Crossing the Chasm of Semantic Despair: Integrating Knowledge and Data from Science and Clinical Practice

CG Chute, Johns Hopkins

Clinical Research Forum – IT Roundtable
Washington, DC
2 Nov 2017
On the Nature of Translational Research

“Translational science is the field of investigation focused on understanding the scientific and operational principles underlying each step of the translational process.” - NCATS web page

• Dimensions of Translation
  – Discovery
  – Dissemination
  – Acceptance
  – Adoption
  – Implementation
Review Paper

Managing Clinical Knowledge for Health Care Improvement

E.A. Balas, S.A. Boren
Center for Health Care Quality, University of Missouri, Columbia, MO, USA

Quality health care rests fundamentally on the achievements of biomedical research. All health outcomes are improved by sound science: health status can be turned around by transplantation when someone's life is in jeopardy due to a diseased organ; social functioning can be improved by shock wave lithotripsy that leads to faster recovery; and satisfaction can be enhanced when children with moderate or severe asthma receive appropriate anti-inflammatory treatment. To improve the quality of health care that patients actually receive, both biomedical research production and especially its introduction into clinical practice need to be examined.

Growth of Clinical Research Production

Over the past 20 years, the number of articles indexed annually in the Medline database of the National Library of Medicine has nearly doubled (Table I). Certainly, the achievements of the Human Genome Project, innovative medical technologies, and scientific discoveries make further investment in biomedical research appealing.

The growth in publications is particularly spectacular in the category of the most rigorous clinical evaluations, randomized controlled clinical trials. Such trials have long been considered sources of the highest quality evidence on the value of a new clinical intervention. Over the past two decades, the number of clinical trials in cardiology has increased five-fold. Similar growth has occurred in many other clinical specialty areas (Table I). Improvement in the quality and efficiency of health care also depends on progress in the science of organizational, reimbursement, workforce, and information system issues. Correspondingly, 10 times more clinical trials are published today than 20 years ago in health services research (e.g., comparisons of inpatient care with outpatient care, physician profiling, and other information interventions). Yet, health care practices appear to be ill prepared to absorb and efficiently introduce this constantly growing amount of information.

Slow Transfer of Research to Practice

In 1843 before the Boston Society for Medical Improvement, Oliver Wendell Holmes read the first of his famous papers on the “Contagiousness of Puerperal Fever” [3]. It advocated hand washing before examining a pregnant woman— a revolutionary idea at the time. Yet, it took decades for his recommendation to become a universally accepted practice and change did not come without resistance. Today, more scientific discoveries are being achieved than ever.
Research translation takes 17 years, the message takes longer...

Managing clinical knowledge for health care improvement

Authors: E Andrew Balas, Suzanne A Boren
Publication date: 2000/1/1
Journal: Yearbook of medical informatics 2000: Patient-centered systems
Description: The authors assess the growth in clinical research studies coupled with a slow rate of adoption, often taking an average of 17 years for new evidence-based findings to reach clinical practice. Some issues discussed include the need for a more efficient information infrastructure to better connect front-line professionals with the research community, difficulties in translating research into practice, and an inadequate system to help health care clinicians evaluate the strength of new study findings.

Total citations: Cited by 854

Scholar articles:
Managing clinical knowledge for health care improvement
Cited by 854  Related articles  All 11 versions
The Continuum Of Biomedical Informatics
Bioinformatics meets Medical Informatics

The Chasm of Semantic Despair
Origins of Big Science: Astronomy

[Historical text and image of Galileo Galilei]
Sloan Digital Sky Survey III – DR9

- Total area of imaging: 31,637 square degrees
- Image field size: 1361x2048 pixels
- Number fields: 938,046 (excluding supernovae runs)
- Catalog objects: 1,231,051,050
  - Number of unique, primary sources:
    - Total: 469,053,874
    - Stars: 260,562,744
    - Galaxies: 208,478,448
    - Unknown: 12,682

- Images
- Spectra
- Object catalog
- Metadata
Rutherford “Table-top” Experiment
That Higgs Boson

- 600 institutions
- 10,000 scientists
- 800 trillion collisions
- 200 PB of data =
  - $2 \times 10^{17}$ bytes of data

Boarding on an astronomical number in its own right!

