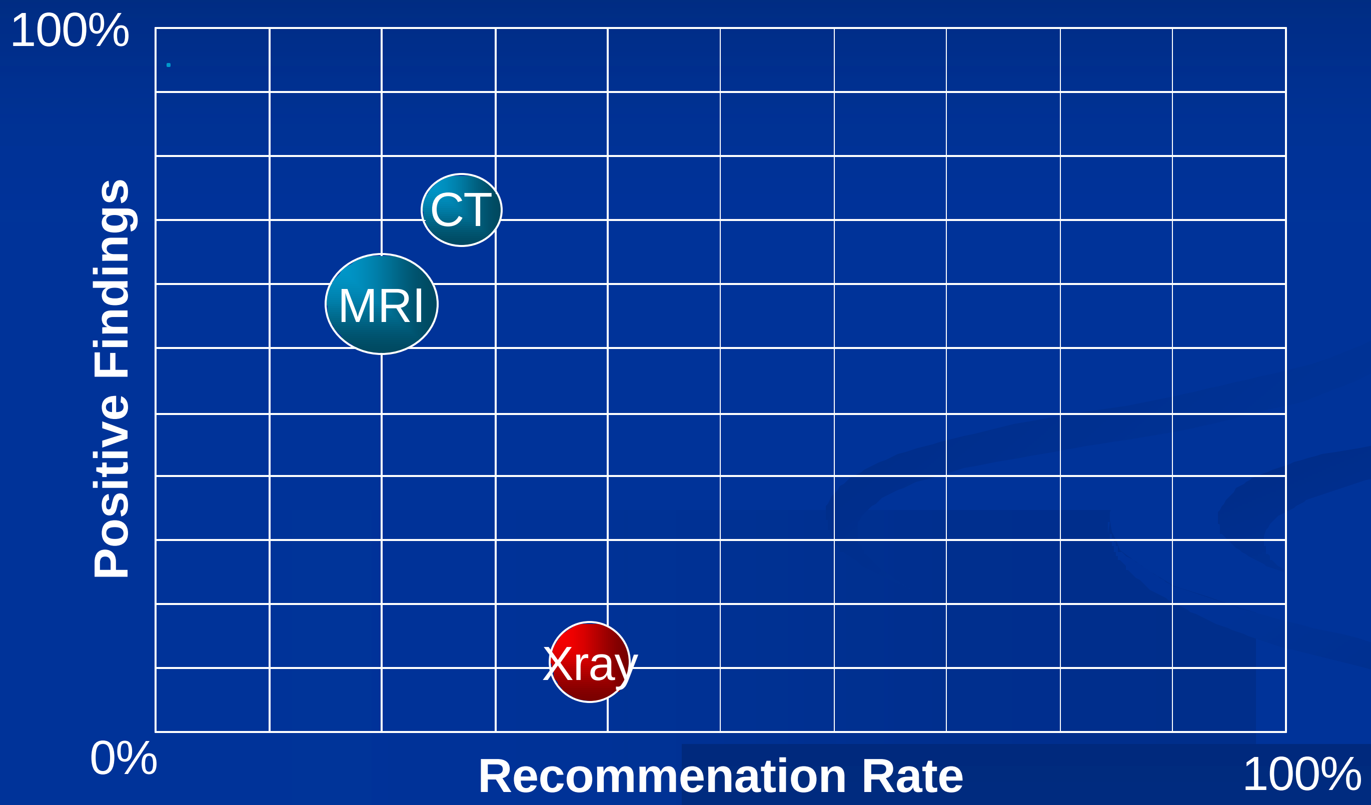
Indication ?   Examination Evaluation

Thunderclap Headaches ?   Head Imaging Analysis
males over 40 years old

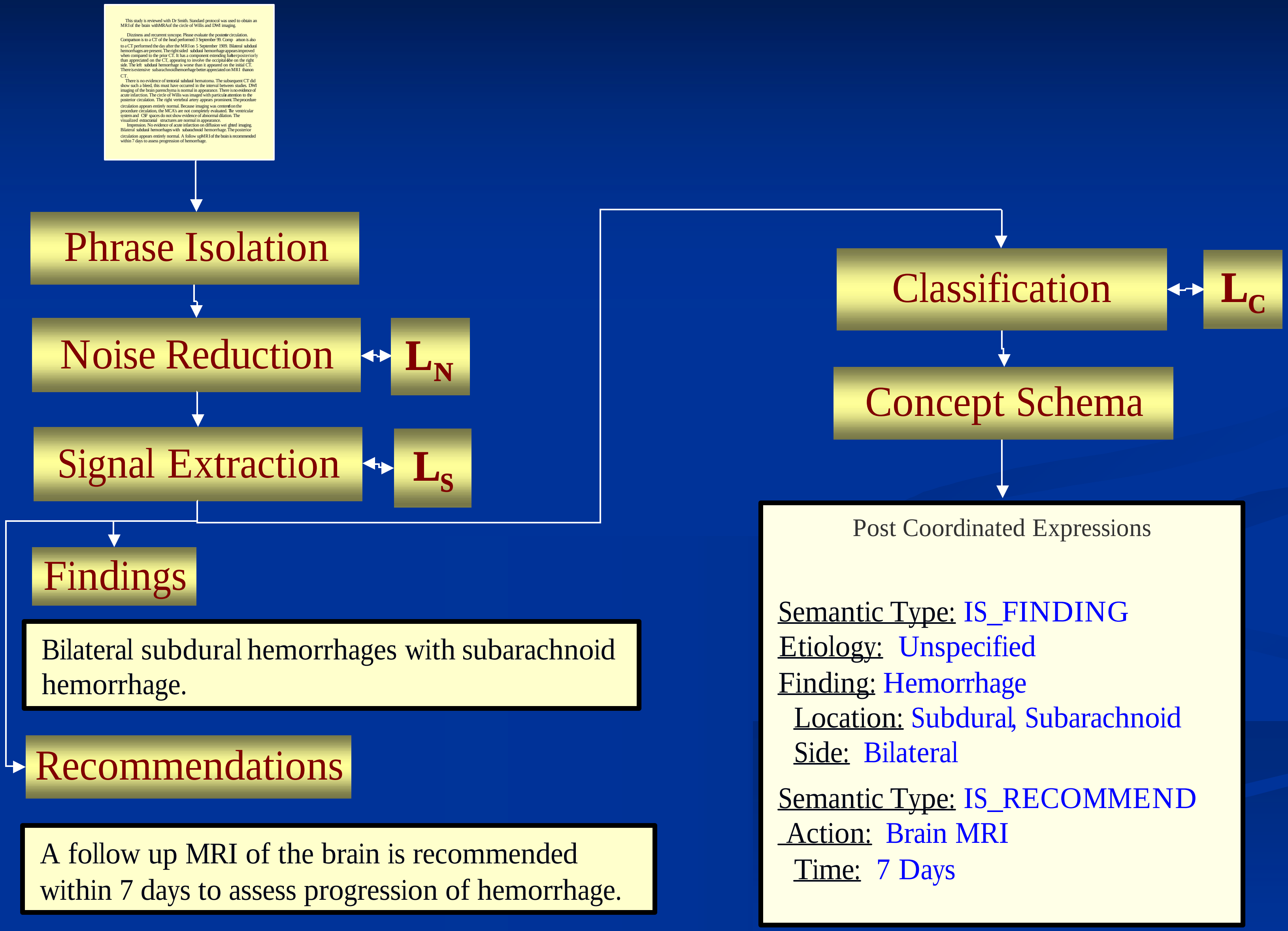


Indication ?   Examination Evaluation

Thunderclap Headaches ?   Head Imaging Analysis
males under 40 years old



Signal Classification



This study is reviewed with Dr Smith. Standard protocol was used to obtain an MRI of the brain withMRAof the circle of Willis and DWI imaging.

Dizziness and recurrent syncope. Please evaluate the posterior circulation. Comparison is to a CT of the head performed 3 September 99. Comparison is also to a CT performed the day after the MRI on 5 September 1909. Bilateral subdural hemorrhages are present. The right sided subdural hemorrhage appears improved when compared to the prior CT. It has a component extending further posteriorly than appreciated on the CT, appearing to involve the occipital lobe on the right side. The left subdural hemorrhage is worse than it appeared on the initial CT. There is extensive subarachnoidhemorrhage better appreciated on MRI thanon CT.

There is no evidence of tentorial subdural hematoma. The subsequent CT did show such a bleed, this must have occurred in the interval between studies. DWI imaging of the brain parenchyma is normal in appearance. There is no evidence of acute infarction. The circle of Willis was imaged with particular attention to the posterior circulation. The right vertebral artery appears prominent. The procedure circulation appears entirely normal. Because imaging was centered on the procedure circulation, the MCA's are not completely evaluated. The ventricular system and CSF spaces do not show evidence of abnormal dilation. The visualized extracranial structures are normal in appearance.

Impression. No evidence of acute infarction on diffusion weighted imaging. Bilateral subdural hemorrhages with subarachnoid hemorrhage. The posterior circulation appears entirely normal. A follow upMRI of the brain is recommended within 7 days to assess progression of hemorrhage.

Phrase Isolation

Noise Reduction ↔ L_N

Signal Extraction ↔ L_S

Findings

Bilateral subdural hemorrhages with subarachnoid hemorrhage.

Recommendations

A follow up MRI of the brain is recommended within 7 days to assess progression of hemorrhage.

