Big Data at Penn Medicine

Patient Care and Research

Brian P. Wells
Associate VP Health Technology and Academic Computing
brian.wells@uphs.upenn.edu
wellsb@mail.med.upenn.edu
Penn Medicine at a glance

Hospital of the University of Pennsylvania (HUP)
695 beds, 1595 physicians
38,213 adult admissions

Penn Presbyterian Medical Center (PPMC)
317 beds, 832 physicians
15,419 adult admissions

University of Pennsylvania School of Medicine
Founded 1765
720 students, 1006 housestaff

Clinical Practices of the University of Pennsylvania (CPUP)
1200 faculty
200+ practice locations
1.4M office visits/yr

Pennsylvania Hospital (PAH)
515 beds, 888 physicians
23,903 adult admissions

Clinical Care Associates (CCA)
250 providers
55+ practice locations
800K office visits/yr

3.8 billion operating budget
Sources of Big Data in Penn Medicine

- Patient Research Data:
  - Tissue/liquid samples, genetics, proteomics, clinical trials, PROs...

- Patient Clinical Unstructured Data:
  - Exam notes, reports (Rad, Path, GI, Derm, Neurology, Cardiology, ...), discharge summary

- Patient Clinical Structured Data:
  - Diagnostic tests, results, medications, infections

- Patient Administrative Data:
  - Demographics, billing
The Opportunity

Combine all the discrete/structured, unstructured and research data into one integrated warehouse of patient information and then utilize it to:

### Patient Care

- Improve patient quality and safety through tracking metrics and benchmarks
  - Infections
  - Falls and other incidents
  - Adverse drug events
  - Evidence based medicine
- Develop impactful clinical decision support rules that evaluate at the point of care within the EMR(s)
- Simplify medication reconciliation
- Improve financial performance

### Research

- Identify cohorts of patients for clinical trials
- Complete Genome Wide Association Studies (GWAS)
- Identify unique predictive biomarkers for specific diseases
- Discover, develop and test molecules that modify genetic and proteomic molecular processes
- Complete observational studies to identify ways to improve care
The Challenges

- **Source systems and data integration**
  - UPHS utilizes over 50 systems to run the clinical and billing aspects of the health system alone
  - HL7 is only the *structural* messaging standard
  - PSOM has many islands of research data

- **Standards**
  - Healthcare in general does not utilize *semantic* data standards beyond those required for billing
    - ICD-9
    - CPT-4
  - Clinical and research data standards are poorly adopted
  - Therefore clinical and research data exchange between disparate systems is nearly impossible
  - The USA do not have a national patient identifier!

“Standards are like toothbrushes – everyone has one but nobody wants to use yours”
Doug Fridsma, Director Office of Standards and Interoperability, ONCHIT
The Challenges

• **Volume increasing**
  - UPHS
    - 4.5 million patients
    - 42 million encounters
      - 2+ million added each year
    - 400 million orders and results
    - 40 million system-to-system messages a month across 350 unique interfaces
  - PSOM
    - 2 million samples
    - Genetic data exploding
      - Half a terabyte of data per full patient genome sequence
      - Rapidly increasing sequencing speed, accuracy and fidelity with decreasing duration and cost

• **Network speeds not following Moore’s law**
  - 10 gigabit per second max between buildings
  - 1 gigabit per second at the wall plate
  - 3 hours per terabyte on a dedicated connection
The Challenges

- **Security tightening**
  - HIPAA
    - Breach notification
  - HITECH
    - Personal liability
    - Fines, imprisonment
  - GINA
    - Needs strengthening
  - Full de-identification of unstructured data requires manual review

- **Science**
  - Microbiome sequencing
  - New genetic biomarkers constantly being discovered

- **Liability / Ethics**
  - If we “know” your entire genetic profile must we notify you when a new marker is discovered that you already possess?
  - Must we notify your offspring? Parents? Can you sue us if we fail to?
What is Penn Medicine Actually Doing?

- **Penn Data Store**
  - Financial
  - Clinical
  - 2 to 3 billion rows of structured information
  - Dashboards and reports
    - Financial
    - Clinical quality
    - Patient satisfaction
    - Research requests
What is Penn Medicine Actually Doing?

- **Penn Data Store**
  - Financial
  - Clinical
  - 2 to 3 billion rows of structured information

- **Work in progress**
  - High Performance Computing
    - New local, $3 million cluster
    - 1024 high memory cores
    - 1 petabyte of spinning disk
    - 3 petabytes of tape archive
  - Best Practice Advisories
    - Clinical trial recruiting
    - Clinical care alerts
  - Predictive analytics
    - Sepsis “sniffer”
  - Unstructured text mining
    - 15 million documents and counting
What is Penn Medicine Actually Doing?

**Future projects**

- Bio-bank operational system (LIMS)
- Simulation
  - What happens if …
- Research section of Penn Data Store
  - Genetic data
  - Bio-bank
  - Tumor Registry
  - Outcomes
  - Use cases
    - Find patients that are poor responders for drug Y and have a mutation in the promoter region of Gene X
    - What is the expression level of TP53 mutants by cancer tissue?
    - How many patients have disease Z, responded to treatment, have a chromosome 18 deletion and have blood samples in the bio-bank?
    - Mine the breast and ovarian TCGA data for the somatic mutation data associated with tumors with germline BRCA1 and BRCA2 mutation
Penn Data Store Expansion

**Penn Data Store - Clinical**

- Patient 4 million
- Encounter 40 million
- Dx & Proc
- Orders & Results 400 million
- Inpatient Infection 50 thousand

**Research Source Systems**

- Oracle Clinical
- VELOS
- LIMS
- ACCARD
- Tumor Registry
- REDCap
- Omics

**Extract, Transform and Load (ETL) Process**

- Honest Broker
  - Linkage of sample to patient
  - De-identification
  - Access logging

**Penn Data Store - Research**

- Subject
- Visit
- Protocol
- Treatment
- Outcome
- Omics

**INFORMATION**

- Dashboards
- Query tools
- Extracts
- Alerts
- Cohort Identification
- Genetics Analytics
Questions?