

Cancer Deep Phenotype Extraction from Electronic Medical Records (DeepPhe)

Guergana K. Savova, PhD

Guergana.Savova@childrens.harvard.edu

Associate Professor
Boston Children's Hospital
Harvard Medical School

Rebecca Jacobson, MD, MS

rebeccaj@pitt.edu

Professor
Department of Biomedical Informatics
University of Pittsburgh Cancer Institute



DeepPhe

BCH

Guergana Savova, MPI
Sean Finan
Timothy Miller
Dmitriy Dligach
Chen Lin
David Harris
James Masanz

University of
Pittsburgh

Rebecca Jacobson, MPI
Harry Hochheiser
Girish Chavan
Eugene Tseytlin
Olga Medvedeva
Melissa Castine
Mike Davis
Adrian Lee
John Kirkwood
Francesmary Modugno

Funding

NCI U24 CA132672 Cancer Deep Phenotyping from Electronic Medical Records (Savova and Jacobson, MPIs)



Background: eMERGE, PGRN, i2b2, SHARP

Manual chart review



EHR Notes

HPI: 48 woman
with
mam Final diagnosis:
Breast, Left,
Needle Biopsy:
In Patient currently on
Ca neoadjuvant therapy with
Taxol. Due to
cardiomyopathy, patient
not candidate for
Transtuzumab...



Phenotype Label
• positive/negative

Automated chart review

EHR Notes

HPI: 48 woman
with
mam Final diagnosis:
Breast, Left,
Needle Biopsy:
In Patient currently on
Ca neoadjuvant therapy with
Taxol. Due to
cardiomyopathy, patient
not candidate for
Transtuzumab...



Doc1, 0, 2, 3...
Doc2, 0, 0, 1...
Doc3, 1,4,0...

Classifier



Phenotype Label
• positive/negative

DeepPhe Project

<http://cancer.healthnlp.org>

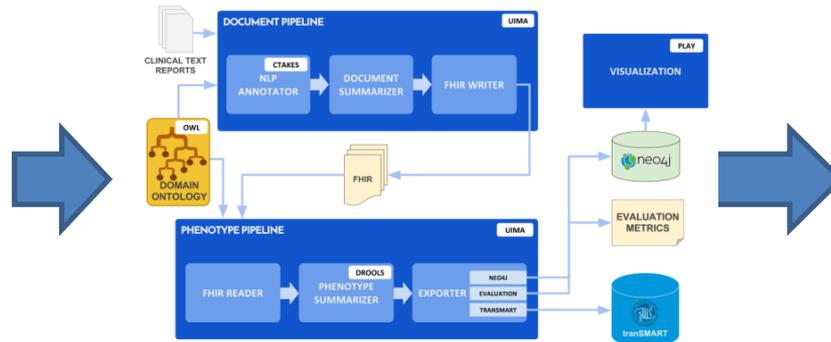
- Goal is to develop next generation cancer deep phenotyping methods
 - No longer dichotomization for a particular phenotype of interest
 - Rather, ***all phenotypes associated with a patient***
- Addresses information extraction but also representation and visualization
- Support **high throughput approach**
 - process and annotate all data at multiple levels (from mention to phenotype) and across time
- Combine IE with structured data (cancer registry)
- Driven by translational research scientific goals as well as surveillance (SEER supplement)

DeepPhe

HPI: 48 woman with right breast mam

Final diagnosis: Breast, Left, Needle Biopsy:

In Patient currently on neoadjuvant therapy with Ca Taxol. Due to cardiomyopathy, patient not candidate for Transtuzumab...



NEOPLASM 1
 Subject: Patient
 Infiltrating Ductal Carcinoma (+3d)
 Location: Right Breast
 Nuclear grade: 2 (+3d)
 ER: negative (+3d)
 PR: negative (+3d)
 Her2/Neu: positive (+3d)
 Size: 2.9 x 2.5 x 2.0 (+3d)

Stage:
 T2N0M0 (interval, +3d, +119d)
 T2N0M1 (+656d)

Lymphadenopathy:
 Negative, Clinical exam (+0d)
 Negative, MRI (+1d)
 Metastasis: Brain (+656d)

TREATMENT 1:
 Neoadjuvant Chemotherapy
 Agents: Paclitaxel (interval, >+3d, >+119d)

