

Big Data and oncology: value and limitations

Philippe Aftimos, MD

Clinical Trials Development Leader Clinical Trials Conduct Unit (CTCU) Institut Jules Bordet







Disclosure information

- Consulting:
 - Boehringer Ingelheim, Macrogenics, Roche, Novartis, Amcure, Servier, G1 Therapeutics, Radius, Deloitte
- Honoraria:
 - Synthon, Amgen, Novartis, Gilead
- Travel grants:
 - . Amgen, MSD, Pfizer, Roche
- Research funding to my institution:
 - . Roche





What is Big Data?

Big data is high-volume, highvelocity and/or high-variety information assets that demand costeffective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

Gartner glossary (circa 2001)



Credit: Ben Chams Fotolia





Big data sources in the treatment of cancer (not comprehensive)







OYour EMR		Palania Brand Sparity 140
Received and the second	Comparing a second	Connext Multication
Pares Profile India Augusta Reg Reg Reg Reg Reg Reg Reg Reg Reg Reg	Prescribe Q	













It's not that simple

Regina Barzilay, an artificial intelligence researcher at MIT who was diagnosed with breast cancer in 2014 at the age of 43, was shocked by the paltry amount of data upon which her doctors based their clinical decision.



Credit: Sam Ogden (DFCI)





Bridging the gap







Leveraging big data in cancer care

- Early diagnosis and prevention
- Digital pathology and molecular pathology
- Drug discovery
- Treatment decisions
- Matching patients to clinical trials
- Anticipating adverse events







What is AI?

- Al refers to the ability of a machine to perform tasks commonly associated with intelligent human behavior.
- AI can be considered a group of iterative, "self-learning" techniques, which discover relationships within data that can evolve and often be performed faster over time.



Russel SJ et al. 2003.



Machine Learning (ML) vs Deep Learning (DL)

• Machine Learning:

- ML algorithms, exposed to training data, are able to appreciate hidden patterns within the data which can then be used to perform a task without explicit programming
- ML tasks are often broadly dichotomized into supervised or unsupervised learning

Deep Learning:

 DL is a form of ML that uses layered "artificial neural networks" to develop sophisticated models with the ability to understand data at different levels of abstraction



Samuel AL. IBM J Res Dev. 1959;3:210-29. LeCun Y, et al. Nature. 2015;521:436-44.



Natural language processing (NLP)

- NLP is any computer-based algorithm that handles, augments, and transforms natural language so that it can be represented for computation
- This technology can harvest important clinical variables trapped in the freetext narratives within electronic medical records







ASCO CancerLinQ

Since 2015, More than **100** organizations **1,426,015** patients with a primary cancer diagnosis.



TABLE 1. Overview of CancerLinQ Quality Data Model Data Model Version 1.6.1

Data Type	Table Name	Contents						
Demographics	Patient	Demographics						
Diagnoses	diagnosis_active	Diagnoses-all types, problem list, adverse events						
Pathology	procedure_performed	Staging, histology, laterality, tumor invasion, surgical margin, tumor size from pathology						
Patient assessments	functional_status_performed	Performance status, therapeutic response, tumor progression						
Risk assessments	risk_assessment	Mental health, pain, and smoking screening, family history						
Physical exam values	physical_exam_performed	Physical examination values for the patient						
Laboratory tests	laboratory_test_performed	Laboratory tests, including genetic tests						
Care plan	care_plan	Treatment intent						
Encounters	encounter_performed	Encounter codes (CPT, SNOMED, HCPCS)						
Interventions	intervention_performed	Indication data of smoking cessation counseling, genetic counseling, chemotherapy, hospice, etc						
Medications (administered)	medication_administered	Medication administered to the patient						
Medications (ordered)	medication_ordered	Medication ordered for the patient						
Radiation therapy	radiation_care_plan	Abstracted summary radiation therapy information, including type, location, dose, and total fractions						
Radiation therapy	procedure_radiation	Transactional radiotherapy information, where available						
Surgery	procedure_surgery	Abstracted summary surgical information						
Imaging results	procedure_imaging	Abstracted imaging activities, including size of tumor and progression						

Abbreviations: CPT, Current Procedural Terminology; HCPCS, Healthcare Common Procedure Coding System; SNOMED, Systemized Nomenclature of Medicine Clinical Terms.