- $13.25B USD
Biology and Medicine have Entered the Era of Big Science

• *Kicking and screaming….*

• The era of lone research discovery is over

• We are in a “big data”, transdisciplinary world

• The interdependence of
  – Large-scale data
  – Large-scale information
  – Large-scale knowledge resources
Some Fashionable Biology Databases This Week

- Genome
- PDB
- dbSNP
- EMBL-EBI
- UniProt
- Panther Classification System
- BioModels.Net
- Catalytic Site Atlas
- PharmGKB
- Nucleotide
Some Fashionable Clinical Databases This Week

- USHIK
- OMIM
- i2b2
- NCHS
- CDC
- NDI
- WHO
- ICD-11
- HCUP
- Sentinel Initiative
- ClinicalTrials.gov
- CDISC
- Orphanet
- DrugBank
- AHRQ
- dbGaP
Some Integration Resources
Consequences of the Chasm: Impedance to Team Science

• Connection and relation of data made tedious
• Integration of databases cannot be fully automated
• Connection of knowledge to data impaired
• Integration for Big Science and Analytics stymied

One can’t put all the pieces together
• Translation of basic science to practice impaired
From Practice-based Evidence to Evidence-based Practice

- Clinical Databases
- Registries et al.

Data

Inference

Comparability and Consistency

Standards

Terminologies & Data Models

Patient Encounters

Decision support

Expert Systems

Clinical Guidelines

Foundations for Learning Health System

Medical Knowledge

Knowledge Management
From Research-based Evidence to Evidence-based Practice

Data → Clinical Databases → Registries et al.

Inference

Comparability and Consistency

Standards

Terminologies & Data Models

Patient Encounters

Decision support

Expert Systems → Clinical Guidelines

Medical Knowledge

Knowledge Management

Foundations for Precision Medicine
Road to NCATS Translator
TransMed Connector Bridging the Chasm

[Graph showing the transition from Biology to Medicine across various research fields]

- Genomics
- Proteomics
- Metabolic Pathways
- Molecular Modeling
- Cellular Models
- Molecular Assays
- Genomic Testing
- Biospecimens
- Lab Data
- Trials Data
- Disease & Syndrome
- Medical Imaging
- Patient Record Data
- EHR Structures

Biology (Orange) vs. Medicine (Green)
## NCATS Translator Program

<table>
<thead>
<tr>
<th>Participating Organization(s)</th>
<th>National Institutes of Health (NIH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Components of Participating Organizations</td>
<td>National Center for Advancing Translational Sciences (NCATS)</td>
</tr>
<tr>
<td>Funding Opportunity Title</td>
<td>Biomedical Data Translator: Technical Feasibility Assessment and Architecture Design Projects (OT3)</td>
</tr>
<tr>
<td>Activity Code</td>
<td>OT3 Multi-component Research Project - Other Transaction Award</td>
</tr>
<tr>
<td>Announcement Type</td>
<td>New</td>
</tr>
<tr>
<td>Related Notices</td>
<td>N/A</td>
</tr>
<tr>
<td>Funding Opportunity Announcement (FOA) Number</td>
<td>OT-TR-16-001</td>
</tr>
</tbody>
</table>

June, 2016
Building a Translational Bridge

Tools

Ontologies

Algorithms

Data

Standards

Basic
- Genomics
- Proteomics
- Metabolic Pathways
- Molecular Modeling
- Molecular Simulation
- Cellular Models
- Molecular Assays

Clinical
- Biospecimens
- Lab Data
- Trials Data
- Disease & Syndrome
- Medical Imaging
- EHR Structures
- Patient Record Data
Deep open source experience, each step of the way

Leadership in existing projects to be leveraged in the translator

1. Extraction of knowledge from diverse sources
2. Data & knowledge integration, algorithms & tools
3. Dissemination / deployment

Leadership in other projects

Contribution in other projects

- VIVO
- OBO Foundry
- BioPerl
- Bioconductor
- Global Alliance for Genomics & Health
- SHARPn.org
- ICD-11
- EXOMISER
- dipper
- monarch initiative
- WIKIDATA
- phenomizer
- EVS
- cAKES

- bioL
- BioPortal
- GMOD
- WETWARE
- CI4CC
- gatk
(1) Extraction of knowledge from diverse sources

- Community Ontologies
- Structured Data
- Omics Databases
- Literature

- Dynamic Ontology
- ETL
- Metadata Registration
- Text Mining

- Tools to integrate vocabularies
- Tools to identify equivalencies:
  - Identifier and synonym alignment
  - Conceptual alignment based upon logic determinations and prior probabilities
- Text mining for clinical concepts and pathway fragments
Data & knowledge integration, algorithms & tools

GRAPH INFERENCEx
- OWL-based reasoning over large graphs
- BELpathway ‘chains of causation’
- Probabilistic inference across disease-phenotype-gene
  - Bayesian Ontology Query Algorithm (Boqa)

GRAPH QUERY
- Query related entities within/across species, sources
- Query similar sets of entities (OWLsim)
  - Sets of phenotypes
  - Expression patterns
  - Pathway modules
(3) Dissemination, deployment & validation

EVALUATION

- Purpose built for clinicians, researchers, & bioinformaticians
- TransMed outputs based on condition set and competency questions are iteratively compared against current phenomenological nosology
- Comparisons of integrated delivery of mechanistic modules against single source inquiries
- T-Score measuring connectivity of clinical and basic sources
Monarch (via Dipper) ingest so far in TransMed

