Large and h**umongous Clinical Data Bases Imagine the Possibilities**

November 12, 2009 Bethesda, MD

Clement J. McDonald, M.D.
Director Lister Hill Center
National Library of Medicine

WHY DO WE CARE

The Shortage of Data for Clinical Decision Making

- Clinicians are faced with zillions of decisions
- Research helps them with a smidgen of these
  - Preventive decisions – but even some of these (pneumonia vaccine) are soft
  - Some cardiovascular interventions
  - Some anticoagulation interventions
- Minimal help with special circumstances – age, co-morbidity
- Little help for decisions about diagnostic testing, surgery, use of devices
- Almost no help with cost/benefits (Biggs- Almost no help regarding cost benefits (Haynes)

Formal research can’t do it all

- Know relatively little about cost/benefits (Biggs-- BMJ. 2000 December 2; 321(7273): 1362–1363.
- Know little about relative advantage of similar treatments- ‘because differences go down; so sample size requirements go up.

Some presumed lousy data are good predictors

- MD's prediction of probability of positive chemistry results was way off – horribly calibrated
  - Yet…strong statistical predictor of a positive result
- Answer to one question patient predicted health – very good predictor of health

Deeper problems

- Sample size requirements for trials become difficult/Impossible
  - When event rates are low and…
  - When difference between “treatment and control” are modest-
    - often the case in comparisons of new with best existing treatment
  - Want we want to accurately quantify the amount of benefit for cost benefit analysis

Deeper problems - more

- A study with 10% event rate and 25% difference (big difference) can require enrollment of 10,000 patients !
- To be 95% sure of finding one case of finding with event rate of 1/25,000 need to observe 63,000 cases (e.g. rhabomyolysis)
- So clinical trials will never find rare adverse events
- And they just can’t cover the water front-
  - Have to think about clinical data bases

Cost matters-

- Limits the numbers and scope of trials.
- Forces look at very narrow (homogeneous) slice of populations to minimize sample size
- Increases the cost of health care (by inflating product costs)
- Other kinds of research (e.g. genome sequencing)is getting more efficient by orders of magnitude
- We need to find comparable efficiencies in clinical research

How to Get More for Less

- Collect less on greater numbers of patients (large simple trials)
- Leverage existing clinical data
  - Use existing to fulfill some data requirements of clinical trials
  - Build Large clinical data bases from it for data base studies
  - Combine and re-use clinical trial data –
- Use the computer methods to accomplish this

The Good News

- 1. Rich lodes of clinical data available for clinical use
- 2. Increasing opportunities to link clinical, and genetic, data
- 3. Opportunities– (In theory) for efficient recruitment of patients
- 4. New potentials through PHR- will not get to thaat

ACADEMIC EMR'S AS DATA SOURCES

EMRs

- Almost every health care institution has one
- Some are quite large, and long standing and tuned for research use (examples)

EMR Data availability

- Lab data (almost always electronic
- Medication orders in patients
- Radiology reports (text)
- Pathology reports (text)
- Dictation (discharge summary)
- EKGs
- Cardiac echoes
- Endoscopy
- Spirometry

Different types of data

- Numbers and coded variables “easy”
- Narrative reports harder to mine
  - Can do broad scope search for words and then review by hand
  - Narrowly targeted NLP works fairly well too- b

