

Challenges of Data Quality in Medical Informatics Data warehouses

ABSTRACT-----

Efforts are currently underway to integrate both historical and current patient data (clinical, histopathological, billing etc.) along with imaging and genotyping data into the data warehouse to improve healthcare delivery. Successful implementation of the data warehouse will play a critical role in preventive medicine, drug discovery and targeted therapeutics.

The success of this enterprise wide data warehouse is critically dependent on successfully integrating data from various sources into a single warehouse that has excellent data quality. This presentation discusses the sources of data error and challenges in integrating the data into a single warehouse. The potential pathways of errors (source system errors, data abstraction) and challenges (non standard representation, patient reported data) in developing rules to capture and represent data in a reliable, interpretable form are also discussed. These issues are highlighted with real world examples. Potential solutions to some of the challenges are also discussed.

BIOGRAPHY-----

Neera Bhansali, PhD

Manager of Data Quality & Curation
Moffitt Cancer Center

Neera Bhansali received her BA from Calcutta University, India; MB and PhD degrees in Business from RMIT University, Australia. She has served in diverse roles in information systems and business divisions in manufacturing, airlines, consulting, media, finance and healthcare sector around the globe. Currently she is with H. Lee Moffitt Cancer Center & Research Institute (an NCI designated comprehensive cancer center). Her areas of expertise are strategic planning, data warehousing, data governance and strategic alignment.



Challenges of Data Quality In Medical Informatics Data Warehouses

Neera Bhansali, PhD
Dept of Biomedical Informatics
Moffitt Cancer Center & Research Institute
12902 Magnolia Drive
Tampa 33612

Neera Bhansali

MIT 2009 IQIS Symposium

1



Neera Bhansali

MIT 2009 IQIS Symposium

2

Moffitt Cancer Center

Mission

To contribute to the prevention & cure of cancer

Vision

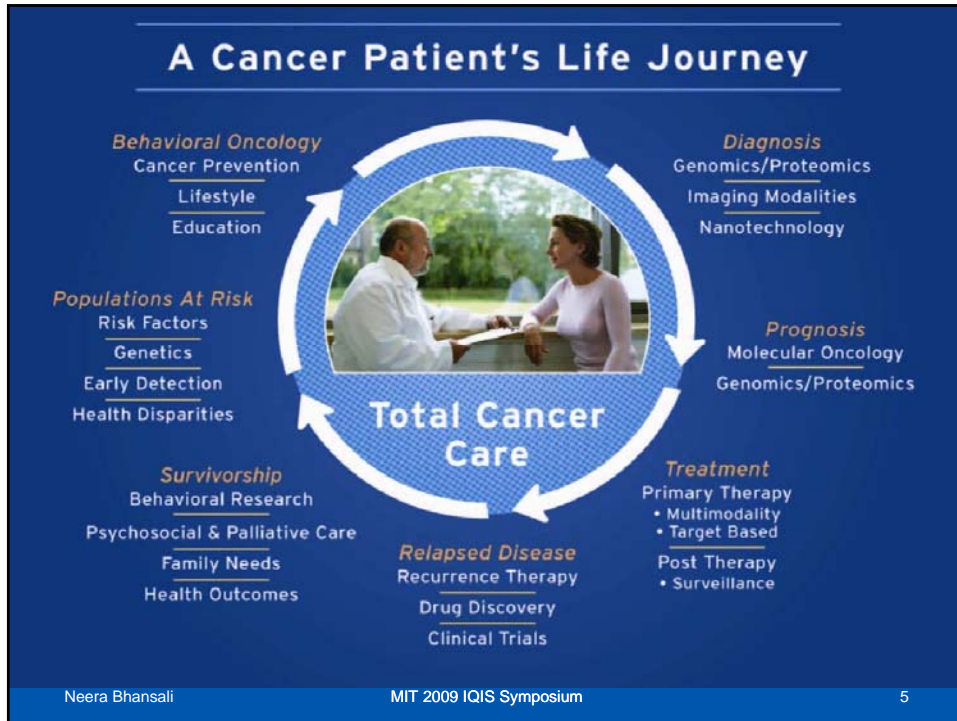
To be the leader in scientific discovery and translation into compassionate care, cures, and prevention of cancer for our community and the world.

The slide features the Moffitt Cancer Center logo in the top left, consisting of the text 'MOFFITT CANCER CENTER' and a stylized 'M' icon. Below the logo is the slogan 'TOTAL CANCER CARE™'. The background is a dark blue gradient with a glowing network of yellow lines and nodes, overlaid on a faint map of the United States. The network lines radiate from a central point in the bottom right, connecting to various locations across the country.