Classification ↔ L_C

Concept Schema

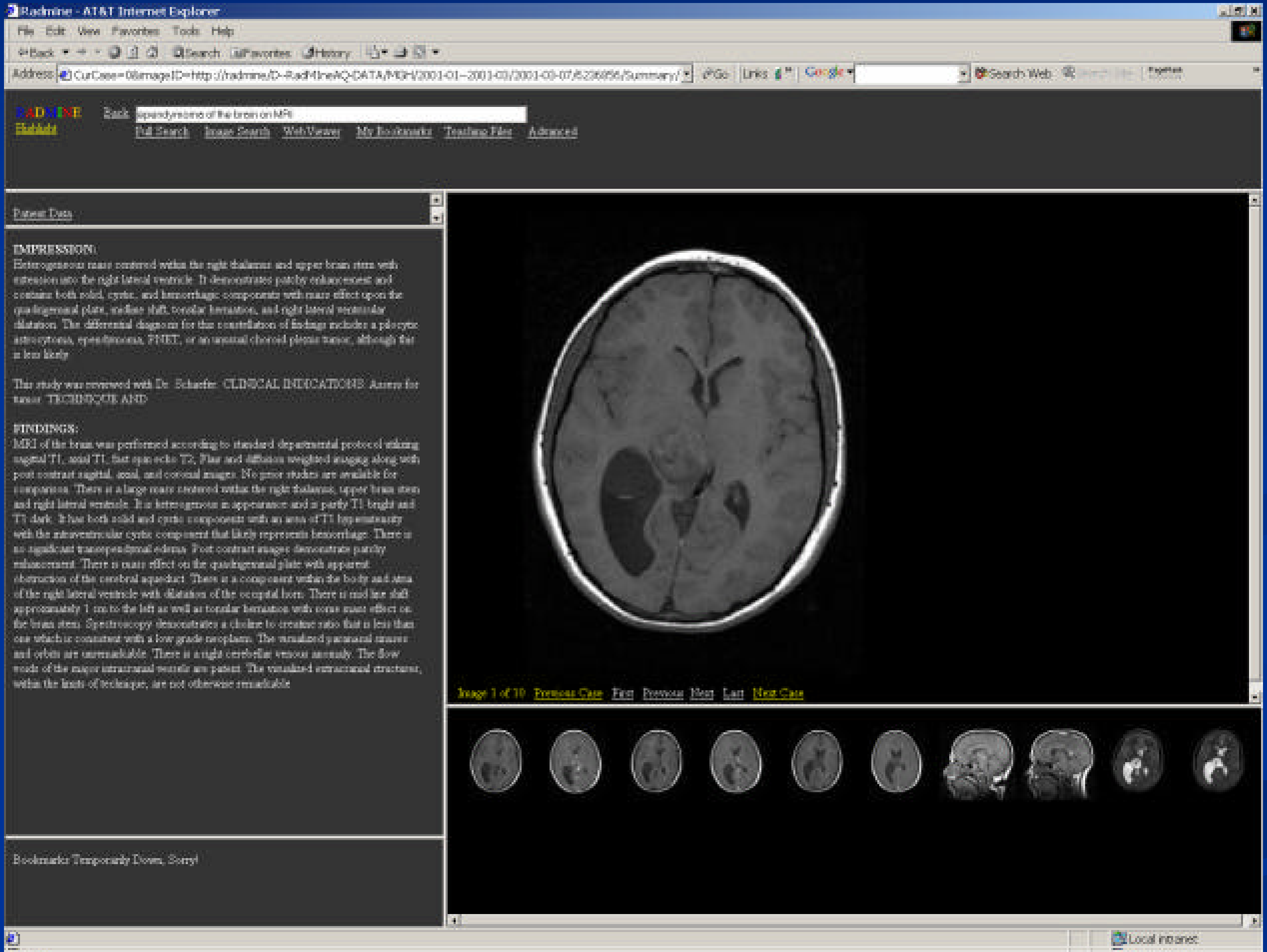
Post Coordinated Expressions

Semantic Type: IS_FINDING
Etiology: Unspecified
Finding: Hemorrhage
Location: Subdural, Subarachnoid
Side: Bilateral

Semantic Type: IS_RECOMMEND
Action: Brain MRI
Time: 7 Days

Extraction of Related Information

ependymoma of the brain on MRI



CurCase=0&imageID=http://radmine/D-RadMineAQ-DATA/MGH/2001-01--2001-03/2001-03-07/6236956/Summary/

ependymoma of the brain on MRI

IMPRESSION:
Heterogeneous mass centered within the right thalamus and upper brain stem with extension into the right lateral ventricle. It demonstrates patchy enhancement and contains both solid, cystic, and hemorrhagic components with mass effect upon the quadrigeminal plate, midline shift, tonsilar herniation, and right lateral ventricular dilatation. The differential diagnosis for this constellation of findings includes a pilocytic astrocytoma, ependymoma, PNET, or an unusual choroid plexus tumor, although the is less likely