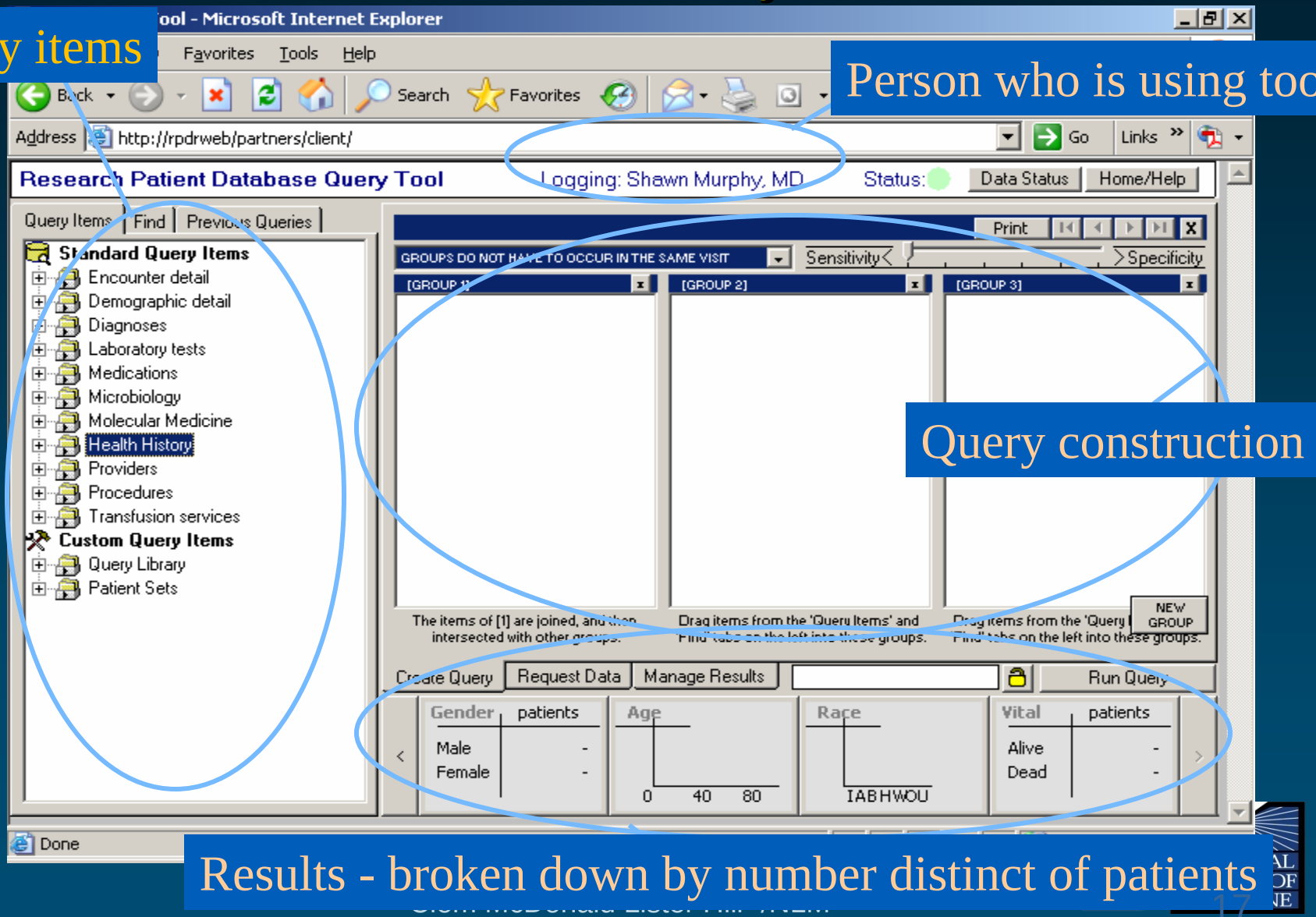
Research uses of EMRs

- To find numbers and statistics needed to plan studies and write grants.
- Help to recruit study patients with providers consent and involvement
  - E.g. Regenstrief , Columbia, Harvard, Vanderbilt ,Kaiser, the VA.
- De-identified studies and statistical analysis (many examples)
- Provide follow up data for longitudinal studies.
- 80% of Human studies used RI Data during some point of their evolution at IU.

Some example academic query tools

- Harvard Partners 4+ million patients 1.2 Billion discrete measurements, drugs information items. Genetic data or samples available on some
- Indiana Health Information Exchange 5 million patients 1.5 billion drugs and stored variables -
- Vanderbilt Medical Center –tied to genetic data
  - Mining narrative for phenotypes
  - Part of a larger consortium-

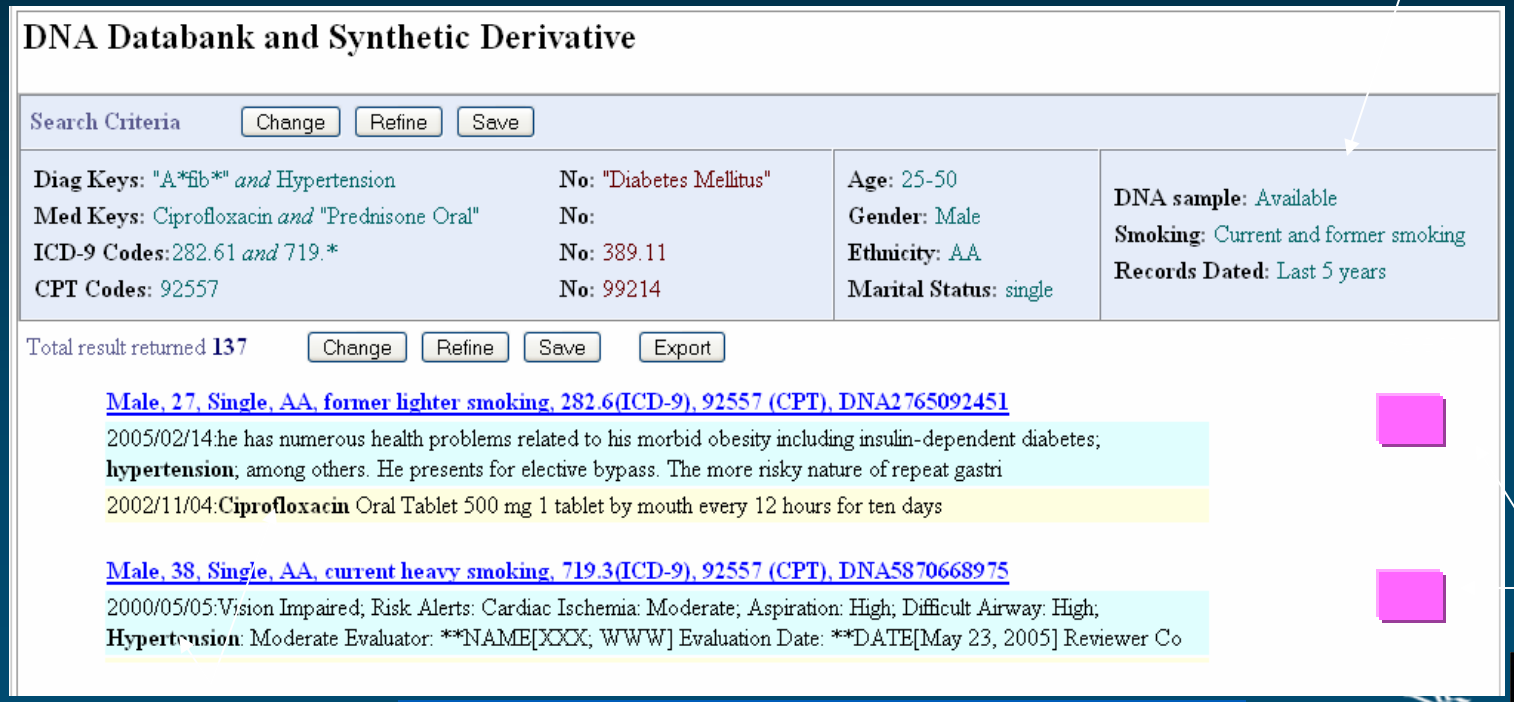
Harvard-Partner's Query tool



Vanderblit's specimen search

- Researcher executes search using defined parameters
- Researcher selects samples