TREATMENT 2:
 Her2Neu MAB

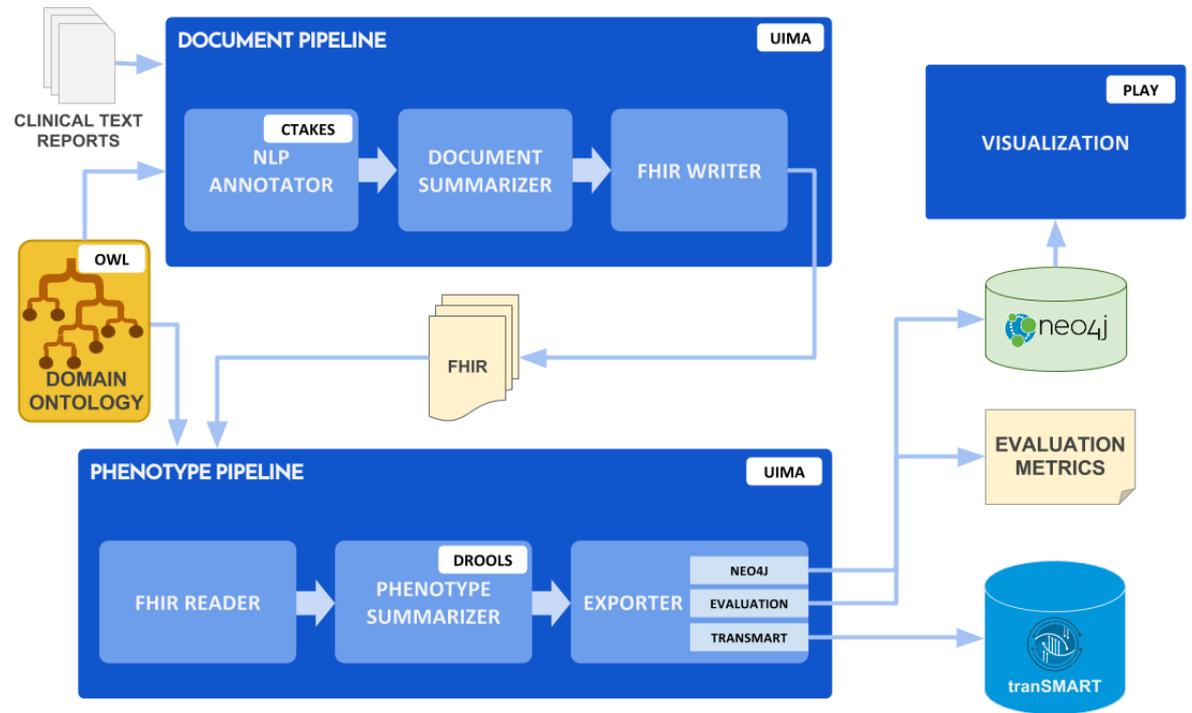
DISEASE 1:
 Subject: patient
 Cardiomyopathy

NEOPLASM 2:
 Subject: mother
 Cancer

.....

Architecture

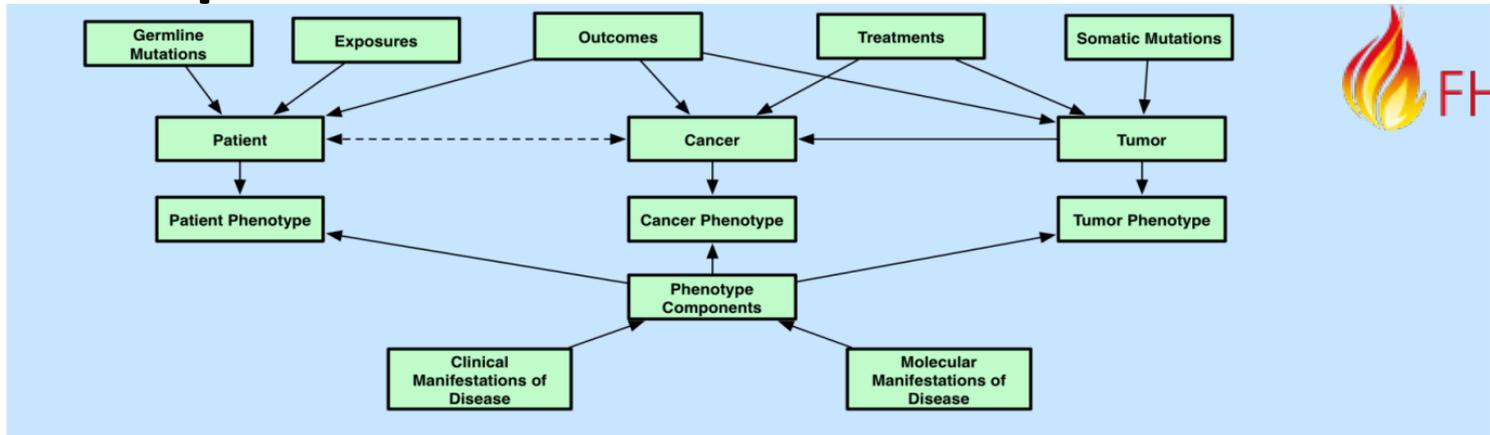
- Portability
- Extensibility
- Modularity
- Ontology driven