Potter D et al. JCO CCI 2020.



CancerLinQ & Home Lo SmartLinQ L Patients S Insights -		CancerLinQ Health System 🎝 👻 Log out [6
SmartLinQ TM All Measures Dashboard	V [All Organization Units (3)] V All Providers (5) V	
Chemotherapy intent (curative vs. non-curative) documented before or within two weeks after administration	HER2/neu Testing for Breast Cancer Patients	Chemotherapy Administered to Patients with Metastatic Solid Tumor with Performance Status of ECOG 3 or 4; KPS 10 - 40; or Undocumented
Staging Documented within One Month of First Office Visit	Opioid Therapy Follow-up Evaluation 51*5 *** Numerator / Decompany 1464 / 2844 *** N.A	KRAS Gene Testing for Metastatic Colorectal Patients
Hepatitis B Testing Prior to Rituximab Administration for NHL	Percentage of patients with initial AJCC stage IV or distant metastatic Lung Cancer whose performance status is documented.	Appropriate Antiemetic Therapy for High- and Moderate-Emetic-Risk Antineoplastic Agents View 1964 / 2717 Actionatie N/A



Potter D et al. JCO CCI 2020.



Standardization efforts for genomic data (Minimum Variant Level Data MVLD)







Genomics databases and selected tools for tertiary analyses





CTGATGGTATGGGGCCAAGAGATA AGGTACGGCTGTCATCACTTAGAC AGGGCTGGGATAAAGTCAGGGC/ CATGGTGCATCTGACTCCTGAGGA CAGGTTGGTATCAAGGTTACAAGA CAGTTGACTCTCTCTGCCTATTGG



dbSNP

dbSNP contains human single nucleolide variations, microsatellites, and small-scale insertions and deletions along with publication, population frequency, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations.

PolyPhen-2 prediction of functional effects of human nsSNPs





UniProt



MY CANCER GENOME ® GENETICALLY INFORMED CANCER MEDICINE

H U.S. National Library of Medicine

ClinicalTrials.gov







FINDING CURES TOGETHER"





AACR PROJECT GENIE: PUBLICATIONS

Linked Entity Attribute Pair (LEAP): A Harmonization Framework for Data Pooling Journal of Clinical Oncology, July 2020

PROJECTGENIE

Genomics Evidence Neoplasia Information Exchange

Characteristics and Outcome of AKTIE17K-Mutant Breast Cancer Defined through AACR Project GENIE, a Clinicogenomic Registry *Cancer Discovery*, Vol. 10, Issue 4, 2020, April, 2020

American Association for Cancer Research Project Genomics Evidence Neoplasia Information Exchange: From Inception to First Data Release and Beyond—Lessons Learned and Member Institutions' Perspectives Journal of Clinical Oncology, February 16, 2018

AACR Project GENIE: Powering Precision Medicine through an Internal Consortium Cancer Discovery, Vol. 7, Issue 8, August 2017





Data Sets Web API R/MATLAB Tutorials FAQ News Visualize Your Data About

The Metastatic Breast Cancer Project (Provisional, October 2017) Query this study

The Metastatic Breast Cancer Project is a patient-driven initiative. This study includes genomic data, patient-reported data (pre-pended as PRD), medical record data (MedR), and pathology report data (PATH). All of the titles and descriptive text for the clinical data elements have been finalized in partnership with numerous patients in the project. As these data were generated in a research, not a clinical, laboratory, they are for research purposes only and cannot be used to inform clinical decision-making. All annotations have been de-identified. More information is available at www.mbcproject.org. Duestions about these data can be directed to data@mbcoriect.org.