Neera Bhansali

MIT 2009 IQIS Symposium

4



Total Cancer Care

Moffitt's comprehensive approach to cancer care and research: Focuses on the individual's needs throughout their lifetime

Neera Bhansali MIT 2009 IQIS Symposium 6

Total Cancer Care Goals

- Identify the needs of the individual patient
- Identify markers to predict needs and risks
- Develop methods of early detection
- Match the right treatment for the right patient
- Improve the performance of clinical trials through molecular profiling and matching
- Create evidence-based guidelines
- Raise the standard of care

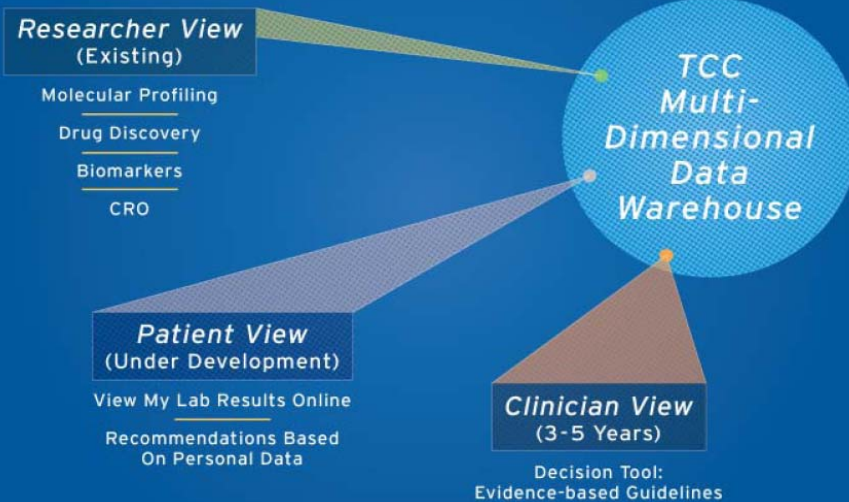
The Building Blocks Of Total Cancer Care™



Approach: Total Cancer Care Personalized Medicine Project (PCC)

- A large prospective translational research project with patient consent
- Collect, relate, and interpret clinical data and molecular data from thousands of patients
 - Tumor, blood, urine samples
 - Clinical data (risk factors, therapies, outcomes)
- Identify molecular signatures for prognosis
- Personalize therapy and follow-up
 - Right care, right patient
- Information->evidence->knowledge->wisdom

Three Portals To TCC



Effectiveness of enterprise wide data warehouse is reduced by poor data quality

Impact of Poor Data Quality

Scientific, educational, and patient care

- Risk of flawed publication based on poor data
- Conclusions that cannot be reproduced/validated clinically
- Risk of wrong diagnosis and treatment of patients
- Risk of using wrong genetic signature to treat patients

Institutional

- Bad business decisions based on incomplete and/or misleading data
- Adversely impacts institutional reputation

Benefits of High Data Quality

- Trust in the data
 - Usable, value added, relevant
 - Available
 - Contextual, Interpretable
 - Consistent, standard representation
- Supports and enhances institutional objectives
 - Cancer care delivery
 - Research
 - Education
 - Reputation

High Data Quality

- Accuracy
- Completeness
- Consistency
- Timeliness
- Relevance
- Accessibility

Data Quality Challenges

- Complexity of data (multiple subject areas)
- Divergent business objectives and needs
- Self reported data (by patients)
- Source data dispersed over multiple systems
- Data integration challenges

Multiple subject areas

- Treatment data
- Research data
- Clinical Trials data
- Imaging data
- Financial data



Multiple Types Data

- Paper Charts
- Electronic Medical Records
- Images



Typical Data Quality Problems

Any difficulty encountered along one or more quality dimensions

- Intrinsic
- Accessibility
- Contextual
- Representational

Example: Gender

- Expected values of Male, Female
- Incomplete data - nulls, blanks, not specified
- Invalid data
- Inconsistent data across different source systems - M/F, 0/1
- Inconsistent data element name – gender, sex
- Different requirements – hermaphrodite, transsexual

Example: Gender

- Errors due to data capture and data entry at source systems
- Inaccuracies due to data transfer and data integration approaches
- Different permissible gender values
- Different aliases

Example: Race

- Race / Ethnicity / Nationality
- Patient self reported
- Different requirements for
 - Billing and Regulatory Filings
 - Cancer Registry
 - NCI clinical trials
- Collected differently in different systems
e.g. Hospitals, surveys
- Multi-racial

Example: Race

- Source system accuracy
- Non standard Race categories across source systems
- Data capture forms following different standards
- Free Form entry
- Multiple medical record numbers for same individual

Example: Race

- Multiple sources of same data
 - Different formats, different platforms, different requirements
- Completeness
 - Not capturing all required data discretely
- Consistency
 - Different standards of data abstractions
- Questionable Objectivity
 - Self reported data

DQ Solutions

- Control techniques
 - Double entry format
 - Validation at data entry: Logic Checks, If-then checks, Range checks, Zero-control test
 - Probabilistic tests – atypical observations, detecting outliers
- Development of standardized definitions, representations and permissible values for the data elements

DQ Solutions (contd.)

- Improve interpretability - standard semantics, aliases
- Database integrity constraints, update controls
- Integration of data – value add

Implications

- To solve data quality problems effectively, entire range of data concerns must be addressed
- As users evaluate data quality relative to their requirements, providing high quality data implies an ever-moving target