DeepPhe Information Model



Level 4
Patients/
Phenotypes

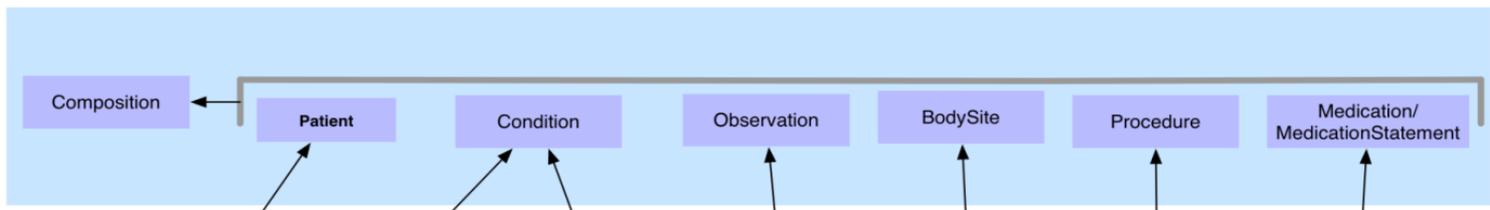


Level 3
Episodes



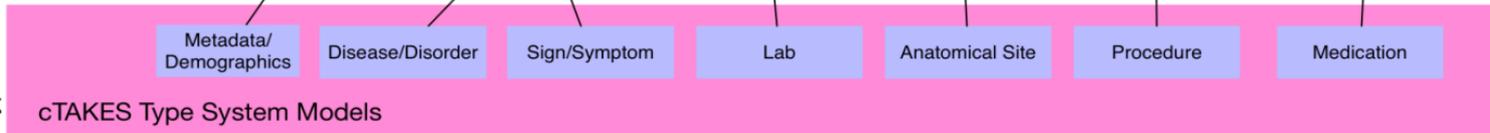
Level 2
Documents

FHIR Models



Level 1
Mentions

CTAKES
Types



DeepPhe NLP Pipeline



Invasive Ductal Carcinoma. 4.4 cm

Tumor is ER-, PR-, Her2-.

Path

58 yo F presents to the ER with slurred speech.

Patient has triple negative breast cancer.

ER

Tumor is ER -, PR -, HER2 -.

Tumor	is	ER	-	PR	-	HER2	-	.
Tumor	is	ER	neg	PR	neg	HER2	neg	.
NN	VBZ	NNP	JJ	NNP	JJ	NNP	JJ	.

Neoplasm C3273930	Estrogen Receptor C0034804	Progesterone Receptor C0034833	erbB2 protein C0069515
----------------------	----------------------------------	--------------------------------------	------------------------------

Patient has triple negative breast cancer.

Patient	has	triple	negative	breast	cancer	.
Patient	has	triple	negative	breast	cancer	.
NN	VBZ	JJ	JJ	NN	NN	.

Patient C0030705	Malignant Neoplasm of Breast C0006142
---------------------	---------------------------------------------

Tumor ER PR Her2

Triple-Negative Tumor



Tumor Phenotype	ER	PR	Her2
	Receptor negative	Receptor negative	receptor negative

Tumor Phenotype	ER	PR	Her2
	Receptor negative	Receptor negative	receptor negative

Tumor Phenotype	ER	PR	Her2
	Receptor negative	Receptor negative	receptor negative

