

CLAMP-Cancer – an NLP tool to facilitate cancer research using EHRs

Hua Xu, PhD

School of Biomedical Informatics
The University of Texas Health Science Center at Houston

Advancing Cancer Pharmacoepidemiology Research Through EHRs and Informatics (1 U24 CA194215-01A1)

- Team
 - UTHealth (Hua Xu)
 - Vanderbilt (Josh Denny)
 - Mayo Clinic (Ping Yang)
- Specific Aims of the study

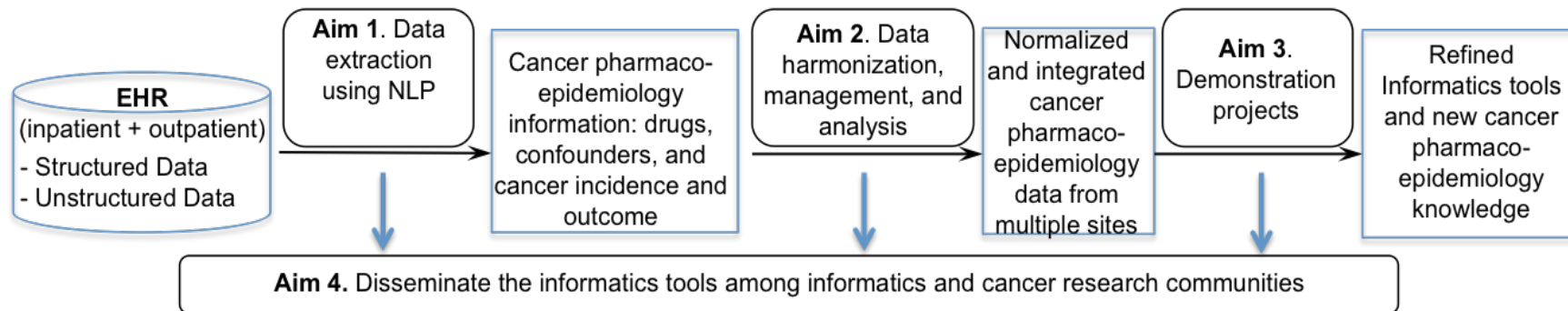


Figure 1. An overview of the proposed specific aims.

EHR - an enabling resource for clinical research

- **CTSA** – Clinical and translational science
- **eMERGE** – EMRs linked with genomic data
- **“All of US”** - Research Program of Precision Medicine Initiative (PMI)
- **PCORnet** - National Patient-Centered Clinical Research Network
- **OHDSI** – Observational health data sciences and informatics

Narratives in EHRs

Admit 10/23

71 yo woman h/o DM, HTN, Dilated CM/CHF, Afib s/p embolic event, chronic diarrhea, admitted with SOB. CXR pulm edema. Rx'd Lasix.

All: none

Meds Lasix 40mg IVP bid, ASA, Coumadin 5, Prinivil 10, glucophage 850 bid, glipizide 10 bid, immodium prn

- Clinical documents contain rich patient information needed for clinical research
- It is costly and time-consuming to extract such information manually

Clinical natural language processing (NLP)

- Methods
 - Named entity recognition (NER), concept encoding, relation extraction (e.g., temporal), co-reference resolution...
- Tools
 - General purpose : MedLEE, MetaMap, cTAKES....
 - Specific purpose : medication, diseases, labs
- Applications
 - Clinical research, clinical decision support, surveillance...

A challenge of clinical NLP - Portability

Need a solution for users to efficiently
build high-performance NLP pipelines for
individual applications!



CLAMP - Clinical Language Annotation, Modeling, and Processing

- A general purpose, high-performance clinical NLP system built on proven methods
- An Integrated development environment (IDE) for building customized clinical NLP pipelines
- A scalable enterprise solution for NLP needs in healthcare organizations
- Available at <http://clamp.uth.edu>



CLAMP – built on proven methods

NLP Tasks		Ranking
Named entity recognition	2009 i2b2, medication	#2
	2010 i2b2 problem, treatment, test	#2
	2013 SHARe/CLEF abbreviation	#1
UMLS encoding	2014 SemEval, disorder	#1
Relation extraction	2012 i2b2 Temporal	#1
	2015 SemEval Disease-modifier	#1
	2015 BioCREATIVE Chemical-induced disease	#1

CLAMP default pipeline performance

- Extract problems, treatments, and tests

Corpus	Entity types	# entity	Exact match			Relaxed match		
			P	R	F1	P	R	F1
MTsamples	treatment, problem, test	25,531	0.841	0.811	0.826	0.921	0.890	0.905
i2b2	treatment, problem, test	72,846	0.891	0.861	0.876	0.958	0.925	0.941
UTNotes	treatment, problem, test	124,869	0.921	0.900	0.910	0.963	0.940	0.951
SemEval 2014	Disease_Disorder	10,077	0.861	0.791	0.824	0.870	0.799	0.833
SemEval 2015	Disease_Disorder	17,333	0.867	0.816	0.840	0.886	0.834	0.859

CLAMP GUI

- An IDE (integrated development environment) for building customized clinical NLP pipelines via GUIs
 - Annotating/analyzing clinical text
 - Training of ML-based modules
 - Specifying rules

Building your own pipeline

The screenshot displays the Clamp Toolkit interface, which is used for building and managing NLP pipelines. The main window shows a list of components and a table of pipeline components.