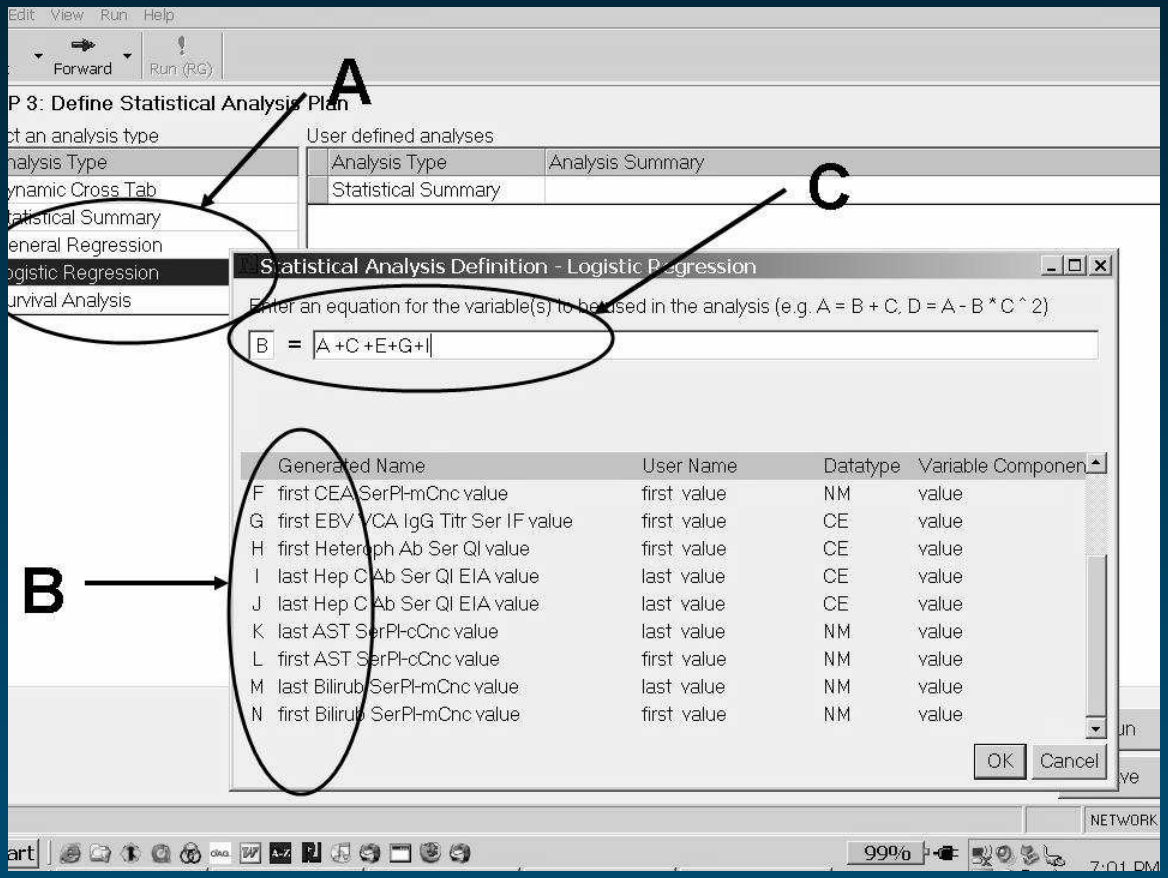
Search requirements specify to return only records with biological samples with a certain volume amount



Researcher selects most appropriate records

Keywords in context provide information for evaluating records

Regenstrief SPIN tool



Some f other medical centers that use their EMRs as research data bases

- Kaiser Permanente (? More then 1 billion for 8.7 million outpatients (as of 2008)
  - $25 M grant to genotype and phenotype (much through EMR) an enrollee sample
- Mayo Clinics– VERY long term history on patients in immediate county – huge publication list
- Veterans Administration- the largest- also a long history of research using their central data bass

Good news on Connections-

- Almost every clinical system marketed to hospitals or large group practices can deliver data in a standard message format –HL7 –
- And can send to researchers
- For result messages these work quite well except when senders jam things into the wrong fields

HEALTH INFORMATION EXCHANGES

IHEs

- Also called Regional Health information Organization (RHIO)
- Think of them a EMRs expanded to contain “all” (much) of the medical record data from a community or a region

An example

- Indiana Network for Patient Care
  - 46 Hospitals at last count) plus other sources The above screens- exemplify
  - Report push to more than 13,,000 MDs in central Indiana
  - Integrated public health functions
  - Citywide research database
  - More

HIE's are population based

- Very important
- Hospitals know about the hospital course-- can identify in hospital outcomes, but not long term outcomes
- Need special effort (and funding) to trace and get follow up - ala the Duke Cardiology data base to get unbiased outcomes
- HMOs- and similar (VA) have turn over and exclusions so can be patchy without same special effort