Boundary detection

Tokenization

Normalization

POS tagging

Entity Recognition

Entity Properties

deepPH

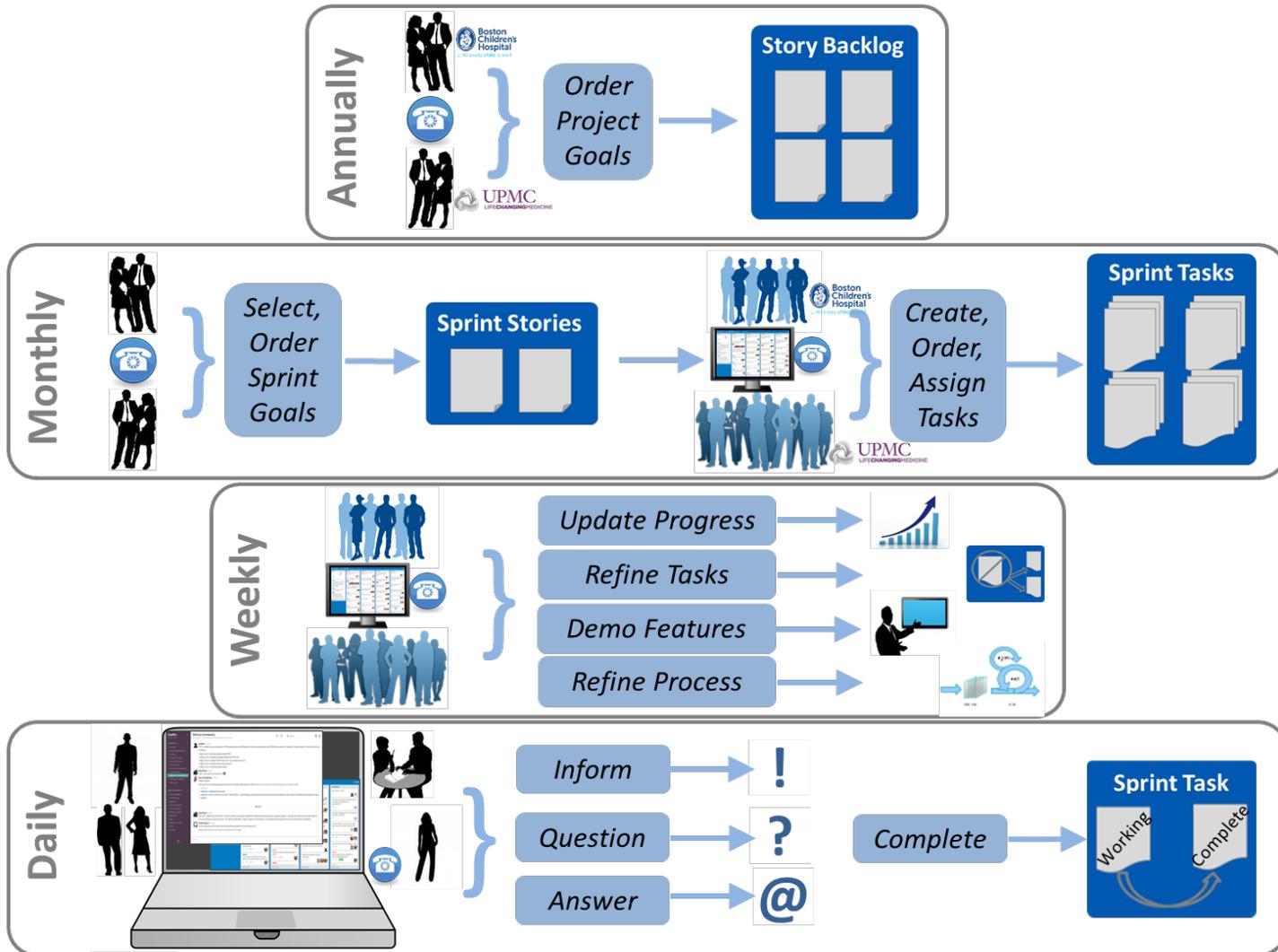
Relation Extraction

Document Summary

Phenotype Summary



Software Development Process

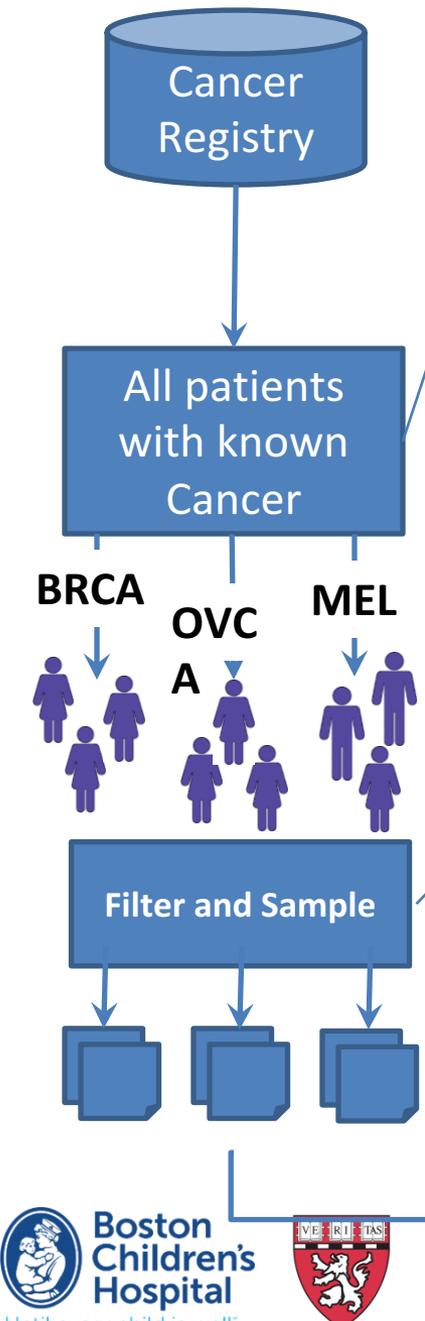


Y3 Developments

- IE methods
 - Coreference
 - Temporal relations
 - Template filling improvement
- Additional templates for Procedures, Medications, Tumor size
- Breast imaging reporting and data system (BIRADS) annotations
- Clinical genomics – gold standard
- New model for Melanoma
- Visualization of patient data from graph db
- Tooling for faster, more efficient evaluations

DeepPhe Gold Set

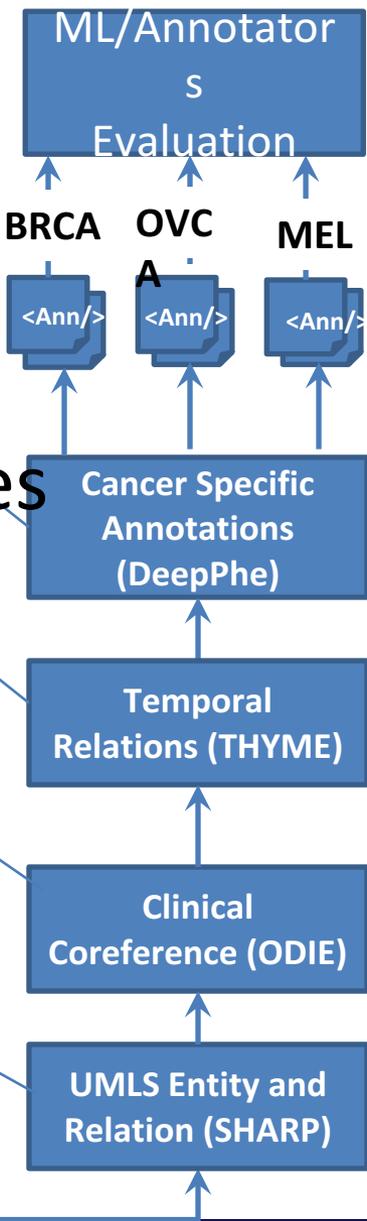
(>500 documents)



- Final contact < 12/31/13
- ~165 Breast Cancer Notes (e.g. TNM, Stage, Metastasis)
- "Analytic" – all care within system
- ~165 Ovarian Cancer Notes (e.g. identifying temporal relationship between two events)
- ~165 Melanoma Notes (e.g. identifying chains relating tumor, mass and cancer)

- Select patients within 2 SD of mean # reports
- Random Sample Text reports
- Collect DS, RAD, PATH, PGN
- Filter reports to specific windows of interest

Deidentify and Preprocess