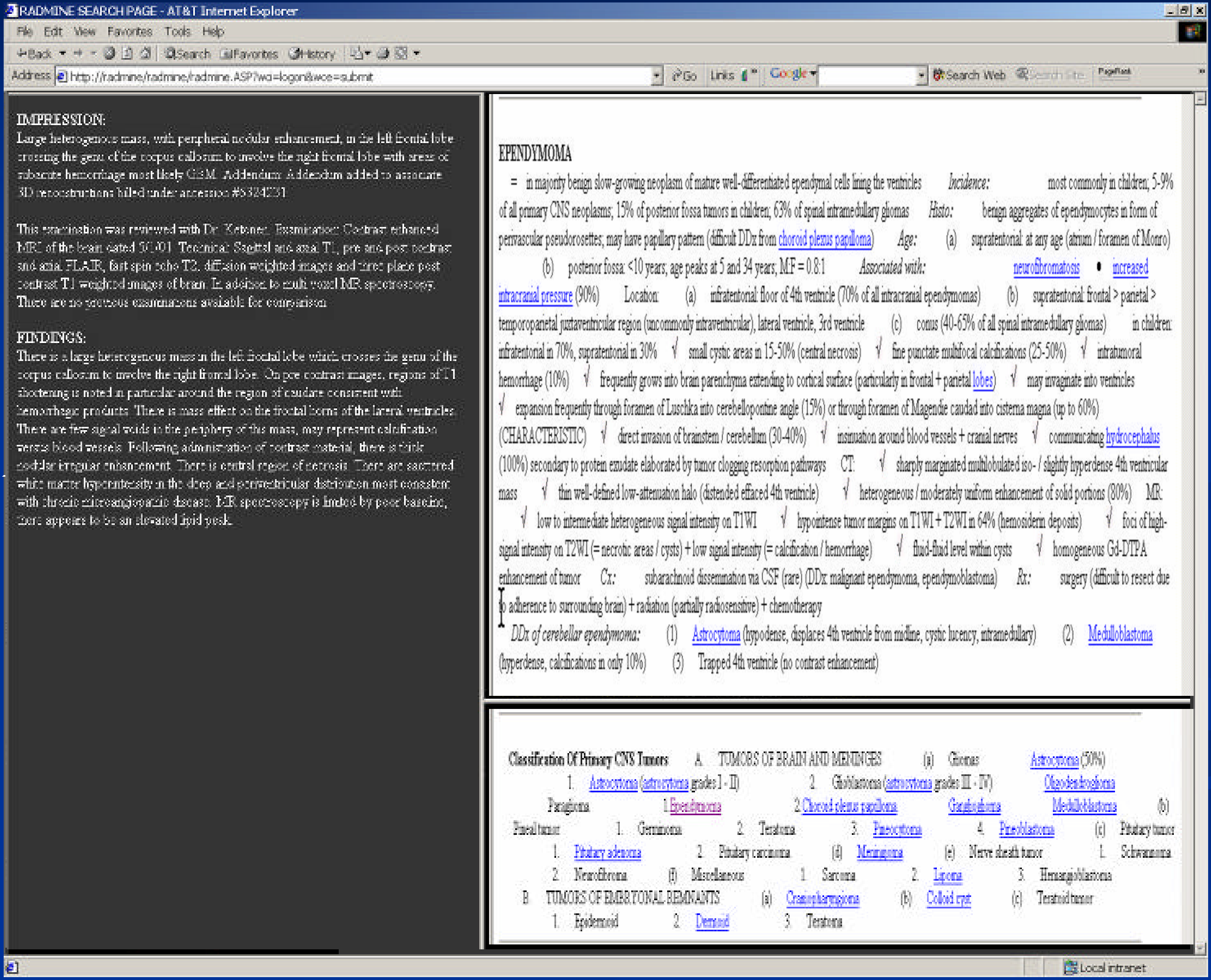
This study was reviewed with Dr. Schaefer. CLINICAL INDICATIONS: Assess for tumor. TECHNIQUE AND

FINDINGS:
MRI of the brain was performed according to standard departmental protocol utilizing sagittal T1, axial T1, fast spin echo T2, Flair and diffusion weighted imaging along with post contrast sagittal, axial, and coronal images. No prior studies are available for comparison. There is a large mass centered within the right thalamus, upper brain stem and right lateral ventricle. It is heterogeneous in appearance and is partly T1 bright and T1 dark. It has both solid and cystic components with an area of T1 hyperintensity with the intraventricular cystic component that likely represents hemorrhage. There is no significant transependymal edema. Post contrast images demonstrate patchy enhancement. There is mass effect on the quadrigeminal plate with apparent obstruction of the cerebral aqueduct. There is a component within the body and atria of the right lateral ventricle with dilatation of the occipital horn. There is mid line shift approximately 1 cm to the left as well as tonsilar herniation with some mass effect on the brain stem. Spectroscopy demonstrates a choline to creatine ratio that is less than one which is consistent with a low grade neoplasm. The visualized paranasal sinuses and orbits are unremarkable. There is a right cerebellar venous anomaly. The flow voids of the major intracranial vessels are patent. The visualized extracranial structures, within the limits of technique, are not otherwise remarkable.

Image 1 of 10

Bookmarks Temporarily Down, Sorry!