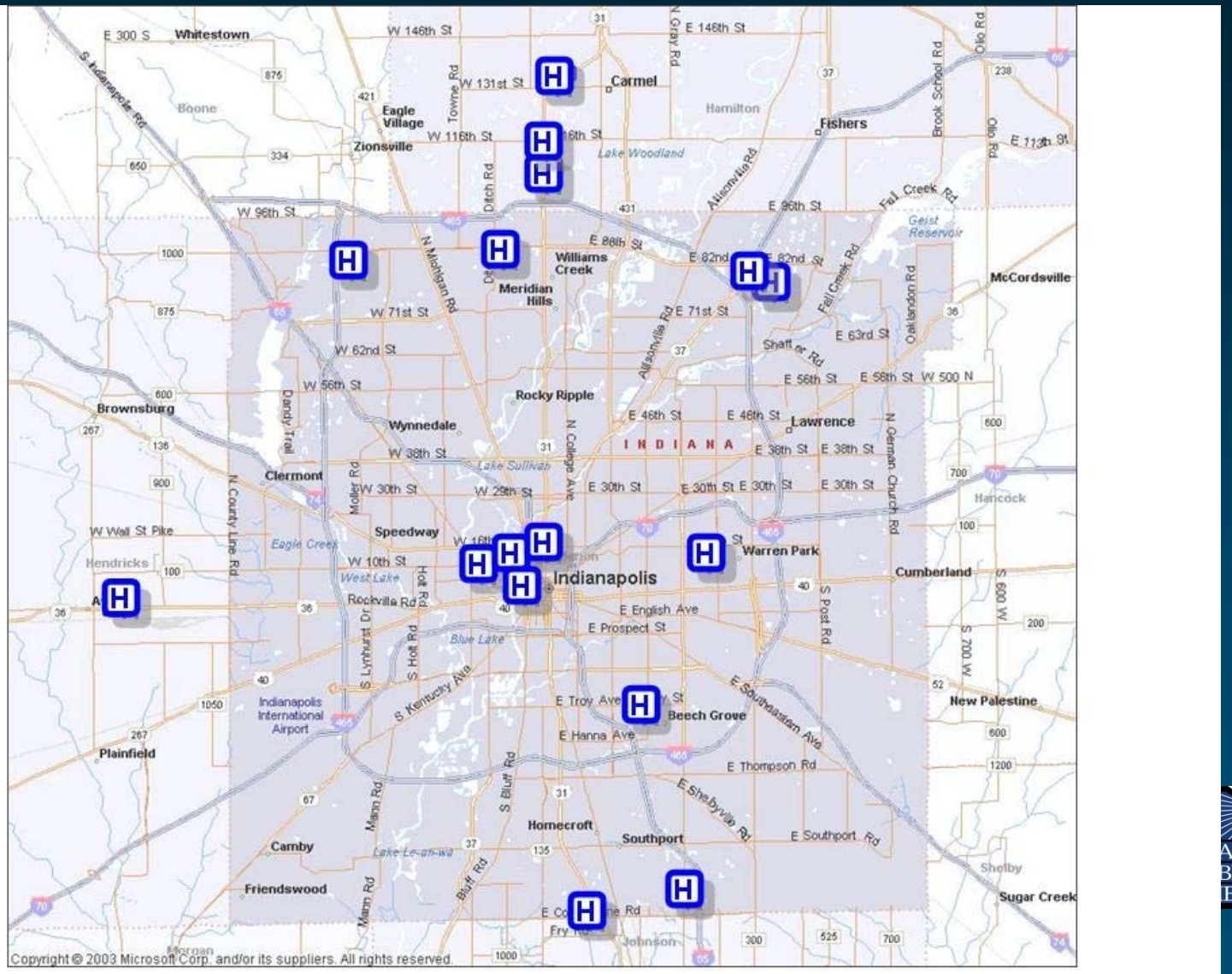
Population based more

- They will can pick up most of the outcomes in a community
  - Patients will go some where for a failed surgery (not necessarily the original place