Evaluation Results: Entity Level (BrCa)

Results on BrCa Test Split									
	<i>overlapping span of template anchor (mention instance)</i>								
	<i>stage</i>	<i>TNM</i>	<i>receptors</i>	<i>metastasis</i>	<i>size</i>	<i>procedures</i>	<i>neoplasms</i>	<i>medications</i>	<i>docTimeRel</i>
#instances in test split	10	48	79	42	55	728	217	121	713
computational method	dictionary lookup	patterns	patterns	dictionary lookup	patterns	dictionary lookup	dictionary lookup	dictionary lookup	machine learning
precision/PPV	1	0.98	1	1	1	0.81	0.87	0.62	0.67
recall/sensitivity	0.9	0.94	0.66	0.79	0.75	0.93	0.86	0.94	0.67
F1	0.95 (1)	0.96 (1)	0.79 (0.89)	0.88 (0.46)	0.85	0.87	0.86 (0.91)	0.75	0.67 (0.65)
	<i>attribute accuracy</i>								
*conditional	1	n/a	1	1	n/a	0.37	1	0.33	
*uncertainty	1	1	1	0.72	n/a	0.99	0.97	1	
*negation	1	n/a	1	0.99	n/a	0.77	0.98	0.96	
*subject	1	n/a	1	1	n/a	0.67	1	1	
*generic	1	n/a	1	0.39	n/a	0.28	0.98	0.46	
associated neoplasm	0.67	0.78	0.67	0.27	0.17	n/a	n/a	n/a	
body location	n/a	n/a	n/a	0.65	n/a	n/a	0.59	n/a	
test method	n/a	n/a	0.62	n/a	0.61	n/a	n/a	n/a	
value	n/a	n/a	1	n/a	n/a	n/a	n/a	n/a	
<i>* indicates weighted accuracy per SemEval 2015 to take into account default value prevalence rates</i>									