Extraction of Related Information


RADMINE SEARCH PAGE - AT&T Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://radmine/radmine/radmine.ASP?wci=logon&wce=submit
Go Links Google Search Web
Local intranet

IMPRESSION:
Large heterogenous mass, with peripheral nodular enhancement, in the left frontal lobe crossing the genu of the corpus callosum to involve the right frontal lobe with areas of subacute hemorrhage most likely GBM. Addendum: Addendum added to associate 3D reconstructions filed under accession #5324231

This examination was reviewed with Dr. Ketonen. Examination: Contrast enhanced MRI of the brain dated 5/1/01. Technical: Sagittal and axial T1, pre and post contrast and axial FLAIR, fast spin echo T2, diffusion weighted images and three plane post contrast T1 weighted images of brain. In addition to multi voxel MR spectroscopy. There are no previous examinations available for comparison.

FINDINGS:
There is a large heterogenous mass in the left frontal lobe which crosses the genu of the corpus callosum to involve the right frontal lobe. On pre contrast images, regions of T1 shortening is noted in particular around the region of caudate consistent with hemorrhagic products. There is mass effect on the frontal horns of the lateral ventricles. There are few signal voids in the periphery of this mass, may represent calcification versus blood vessels. Following administration of contrast material, there is thick nodular irregular enhancement. There is central region of necrosis. There are scattered white matter hyperintensity in the deep and periventricular distribution most consistent with chronic microangiopathic disease. MR spectroscopy is limited by poor baseline, there appears to be an elevated lipid peak.

EPENDYMOMA

= in majority benign slow-growing neoplasm of mature well-differentiated ependymal cells lining the ventricles *Incidence:* most commonly in children; 5-9% of all primary CNS neoplasms; 15% of posterior fossa tumors in children; 63% of spinal intramedullary gliomas *Histo:* benign aggregates of ependymocytes in form of perivascular pseudorosettes; may have papillary pattern (difficult DDx from choroid plexus papilloma) *Age:* (a) supratentorial: at any age (atrium / foramen of Monro) (b) posterior fossa: <10 years; age peaks at 5 and 34 years; M:F = 0.8:1 *Associated with:* neurofibromatosis • increased intracranial pressure (90%) Location: (a) infratentorial: floor of 4th ventricle (70% of all intracranial ependymomas) (b) supratentorial: frontal > parietal > temporoparietal juxtaventricular region (uncommonly intraventricular), lateral ventricle, 3rd ventricle (c) conus (40-65% of all spinal intramedullary gliomas) in children: infratentorial in 70%, supratentorial in 30% √ small cystic areas in 15-50% (central necrosis) √ fine punctate multifocal calcifications (25-50%) √ intratumoral hemorrhage (10%) √ frequently grows into brain parenchyma extending to cortical surface (particularly in frontal + parietal lobes) √ may invaginate into ventricles √ expansion frequently through foramen of Luschka into cerebellopontine angle (15%) or through foramen of Magendie caudad into cisterna magna (up to 60%) (CHARACTERISTIC) √ direct invasion of brainstem / cerebellum (30-40%) √ insinuation around blood vessels + cranial nerves √ communicating hydrocephalus (100%) secondary to protein exudate elaborated by tumor clogging resorption pathways CT: √ sharply marginated multilobulated iso- / slightly hyperdense 4th ventricular mass √ thin well-defined low-attenuation halo (distended effaced 4th ventricle) √ heterogeneous / moderately uniform enhancement of solid portions (80%) MR: √ low to intermediate heterogeneous signal intensity on T1WI √ hypointense tumor margins on T1WI + T2WI in 64% (hemosiderin deposits) √ foci of high-signal intensity on T2WI (= necrotic areas / cysts) + low signal intensity (= calcification / hemorrhage) √ fluid-fluid level within cysts √ homogeneous Gd-DTPA enhancement of tumor *Cx:* subarachnoid dissemination via CSF (rare) (DDx: malignant ependymoma, ependymoblastoma) *Rx:* surgery (difficult to resect due to adherence to surrounding brain) + radiation (partially radiosensitive) + chemotherapy

DDx of cerebellar ependymoma: (1) Astrocytoma (hypodense, displaces 4th ventricle from midline, cystic lucency, intramedullary) (2) Medulloblastoma (hyperdense, calcifications in only 10%) (3) Trapped 4th ventricle (no contrast enhancement)

Classification Of Primary CNS Tumors A. TUMORS OF BRAIN AND MENINGES (a) Gliomas Astrocytoma (50%) 1. Astrocytoma (astrocytoma grades I - II) 2. Glioblastoma (astrocytoma grades III - IV) Oligodendroglioma Paraglioma 1 Ependymoma 2 Choroid plexus papilloma Ganglioglioma Medulloblastoma (b) Pineal tumor 1. Germinoma 2. Teratoma 3. Pineocytoma 4. Pineoblastoma (c) Pituitary tumor 1. Pituitary adenoma 2. Pituitary carcinoma (d) Meningioma (e) Nerve sheath tumor 1. Schwannoma 2. Neurofibroma (f) Miscellaneous 1. Sarcoma 2. Lipoma 3. Hemangioblastoma B. TUMORS OF EMBRYONAL REMNANTS (a) Craniopharyngioma (b) Colloid cyst (c) Teratoid tumor 1. Epidermoid 2. Dermoid 3. Teratoma