HIE's Some Examples

- INPC central Indiana **(McDonald 2006 Health Affairs)**
- The one I know best

Indianapolis- Data from all hospitals (1B) in one Repository



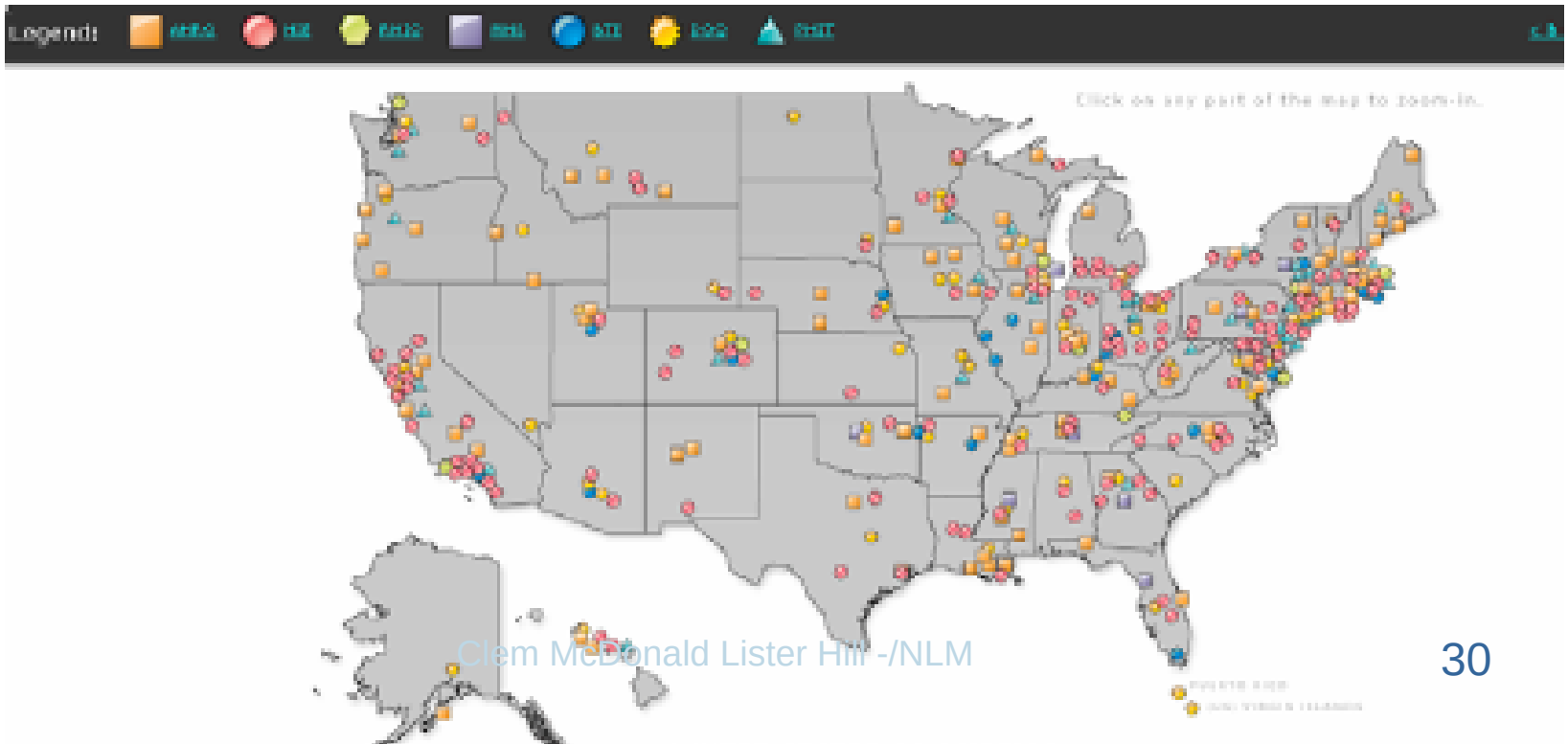
Other active IHEs

- Memphis (Venderbilt)
- The Ontario Children's network (all test results from all pediatric hospitals made available to all pediatricians) Gill Hill
- Massachusetts e-Health project- five practices many sites (David Bates)
- **Many other starting- NY, California, Rhoad Island**

More RHIOs on the way

HIMSS
Transforming healthcare through IT
architects of change

State and Regional Initiatives



Catchment problems

- However catchment is never complete
  - Patients migrate
  - Get procedures at referral centers outside of their catchment area (their IHE)
  - So need to analyze appropriate subsets
  - And look for external adjustment data (Later)

DATA BASES THAT CAN MITIGATE THE CATCHMENT PROBLEM

Death tapes

- Death is a key outcome- useful add-on to any observational data base
- Social security provides tapes carrying records of - 83 million deaths since inception of SS
- Includes SS#, name, zip code, birth date and death date

Death tapes- more

- Various subscription mechanism – 1 on line access is about $1K
- http://www.ntis.gov/products/ssa-weekly.aspx
- CDC has similar content based on death certificates and add cause of death.- Believe it's a per person searched cost

Medicare Data Bases

- Seven linked files (not counting drugs etc)
- Part A (hospital events) extrapolating from Indiana's numbers
  - 12-14 M hosp admissions per year
- Part B – Office and professional charges
  - 800-900 M charge events per year
- Data goes back (in some form) to at least 1994
- So about 330 M hospitalizations and Billions and billions of part B charge records.

Medicare Part D Drug data

- Regulation enabling its use for research is now final and published
- http://www.cms.hhs.gov/PrescriptionDrugCovGenIn/08_PartDData.asp
- All kinds of research opportunities-

More detailed clinical data may be coming from Medicare

- CARE prototype Post acute care – record
  - Long
  - Very rich data – including admission and discharge Dx's. , drugs, lab tests, complications –lots of survey instruments
  - Could be 14 billion distinct observations per year
- NLM provided some of the standards for this system
  - RxTerms For recording medications
- LOINC for each of the distinct questions – See hand out.

Medicare how to get

- For certain categories- provides confirming info and near complete catchment- lacks 3-6% > 65 hospitalizations (VA and non Medicare)
- Marvelous mining work by Wennberg
- Is available for research- with cost.
- Find out more at RESDAC
  - http://www.resdac.umn.edu/

More prescription information

- SureScripts Rx.Hub – a consortium of pharmacy benefit managers
  - 2.5 billion prescription records per year- 60-70% of the national volume
  - All facilitated by a standards from NCPDP
  - Available for clinical use
  - Constraints on research use

Other large and growing national **data bases**

- Tumor registries
  - SEER national data base in 2000 about 6 million
  - All states roughly 26 million over 15 years
- Cardiology data bases (ACC, ATS , etc) whole country
- Federal ESRD base

LONGITUDINAL RESEARCH DATA BASE

The best kind of observational data

- Collected prospectively at regular intervals
- fairly complete compared to EMR data
- Specific clinically descriptive data included
- Genetic samples and data may also be available
- Some of these are huge

.

Longitudinal Research Data Bases some examples

- Woman's health initiative (WHI)
  - > 30 years
  - > 30K+ patients
  - > 2400+ variables
  - > 600 million rows
- Framingham – 15-25 K patients depending upon phase
  - Similar in data mass to WHI

You can get to much of this data

- Through Db GAP-
  - An NLM-NCBI service
  - http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap
- Includes Framingham and many hundreds of other GWAS (Genome Wide Association Studies)
- Hand out
- You can see the details of what data was collected Down to the exact question and answer menus.
- Can request access to summary data, the patient level data and/or genetic data-

The Research Possibilities

- Epidemiology (in general)
- Early discovery of drug toxicities(Viox)
- Cost benefit and variation (Think Wennberg)
- Value of new technology, treatment
- Recruitment of patients into studies
- Longitudinal follow up
- Large espescially simple clinical trials
  - Randomize and watch the Medicare encounters and Social Security death tapes

Back Up a Little

- The data from many sources has have to be combined to answer new and important research questions

Combining data

- Just one wild example- Medicare data could provide long term outcomes for short term clinical trials and/or GWAS studies
- To get reasonably complete medication data for a population of patients, need more then SureScripts – (Medicaid, VA, other insurance carriers)
- To construct an IHE –have to combine data from lots of sources.

SO WE NEED STANDARDS

The three most important items are

- Standard patient ID (or sufficient identifying information to link them) to link data from one person from different sources.
- Standard packages for shipping data from one place to another (Message standards)
- Standard codes for variables and other things

Standard identifiers for patients

- Political forces preclude the option of universal patient identifiers – though they are the rule in most developed countries.
- Linking is accurate enough for research
- Need to preserve the option of linking and then de-identifying – Something that would simplify research.
- Vanderbilt researchers have devised a one way hashing mechanism for linking data in this way within their institutions
- Complexities abound.