Extended Test Set Evaluation

Evaluation Results: Phenotype Level Cancer

Cancer Template Distribution: BrCa	
	#instances in corpus
cancer	52
body location	52
body location side	52
clinical stage	35
cT value	34
cN value	34
cM value	32
pT value	34
pN value	34
pM value	15

corpus: 50 patients, 1881 documents

Results on BrCa Test Set: Phenotype System vs. Gold (results in parentheses are for inter-annotator agreement)			
	Precision/PPV	Recall/Sensitivity	F1 measure
cancer	0.87 (1)	0.83 (1)	0.85 (1)
body location	1.00 (1)	1.00 (1)	1.00 (1)
body location side	1.00 (n/a)	1.00 (n/a)	1.00 (n/a)
clinical stage	0.35 (0.80)	0.93 (1)	0.51 (0.89)
cT value	0.42 (0.89)	0.89 (1)	0.57 (0.94)
cN value	0.71 (0.89)	0.96 (1)	0.82 (0.94)
cM value	0.90 (0.89)	1.00 (1)	0.95 (0.94)
pT value	0.80 (0.89)	0.93 (1)	0.86 (0.94)
pN value	0.66 (0.78)	0.83 (0.88)	0.74 (0.82)
pM value	0.92 (0.62)	0.85 (1)	0.88 (0.77)

Evaluation Results: Phenotype Level Tumor

Tumor Template Distribution: BrCa	
	#instances in corpus
tumor	127
body clock face	44
body quadrant	27
diagnosis	127
tumor type (primary; regional or distant metastasis; local, regional, distant recurrence)	127
er interpretation	51
pr interpretation	51
her2 interpretation	48
calcifications	127
corpus: 50 patients, 1881 documents	

Results on BrCa Test Set: Phenotype System vs. Gold (results in parentheses are for inter-annotator agreement)			
	Precision/PPV	Recall/Sens.	F1 measure
tumor	0.56 (0.79)	0.36 (0.88)	0.44 (0.84)
body clock face	0.50 (0.89)	0.02 (0.73)	0.05 (0.80)
body quadrant	0.52 (0.73)	0.57 (0.80)	0.54 (0.76)
diagnosis	0.77 (0.93)	0.78 (0.93)	0.77 (0.93)
tumor type	0.92 (1)	0.92 (1)	0.92 (1)
er interpretation	0.91 (1)	0.93 (1)	0.92 (1)
pr interpretation	0.84(1)	0.84 (1)	0.84 (1)
her2 interpretation	0.61 (1)	0.55 (1)	0.58 (1)
calcifications	0.67 (n/a)	0.67 (n/a)	0.67 (n/a)

SEER Dev Subset Evaluation

Evaluation Results: Phenotype Level Cancer

Cancer Template Distribution: BrCa	
	#instances in corpus
cancer	240
body location	239
body location side	218
clinical stage	8
cT value	2
cN value	2
cM value	2
pT value	79
pN value	72
M value	12
corpus: 231 patients, 254 documents	

Results on BrCa Simple Train/Dev: Phenotype System vs. Gold (results in parentheses are for inter-annotator agreement)			
	Precision/PPV	Recall/Sensitivity	F1 measure
cancer	0.68 (1)	0.63 (1)	0.65 (1)
body location	1.00 (1)	1.00 (1)	1.00 (1)
body location side	1.00 (n/a)	1.00 (n/a)	1.00 (n/a)
clinical stage	1.00 (0.80)	1.00 (1)	1.00 (0.89)
cT value	1.00 (0.89)	0.50(1)	0.67 (0.94)
cN value	1.00 (0.89)	0.50(1)	0.67 (0.94)
cM value	0.33 (0.89)	0.50(1)	0.40 (0.94)
pT value	0.67 (0.89)	0.70 (1)	0.69 (0.94)
pN value	0.61 (0.78)	0.69 (0.88)	0.65 (0.82)
pM value	0.75 (0.62)	1.00 (1)	0.86 (0.77)