Selected: 103 samples / 78 patients (2) (4)		ary genes - click to expand				Select cas	Add Chart							
		# - Freq		Mutated Genes (103 Gene # Mut		profiled samples)		0 초 ÷ × Freq	CNA Genes		(103 profiled samples) CNA #			✓ Freq
Breast Invasive Ductal	67	0	65.05%	TP53(9)	30	30		29.13%	CCND16	11q13.3	AMP	38	0	36.89%
Invasive Breast Carcinoma	16		15.53%	PIK3CA 9	29	27		26.21%	NDRG1	8q24.22	AMP	34		33.01%
Breast Mixed Ductal and	11	0	10.68%	TTN	27	24	0	23.30%	FGF19	11q13.3	AMP	33	0	32.04%
Breast Invasive Lobular	9		8.74%	OBSCN	13	12	0	11.65%	ERBB2	17q12	AMP	32		31.07%
Search				RYR1	11	10	0	9.71%	FGF4	11q13.3	AMP	31	0	30.10%
				CDH1	9	9		8.74%	FGF3	11q13.3	AMP	31		30.10%
				KIAA1109	10	9	0	8.74%	IKZF3	17q12-q	AMP	31		30.10%
				KMT2C	9	9	0	8.74%	CDK12	17q12	AMP	29	0	28.16%
				ACAN	8	8		7.77%	RAD21	8q24.11	AMP	29		28.16%
				SYNE1	8	7	0	6.80%	MUC1©	1022	AMP	28		27.18%
				АРОВ	7	7	0	6.80%	PDPK1	16p13.3	AMP	27	0	26.21%
				Search					Search					
MedR Age at Diagnosis	With	103	lon Data	40 - 30 - 20 - 10 - 0 - wedR T	20 40 6	0 80 100 c Diagnosis (1	120 Calculate	>120 NA	MedR St	age at Diagnosis			83	e Location
With CNA Data	# of Sa	mples I	Per Patient						PATH S	ample Histology		PATH	Sampl	e Grade
103		19	57	30 - 25 - 20 - 15 - 10 - 5 - 0 - <=20	20 40 60 1	80 100 120	140 16	0 180 >180		11 12 67			20	37





AURORA: Aiming to Understand the Molecular Aberrations in Metastatic Breast Cancer





Patient 1345 - pathology images







Aftimos P et al. Manuscript submitted.



Applications in the diagnosis and treatment of cancer



Microsoft researchers detect lungcancer risks in web search logs



Eric Horvitz, technical fellow and managing director of Microsoft's research lab in Redmond, Washington, says search queries may be an early warning of lung cancer. (Photography by Scott Eklund/Red Box Pictures)



ONLINE FIRST

November 10, 2016

Evaluation of the Feasibility of Screening Patients for Early Signs of Lung Carcinoma in Web Search Logs

Ryen W. White, PhD1; Eric Horvitz, MD, PhD1

» Author Affiliations

JAMA Oncol. Published online November 10, 2016. doi:10.1001/jamaoncol.2016.4911







Detecting and classifying breast lesions using a deep convolutional neural network







Ribli D et al. Sci Rep 2018.

A radiomics signature predicting CD8+ TILs and response to CPIs





Sun R et al. Lancet Oncol 2018.



Histopathology reveals driver mutations (but also tumor composition, gene expression and prognosis)





Fu Y et al. Nature Cancer 2020.



A deep learning tool to predict outcome of clinical trials









lalte

Artemov AV et al. December 29, 2016. <u>https://www.biorxiv.org/content/10.1101/095653v2</u>.

MatchMiner

Developed at Dana Farber Cancer Institute and now Open Source

MatchMiner **Clinical Trial Investigator Mode** Patient Patient Trial Patient Patien **Oncologist Mode** Trial Trial Patient Trial







Georgetown Immuno-Oncology Registry









Courtesy Subha Madhavan



Predicting irAE

Can we predict irAE using a machine learning

irAE Risk Calculator

Can we build a risk

calculator for

irAE?

IO Agent Nicolumab Number of doses 2 Current Age 0 Race 1 Bisck/African American 0 Other 1 Bisck/African American 0 Other 1 White/Caucasian Age at Disgnosis 0 Initial Clinical Stage Initial Clinical Stage EGCG status

0





Courtesy Subha Madhavan



Limitations

Proving generalizability and real-world applications

Data access and equity

Interpretability and the black box problem

Education and expertise





mCODE[™]: Minimal Common Oncology Data Elements

The Initiative to Create a Core Cancer Model and Foundational EHR Data Elements

ISTITUT



iris



Annals of Oncology 2020. Article in press.









Credit: Andre Kahles, Gunnar Rätsch, Chris Sander



The Institut Jules Bordet Clinical Trials Conduct Unit (CTCU)



Let's interact...