STANDARD DELIVERY PACKAGES

Some back ground- Flat vs. Stacked data bases

- Typical research data base flat data structure-- one-record per *encounter* –. The field name defines the value
- Typical medical record systems use a stacked data structure with one record per *observation* – One field in the record defines the value
- HL7's OBX is a stacked data structure

Flat structure

Pat ID	Name	surgery date	Hb	DBP	# of BPU	Bypass Minutes	Choles-terol
1234-5	Doe , Jan	12May95	13	95	3	80	180
9999-3	Jones , T	1Aug95	12.5	88	2	90	230
8888-3	Doe Sam	4June95	16	78	0	80	205

Stacked structure

Operational Data Base: One Record Per Observation

Pt ID	Relevant Date	Observation ID	Value	Units	Normal Rang	Place	Observer
Doe J	12-May-95	**Hemoglobin**	13	mg/dl	12.5-15	St Francis	Dr Smith
Doe J	12-May-95	**Hemoglobin**	11.5	mg/dl	12.5-15	St Francis	Dr Smith
Doe J	12-May-95	**Dias BP**	95	mm/Hg	80-140	St Francis	Dr Smith
Doe J	12-May-95	**Dias BP**	110	mm/Hg	80-140	St Francis	Dr Smith
Doe J	13-May-95	**Bypass minutes**	80	min		St Francis	Dr Sleepwell
Doe J	12-May-95	**Cholesterol**	180			St Francis	Dr Bloodbank

HL7- the ISO shipping container for results

HL7 is a stacked structure

- So are most of the message standards
  - CDSC
  - NCPDPs prescription records

The ISO shipping container for results – example with blood count-and standard code

Patient level

PID|||0999999^6^M10||TEST^PATIENT^||1992022
5|F||B|4050 SW WAYWARD BLVD |

Order/report level t

- OBR|||H9759-0^REG_LAB|24358-4 ^Hemogram^LOINC
- Discrete Results

OBX|2|NM||789- 8^RBC^LOINC||4.9|M/mm3| 4.0-5.4
OBX|3|NM|718-7^HGB^LOINC||12.4|g/dL|12.0- 5.0||||F|
OBX|4|NM||20570-8^HCT^LOINC||50|%|35-49|H|||F|
OBX|5|NM||30428-7^MCV^LOINC||81|fL|80-94||||F|

Notice

- HL7OBX-3 the observation ID (the yellow one) is the question - (always coded
- OBX-5 is the answer
  - May be numeric (glucose)
  - May be coded (Discharge diagnosis)
  - May be text
- When the answer is a code- also need a standard code – SNOMED CT would be the choice

Google XML- “same’ content-

</Test> </Result> </Results <Text>**RBC**</Text>
Code> <Value>**789-8**</Value>
<CodingSystem>**LOINC**</CodingSystem> </Code>
</Description>
<TestResult> <Value>4.9</Value>
<Units>**M/mm3**</Units>
</TestResult> <NormalResult>RBC : 4.0 – 5.2
M/mm3
</NormalResult>
</Test> <Test> <Description> < <Text>**Hemoglobin**</Text>
Code> <Value>**718-7**</Value>
<CodingSystem>**LOINC**</CodingSystem> </Code>
</Description> <TestResult> <Value>**12.4**</Value>
<Units>**gm/dL**</Units> </TestResult>
<NormalResult>...</NormalResult>

Good News About Tapping into these Sources

- Almost every clinical system marketed to hospitals or large group practices can pump out data and do so in a standard message format –Using HL7 –
- Enables a Vulcan “Mind Meld” among clinical systems (and other systems)

The packaging Structure

- Its basically solved – for most of the clinical space
  - HL7
  - NCPDP
- Empirically - 98.5 to 99.5 of messages well formed and good.
- Syntax is not the problem
  - The 0.5% to 1.5% bad are egregious violations of crystal clear instructions

STANDARD CODES

Drug codes

- Close to done
- Rx.Norm
  - http://www.nlm.nih.gov/research/umls/rxnorm/overview.html
  - Or subset called Rx.terms
  - Available for simple download here: https://wwwcf.nlm.nih.gov/umlslicense/rxtermApp/rxTermCondition.cfm
  - Demo Medication Order Entry Tool can be tested at: http://rxterms.nlm.nih.gov:8080/

Joining data across institutions requires more

- Have to map the codes for Observations to a standard code system– so that a hemoglobin is a hemoglobin where ever it comes from--- to finish the “mind meld’
- LOINC to the recue
- It is the standard for identifying the observation (OBX-3)
- Essential for mind melds across institutions

What is LOINC

- A 57,000 record data base of universal names and codes for identifying discrete observations and for packages of those discrete observations as panels or survey instruments
  - e.g. the Glasgow Coma score
  - OASIS functional status
  - CBC
  - Standard measures for GWAS studies

LOINC Web Site

- www.regenstrief.org/loinc/:
- Or Enter LOINC in Google
  - Get right to LOINC
- Down load
  - Mapping browser program – called Relma
  - LOINC table
  - Other items

Where is LOINC required/used

- Required by
  - Federal health care systems (CHI)
  - HEDIS – quality
  - Requir4ed in HHS accepted standard HL7 message
  - CDC/ VA
- Used by
  - Available from the largest laboratory service s – e.g. Quest, LabCorp, ARUP
  - Used by major Payers (e.g. United Health)
  - Large research organizations- e.g. Partners, IU-Regenstrief. Intermountain . VA
  - Wide use internationally (more than 6 languages)

LOINC Web site

Higher level standardization

- Do clinical measures (at least for research) in one standardized way
- PhenX- NIH project to develop standard measures for data items collected in GWAS studies- https://www.phenx.org/Default.aspx?tabid=59
- At least 12 clinical dimensions being addressed
- These are being represented in LOINC
- Anthropomorphic measures is one Axis.
  - See hand out.