Common mechanistic underpinnings of rare & common/complex disease

<table>
<thead>
<tr>
<th>Rare, Beneficial</th>
<th>Common, Subtle</th>
<th>Rare, Catastrophic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beneficial</td>
<td>Neutral</td>
<td>Detrimental</td>
</tr>
<tr>
<td>Acetaldehyde metabolism mutations: 500 Million People Affected</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fanconi Anemia: 1 in 160,000 individuals worldwide</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALDH KO: ?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Fanconi Anemia mechanism

**Exposures**
- FA or ALDH variants
- Normal FA / ALDH

**Defective**
- ALDH: More damage
- More reactive aldehyde

**Normal**
- ALDH: Less damage
- Less reactive aldehyde

**Mechanism**
- Defective repair
- Sustained damage

**Outcomes**
- More, earlier solid tumors
- Bone marrow failure
- More and earlier leukemia
- Developmental Abnormalities
Fanconi Phenotypes

Adapted from Garaycoechea et al., 2012
1) Environmental risk factors

- Alcohol
- Smoking
- Fermented foods
- Proximity to aldehyde dumping site

2) Genes (22 implicated)

3) Phenotypic outcomes

(bit.ly/fanconi-cq)
How can we generalize the TransMed Pilot to the community?

Embrace the Center for Data to Health CD2H
How do we achieve this?

We collectively have a remarkable opportunity to build upon the excellence of the CTSA Program to innovate, pilot, implement, and sustain new informatics technologies to improve human health.
Each CTSA is a workbench

1: Harmonize the data ecosystem
2: Realize a software tool ecosystem
3: Synthesize a people ecosystem
4: Catalyze technical and cultural evolution
Aim 1: Harmonize the data ecosystem

✓ Goldilocks solutions to data dissonance; development of Good Data Practice (GDP); promotion of software standards for interoperability

✓ Apply standards to harmonize and integrate data

✓ Improve navigation of the data sharing odyssey
Aim 2: Realize a software tool ecosystem

✓ Assemble communities of practice
✓ Compile software best practices and standards
✓ Configure Synapse Platform for data sharing and provide training
✓ Deploy CIELO: a software “storefront”
✓ Verify & validate existing/emergent software
Aim 3: Synthesize a people ecosystem

- Extend representation of expertise and services across the CTSA consortium
- Enable discovery of services/data/tools/people across the CTSA consortium
Aim 4: Catalyze technical & cultural evolution

Demonstration disease areas:

✓ Rare disease and eliminating the diagnostic odyssey
✓ Improve the impact of clinical research across the lifespan
The translational workforce: It takes a village to solve disease

Clinicians/care team, Pathologists, Ontologists, Informaticians, Curators, Basic Research
Synergistic Informatics Education

**Informatics & data science competencies**
Mapping across current competency efforts (AMIA, CTSAs, ISCB, AMSTAT, etc)

**Modular, dynamic training library**
Leveraging existing materials (BD2K, AMIA, ONC etc)

**Module annotations**
Improve discovery/reuse coordinate BD2K erudite, others

**Interdisciplinary collaboration, open science, & entrepreneurship**
Partnerships with CTSA Workforce Development, AMIA Informatics Educators and others

**Mentoring network**
Synergize with NIH, BD2K Data Science RoAD-Trip, AMIA Mentor Match etc.

**“Packaging” of Learning Modules into LMS**
Collaboration with N-lighten to get training modules into learning management systems (LMS)
Opportunities for CTSA Analytics & Data

**Optimize dissemination**
Dissemination and impact are inherently linked; build on best practices for a range of outputs

**Attribution**
Enable an attribution framework for a range of products & processes for the translational workforce

**Modular, dynamic evaluation library**
Support local hubs in data-driven CQI through improved data, best practices for analytics and data viz

**Accountability**
Facilitate innovative accountability through open project dashboards

**Peer review**
Enable consortium-wide peer review, (NITRO-Competitions)

**Collaborative Innovation**
Develop a model to support and grow CI across the CTSA Program
Continuing the Conversation

- Workgroups & collaborations
- Socio-technical communication environments
- Listservs, project email, and activities calendar
- We would love to come and visit your team!
- Interested in learning more? Please fill out the intake form ([bit.ly/cd2h-kickoff](http://bit.ly/cd2h-kickoff)) or email us at data2health@gmail.com

@Data2Health
Observations

• Biomedical databases are hugely rich and growing
• Information and knowledge remains latent
• Integration of these resources is a necessary first step
• Query and exploration frameworks follow
• TransMed is one of the pilot efforts in NCATS Translator to develop and demonstrate infrastructure
• CD2H can help generalize this across CTSAs