Knowledge Discovery

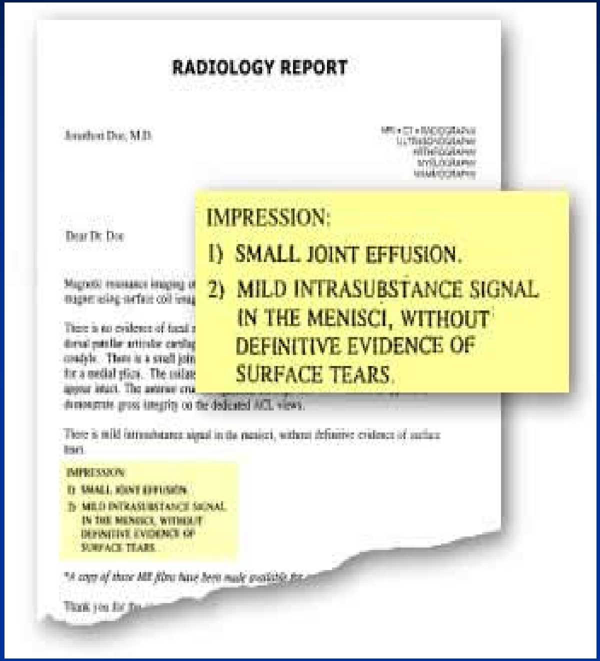
Knowledge Management

- No standards for existing knowledge
- Difficulties in representing knowledge
- Difficulties obtaining new knowledge

Ontology

- The hierarchical structuring of knowledge about things through sub-categorization according to their essential qualities.

Create Structure



RADIOLOGY REPORT

Jonathan Doe, M.D.

MRI • CT • RADIOGRAPHY
ULTRASONOGRAPHY
ARTHROGRAPHY
MYELOGRAPHY
MAMMOGRAPHY

Dear Dr. Doe

Magnetic resonance imaging [illegible] magnet using surface coil imag[illegible]

There is no evidence of focal [illegible] dorsal patellar articular cartilag[illegible] condyle. There is a small join[illegible] for a medial plica. The collate[illegible] appear intact. The anterior cru[illegible] demonstrate gross integrity on the dedicated ACL views.

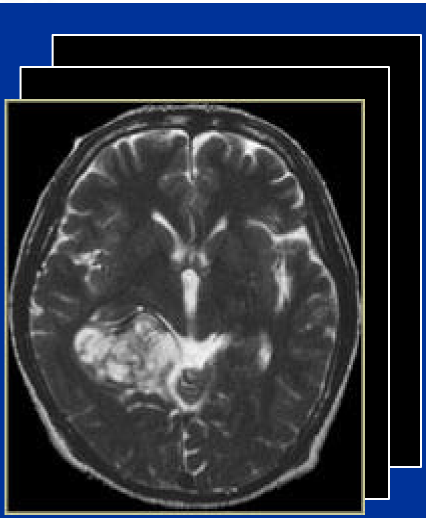
There is mild intrasubstance signal in the menisci, without definitive evidence of surface tears.

IMPRESSION:
1) SMALL JOINT EFFUSION.
2) MILD INTRASUBSTANCE SIGNAL IN THE MENISCI, WITHOUT DEFINITIVE EVIDENCE OF SURFACE TEARS.

*A copy of these MR films have been made available for [illegible]