Caution “carve outs”

- Medicare Part A (hospitalization) data lacks Medicare HMO (??) and VA hospitalizations (3% of > 65 hospitals)
- 3% of over 65 are not in Medicare
- Not all patients sign up for Part D

Successes

- Observational data
  - Smoking and Cancer
- EMR data
  - Erythromycin and Pyloric Stenosis
    - CDC – case report
    - Regenstrief-Wishard data base
    - Medicaid TN (Wayne Ray)
- Medicaid data
  - Lots of studies (Wayne Ray)-may
- Medicare - Wennberg's magnification variation studies

PROBLEMS

Missing data –

- Clinical data is always missing some information
  - Measurements are performed when problems occcur- not at regular intervals
  - Can't reach all data that does exist
- Many well known biases

Solutions

- Pick questions that can be answered with what you have
- Use known risk factors to adjust
- Find controls that are “fair”
- Analyze the space more than one way
- Use time associations
- Perhaps, use differences in adoption by MD’s to (or by region) to compare
- Propensity scores and other sophisticated adjusmtents

We are recruiting for data base research help

- To look at a very rich- 8 year data base of intensive care admissions- mapped to SS death tapes. (50% death rate)
- To try to assess cost and utility of various intensive care interventions What difference does it really make in the short term.

Take Hope from Astronomy

- Everything is observational
- No options for controlled experiments on the cosmos
- Consider one of their tricks

Take hope from astronomey

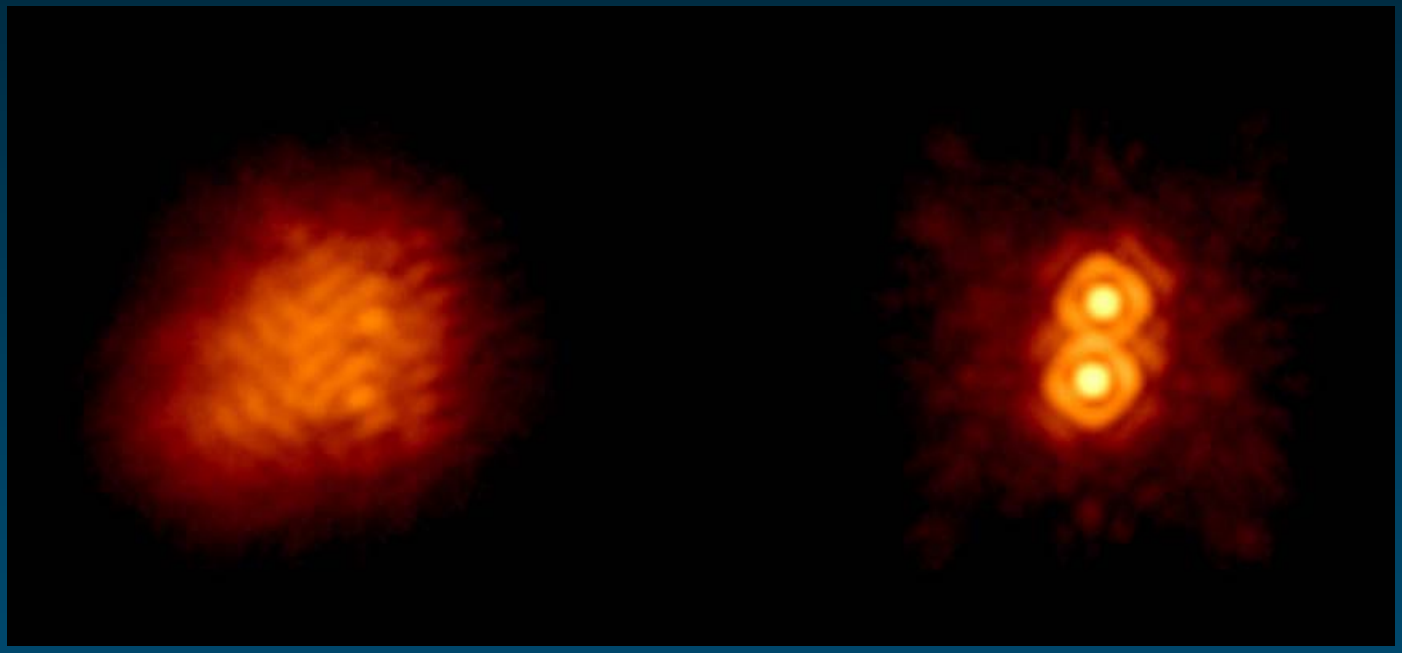
- Where everything is observational- No randomization of universes
  - Yet they have discovered the most primal phenomona .
- And use one of their tricks