Evaluation Results: Phenotype Level Tumor

Tumor Template Distribution: BrCa	
	#instances in corpus
tumor	270
body clock face	78
body quadrant	15
diagnosis	269
tumor type	270
er interpretation	127
pr interpretation	125
her2 interpretation	96
calcifications	270
corpus: 231 patients, 254 documents	

Results on BrCa Simple Train/Dev: Phenotype System vs. Gold (results in parentheses are for inter-annotator agreement)			
	Precision/PPV	Recall/Sensitivity	F1 measure
tumor	0.64 (0.79)	0.52 (0.88)	0.57 (0.84)
body clock face	0.67 (0.89)	0.03 (0.73)	0.06 (0.80)
body quadrant	1.00 (0.73)	0.77 (0.80)	0.87 (0.76)
diagnosis	0.68 (0.93)	0.59 (0.93)	0.63 (0.93)
tumor type	0.96 (1)	0.96 (1)	0.96 (1)
er interpretation	0.78 (1)	0.26 (1)	0.38 (1)
pr interpretation	0.74 (1)	0.27 (1)	0.39 (1)
her2 interpretation	0.67 (1)	0.30 (1)	0.41 (1)
calcifications	0.86 (n/a)	0.86 (n/a)	0.86 (n/a)

Started Processing Melanoma

localhost

DeepPhe Explorer Patient19

Patient19

Body Site	Hand_Digit_1(Right)
Pathologic M Classification	pMX
Pathologic N Classification	pNX
Pathologic T Classification	pT4a
Treatment	OtherTherapeuticProcedure OtherMedication

Tumors

Hand_Digit_1(Right) Arm(Right)

Body Site	Hand_Digit_1(Right)
Diagnosis	Acral_Lentiginous_Melanoma Spindle_Cell_Melanoma
Pathologic Tumor Size	Pathologic Tumor Size Breslow Thickness
Perineural Invasion	Perineural Invasion
Regression	Regression

Fact Information

ID:MedicalRecord_Acra_Lentiginou:

Name:Acral_Lentiginous_Melanoma

Type: Fact

Rules Applied: set-Tumor-Diagnosis

Text Provenances

[MALIGNANT MELANOMA, ACRAL LENTIGINOUS](#) | [MELANOMA](#) | [MELANO](#)

Reports

002_SP 003_RAD 004_RAD 005_SP 006_SP 007_SP 008_RAD

patient19_report002_SP

Report ID.....490,G3aruv/r1nAc
Patient ID.....G3aruv/r1nAc
Patient Name.....Patient19
Principal Date.....20090106 1738
Record Type.....SP

[Report de-identified (Limited dataset compliant) by De-ID v.6.24.5.1]

PATIENT HISTORY:

The patient has a lesion on the right thumbnail: Rule out squamous cell carcinoma versus onychomycosis.

FINAL DIAGNOSIS:

SKIN, RIGHT THUMB, PUNCH:

- [MALIGNANT MELANOMA, ACRAL LENTIGINOUS](#) AND SPINDLE CELL TYPE.
- CLARK' S LEVEL = AT LEAST IV; THE DEPTH OF INVASION (Breslow' s thickness) IS > 4 mm (see comment).
- SURFACE ULCER IS NOT IDENTIFIED.
- THE [MELANOMA](#) IS IN VERTICAL GROWTH PHASE.
- SLIDE SECTION MARGINS ARE INVOLVED BY [MELANOMA](#) (see also synoptic).
- SOLAR (ACTINIC) ELASTOSIS IS ABSENT.
- ONYCHOMYCOSIS.

COMMENT:

Visual Analytics Initial Prototype

- Support display of patient phenotype and all the reports available for the patient.
- Show cancer and tumor information with the ability to view evidence used to derive the information for each element.
- Highlight evidence in the report text.

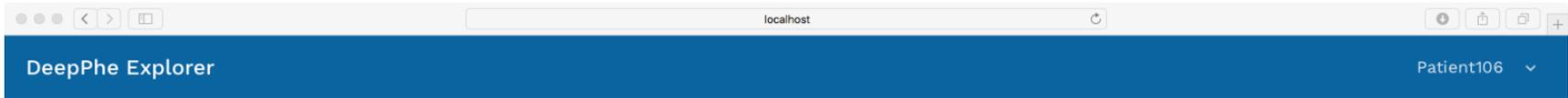
Visual Analytics Architecture

Database: A Neo4J graph database was chosen to most naturally represent relationships between the clinical manifestations, phenotype components and the supporting evidence.

Backend Services: Play Framework. Allows for Java based controllers and backend code.

Front End: React Framework – Allows for modular frontend components.