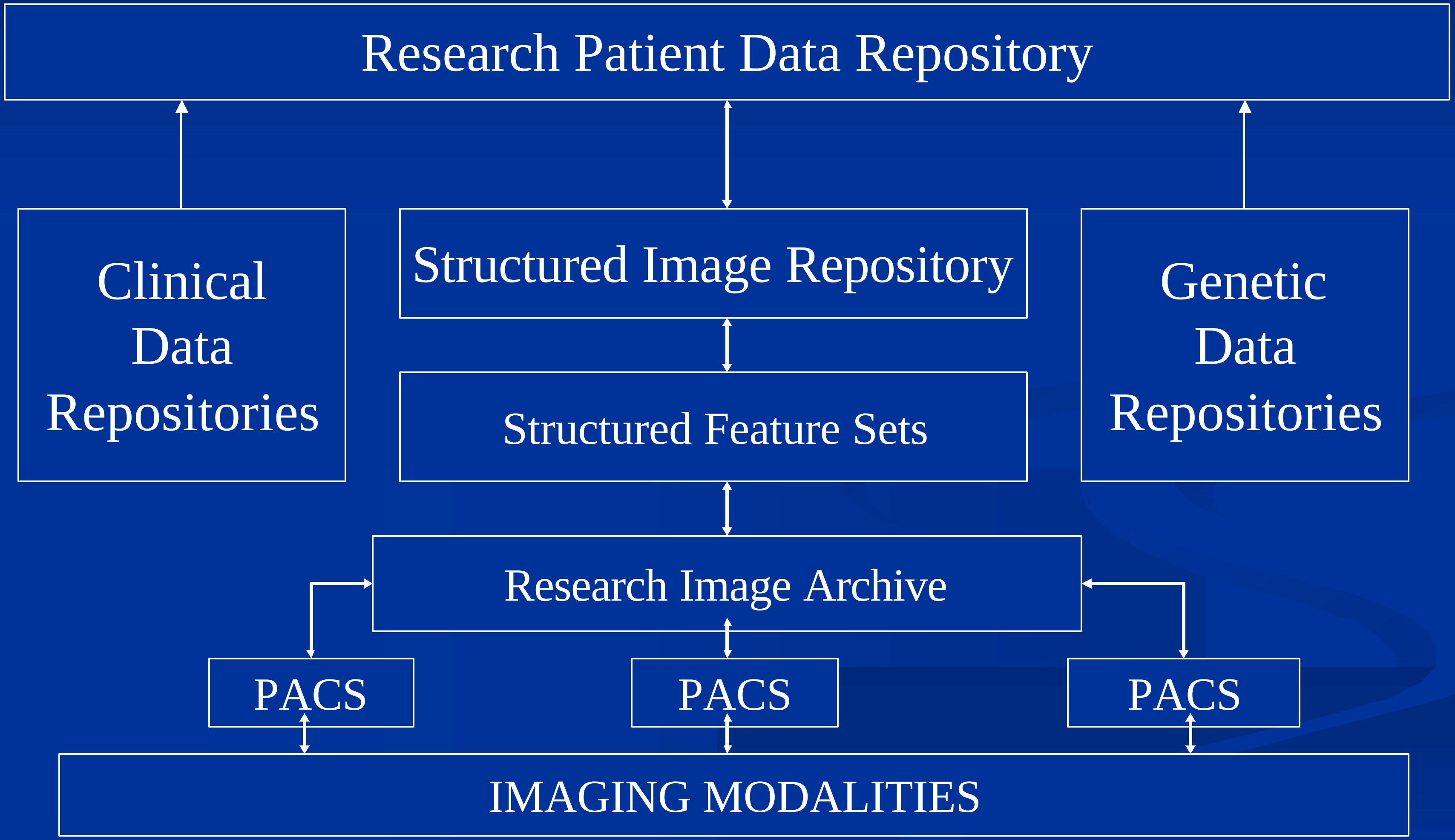
Thank you for [illegible]

IMPRESSION:
1) SMALL JOINT EFFUSION.
2) MILD INTRASUBSTANCE SIGNAL IN THE MENISCI, WITHOUT DEFINITIVE EVIDENCE OF SURFACE TEARS.



- **Classifying Interpreted Findings**
  - LEXIMER (Lexicon Medicated Entropy Reduction)
  - Codification - CPT, ICD-9, RADLEX, SNOMED
- **Image Feature Extraction**
  - Modern CAD Applications
    - Digital Mammography
      - Micro-Calcification Detection
    - CT Chest
      - Lung Nodule Detection
  - Future Application Examples
    - Brain CT, MRI
      - Sagittal asymmetry index, Tumor detection
    - Cardiac MRI, US, NM
      - Myocardial wall thickness quantification, HCM
    - Future Imaging Modalities
      - Molecular, Fusion, 3D Imaging
- **Essential for knowledge discovery**

Clinical Data System Integration



Research Patient Data Repository

Clinical Data Repositories

Structured Image Repository

Structured Feature Sets

Genetic Data Repositories

Research Image Archive

PACS

PACS

PACS

IMAGING MODALITIES

Imaging and Genetics

- Genotype
  - An individual's genetic composition.
  - Identified by genetic analysis.

  +

- Penetrance, Developmental and Environmental

  =

- Phenotype
  - The features of health and disease expressed throughout life.
  - Can be identified, in large part, by medical imaging.

Genetics 101

- Allele
  - One of at least two forms of any individual gene.

- SNP (Single Nucleotide Polymorphism)
  - Small DNA regions that vary between individuals.

- It is these differences that represent underlying disease susceptibility and drug responsiveness.