Mount Palomar laser guide

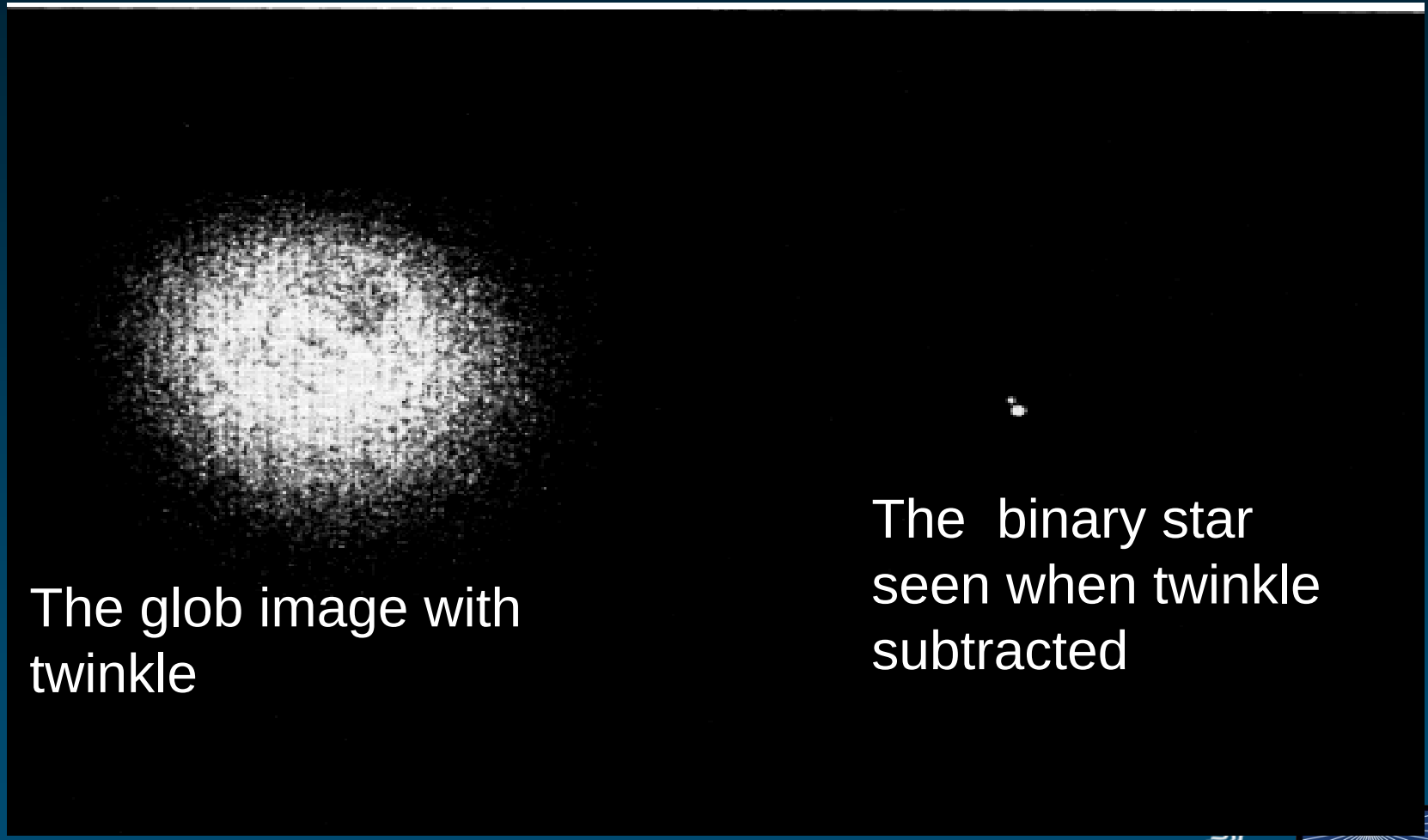
Adjust for twinkle via a reference beam

- They know everything about the wavelength and phase of this laser beam
- So can calculate exactly the twinkle (by subtracting the original from the reflected laser light)
- That same twinkle function can be subtracted from the astronomic images they are capturing simultaneously

Not one but two blurry stars



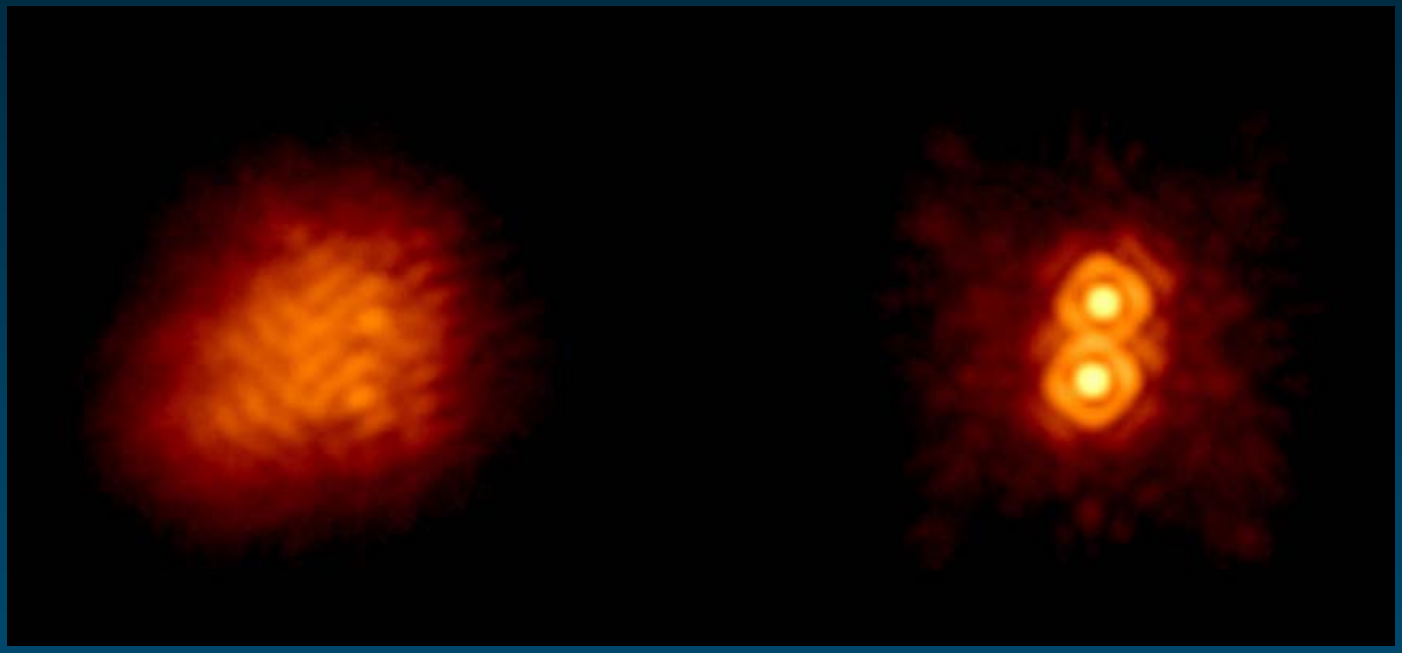
Then They See Clearly



The glob image with twinkle

The binary star seen when twinkle subtracted

Closer



Medicare data could be our laser beam

- Use medicare events data which is complete information to adjust for the richer clinical data which is incomplete