Screenshot



Patient106

Tumors

Lymph_Node Lymph_Node **Breast**

B I R A D S Category	Breast Imaging Reporting and Data System
Body Site	Breast(Left, Upper-Outer Quadrant)
Calcification	Calcification
Cancer Cell Line	Adenocarcinoma Carcinoma
Diagnosis	Invasive_Ductal_Carcinoma_Not_Otherwise_Sp Ductal_Breast_Carcinoma_In_Situ
E R Status	Estrogen Receptor Status(Positive, Positive)
Her2 Status	HER2/Neu Status(Negative, OtherDiagnositcProcedure, OtherDiagnositcProcedure, Negative)
Histologic Type	Ductal
Ki67 Status	Ki-67 status
Lymphovascular Invasion Status	Lymphatic Invasion
Margin Status	Surgical margins
Nuclear Grade	Nuclear Grade

Fact Information

ID:MedicalRecord_Estrogen_Recept

Name:Estrogen Receptor Status

Type: Observation

Rules collect-all-tumor-
Applied: ERStatus

OrdinalInterpretation: Positive |
Positive |

Text Provenances

positive | ESTROGEN | ESTROGEN |
positive |

Total Nottingham score: 5
Nottingham grade (1, 2, 3): 1
ANGIOLYMPHATIC INVASION: Yes
DERMAL LYMPHATIC INVASION: Not applicable
CALCIFICATION: Yes, malignant zones
TUMOR TYPE, IN SITU: Cribriform
Micropapillary
SURGICAL MARGINS INVOLVED BY INVASIVE COMPONENT:
No
Distance of invasive tumor to closest margin: 4 mm
SURG MARGINS INVOLVED BY IN SITU COMPONENT:
No
LYMPH NODES POSITIVE: 1
LYMPH NODES EXAMINED: 4
METHOD(S) OF LYMPH NODE EXAMINATION:
H/E stain
SENTINEL NODE METASTASIS: Yes
ONLY KERATIN POSITIVE CELLS ARE PRESENT:
No
SIZE OF NODAL METASTASES: Diameter of largest lymph node metastasis: 8 mm
LYMPH NODE METASTASIS(-ES) WITH EXTRACAPSULAR EXTENSION:
No
T STAGE, PATHOLOGIC: pT1c
N STAGE MODIFIER: (sn)
N STAGE, PATHOLOGIC: pN1a
M STAGE: Not applicable
ESTROGEN RECEPTORS: positive, previously performed, H-score: 240
PROGESTERONE RECEPTORS: positive, previously performed, H-score: 135
HER2/NEU: 2+
HER2/NEU (FISH): Not amplified

PATIENT HISTORY:

Publications and Collaborations

1. Chen, Lin; Miller, Timothy; Dligach, Dmitriy; Bethard, Steven; Savova, Guergana. 2016. Improving Temporal Relation Extraction with Training Instance Augmentation. BioNLP workshop at the Association for Computational Linguistics conference. Berlin, Germany, Aug 2016
2. Hochheiser, Harry; Castine, Melissa; Harris, David; Savova, Guergana; Jacobson, Rebecca. 2016. An Information Model for Cancer Phenotypes. BMC Medical Informatics and Decision Making.
3. Ethan Hartzell, Chen Lin. 2016. Enhancing Clinical Temporal Relation Discovery with Syntactic Embeddings from GloVe. International Conference on Intelligent Biology and Medicine (ICIBM 2016). December 2016, Houston, Texas, USA
4. Dligach, Dmitriy; Miller, Timothy; Lin, Chen; Bethard, Steven; Savova, Guergana. 2017. Neural temporal relation extraction. European Chapter of the Association for Computational Linguistics (EACL 2017). April 3-7, 2017. Valencia, Spain.
5. Towards Portable Entity-Centric Clinical Coreference Resolution (Journal of Biomedical InformaticsC
6. Castro SM, Tseytlin E, Medvedeva M, Mitchell KJ, Visweswaran S, Bekhuis T, **Jacobson RS**. Automated annotation and classification of BI-RADS assessment from radiology reports. Journal of Biomedical Informatics 2017 (in press). DOI: 10.1016/j.jbi.2017.04.011
7. DeepPhe system paper (submitted to Special Issue of Cancer Research)
8. Collaboration with THYME (thyme.healthnlp.org)

Goals for Next Year (Y4)

- Extraction of treatment regimens from constituent medications and procedures
- Episode classification and extraction (pre-diagn, diagn, treatment, followup)
- Clinical genomics result extraction
- Expanding to ovarian cancer
- Enhanced visualization including timeline
- Extrinsic evaluation of system with breast cancer clinical research questions

Demo

[add link]