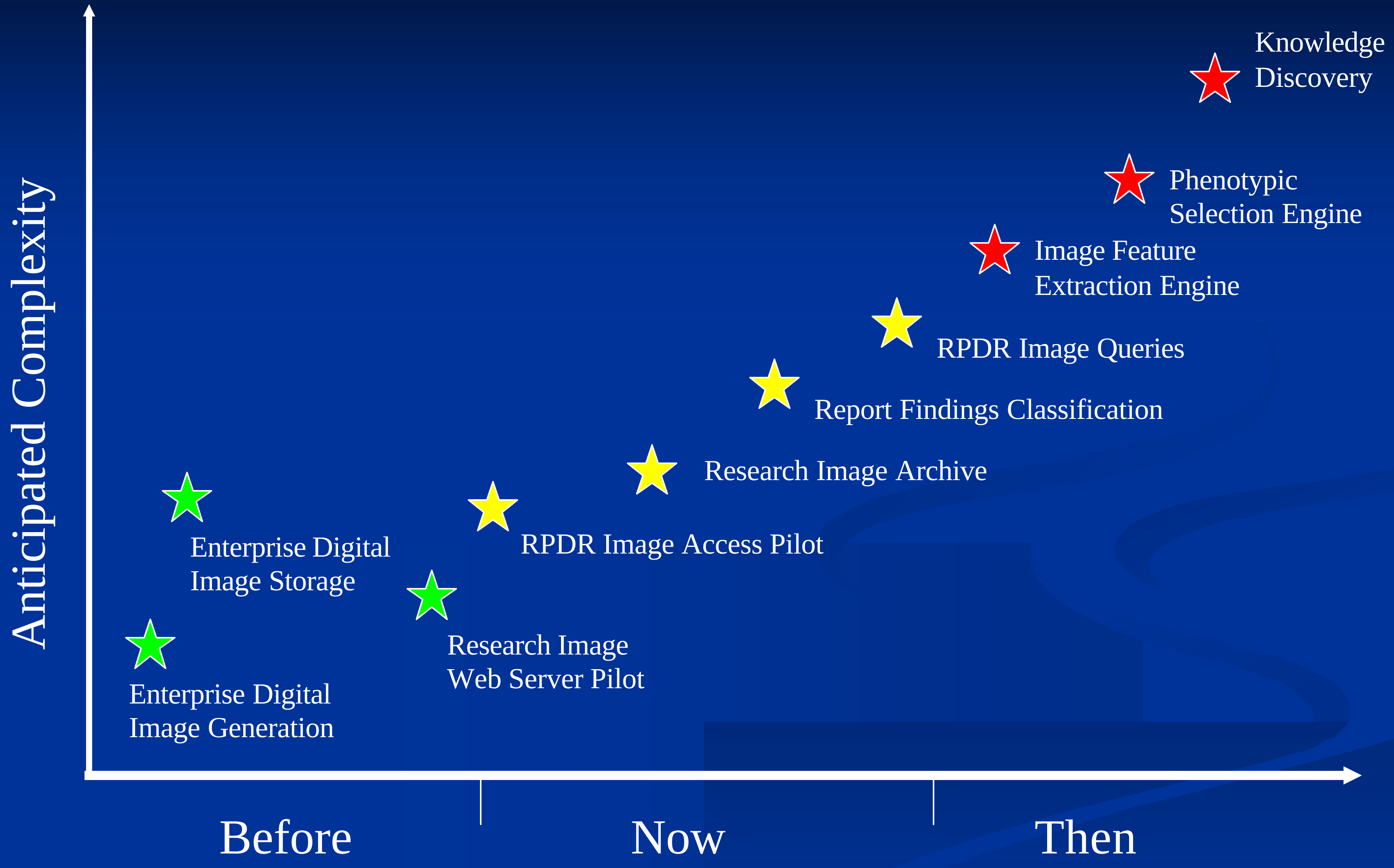
Disease Knowledge Discovery

- **Understand diseases in individuals via medical imaging**
  - Identify phenotypic expression suggestive of disease
  - Explore genetic analysis alternatives
    - Identify a potential genetic penetrance from allele or SNP
  - Evaluate therapy effectiveness

- **Understand diseases in individuals via genetic analysis**
  - Identify genotypic expression suggestive of disease
  - Classify the pathologic spectrum of expression states
    - Standard Imaging - Pathologic Anatomic and Physiologic States
    - Molecular Imaging – Specifically Targeted Expression States
  - Create specific imaging work-ups for genotypes
  - Evaluate therapy effectiveness

- **Discover new diseases in populations using genetic and image analysis**
  - Utilize extensive combinations of genotypic and phenotypic data
    - Perform clustering algorithms to identify classification possibilities
  - Map genetic states to possible phenotypic outcomes
  - Map phenotypic presentations to possible genetic states
  - Monitor therapy alternatives

Summary Timeline



Anticipated Complexity

Enterprise Digital Image Generation

Enterprise Digital Image Storage

Research Image Web Server Pilot

RPDR Image Access Pilot

Research Image Archive

Report Findings Classification

RPDR Image Queries

Image Feature Extraction Engine

Phenotypic Selection Engine

Knowledge Discovery

Before

Now

Then

Clinical Data Mining

"If we knew what it was we were doing, it wouldn’t be called research."
- Albert Einstein

Dr. Keith J. Dreyer
Partners HealthCare System ~ Massachusetts General Hospital
Harvard Medical School