Resource Panel (Left):

- Machine_learning_components
 - NLP_components
 - Assertion_classifier
 - Chunker
 - Named_entity_recogizer
 - POS_tagger
 - Ruta_rule_engine
 - Section_identifier
 - Sentence_detector
- Corpus
 - lab_corpus
 - mtsamples
- Pipeline
 - defaultPipeline
 - my_labtest
 - sfasdf
 - smokedemo
 - Smoking_status

Component Table (Center):

Name	Component	Description
DF_Detect_sentences_by_newline	Sentence detector	Detect sentences by Newline('\n')
DF_Clamp_tokenizer	Tokenizer	Rule based tokenizer
DF_OpenNLP_POS_tagger	POS tagger	OpenNLP based pos tagger
DF_Dictionary_lookup	Named entity recognizer	dictionary lookup algorithm
DF_NegEx_assertion	Assertion classifier	Assertion info detection using NegEx
DF_Ruta_script_file	Ruta rule engine	Ruta script

Context Menu (Overlaid):

- + save as component
- Export as jar
- Copy (⌘C)
- Paste (⌘V)
- Delete (⌘X)
- Move...
- Rename... (F2)
- Import...
- Export...
- Refresh (F5)
- Properties (⌘I)

Console (Bottom Left):

Console
CorpusInput:

Annotating/Re-training

The screenshot displays the Clamp Toolkit interface, which is used for annotating and re-training Named Entity Recognition (NER) models. The main window shows a document titled "0005.xmi" with the following text:

27 The patient is an 80 year old female with breast cancer ,
status post lumpectomy / radiation therapy / Tamoxifen (2000) , hypertension , hyperlipidemia ,
multiple urinary tract infections who presents with a four
day prodrome of dry cough , rhinorrhea , coryza ,
malaise , chills , headache , decreased p.o. intake ,

The text is annotated with NER tags. The tags are displayed as colored boxes above the text: "predict problem" (blue box) and "problem" (green box). The tags are used to identify entities in the text, such as "breast cancer", "hypertension", "hyperlipidemia", "multiple urinary tract infections", "dry cough", "rhinorrhea", "coryza", "malaise", "chills", "headache", and "decreased p.o. intake".

The interface includes a sidebar on the left with a file explorer showing the project structure:

- i2b2corpus
 - corpus
 - test
 - train
 - models
 - model_20150928_
 - output
 - 0004.xmi
 - 0005.xmi
 - 0008.xmi
 - 0010.xmi

Below the file explorer is a "PipelineView" section showing a "newpipeline" entry.

On the right side, there is an "Outline" panel with a "Semantic" section. It contains a list of categories and their sub-items, each with a checkbox:

- ☒ Entity
 - ☒ problem
 - ☐ test
 - ☐ treatment
- ☒ Relation
- ☒ Syntax

At the bottom of the interface, there is a "Console" panel showing the output of the training process:

```
Console
INFO: load from file, filename=[L/Clo
```

Next to the console is a "Progress" panel showing the progress of the training process:

Train project i2b2corpus NER Training, Fold 2

Extracting features: Training NER model...

Specifying rules

The screenshot displays the Clamp Toolkit interface. The main window shows a code editor with a rule definition for a Named Entity Recognition (NER) task. The rule is written in a specific syntax, defining a block for each sentence and applying a regex to identify 'Tamsulosin' as a treatment.

```
TYPESYSTEM ClampTypeSystem;
//Auto generated by rule editor

BLOCK(ForEach) Sentence{FEATURE("segmentId", "medications")}{
  BaseToken{ REGEXP("Tamsulosin") -> UNMARK(ClampNameEntityUIMA, true),
    CREATE( ClampNameEntityUIMA, 1,1,"semanticTag" = "treatment")};
}
```

Below the code editor, a text input field shows the sentence: "1. Tamsulosin 0.4 mg Capsule , Sust . Release 24HR Sig : One (1) Capsule . Sust . Release 24HR PO HS (at bedtime) .". The word "Tamsulosin" is highlighted in green, and a "test" button is visible above it.

In the bottom left, the "PipelineView" panel shows a tree structure of the pipeline components, including "TEST", "Components", "Name entity recognition", "Pos tagger", "script", "default.ruta", "Section header identifier", "Sentence detector", "Tokenizer", "TEST.pipeline", "Data", "Feature", and "Input".

A dialog box titled "Please specify the rule:" is open in the foreground, allowing the user to define the rule conditions and actions. The dialog has two main sections: "IF" and "THEN".

IF Section:

	[TYPE]	[START OFFSET]	[END OFFSET]	[OPERATOR]	[VALUE]	
CONDITION	Token	0	0	=	Tamsulosin	Remove
AND	Section	0	0	=	medications	Remove

THEN Section:

ASSIGN Tamsulosin **TO** treatment

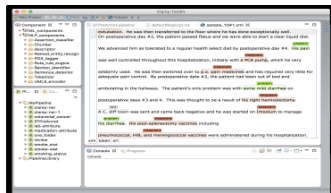
The dialog box includes an "Add condition" button and "OK" and "Cancel" buttons at the bottom.

Two use cases

- Build a rule-based system to extract smoking status in clinical text and classify them into three categories: current smoker, past smoker, and non-smoker
- Build a hybrid (machine learning + rules) system for extracting lab tests and associated values from clinical text
- Videos: <http://clamp.uth.edu/tutorial.php>

CLAMP Enterprise

CLAMP-GUI



NLP algorithm
development



CLAMP-EE



- Deploy at different settings
 - Database
 - Websvc
 - FHIR
- Manage tasks
 - Coordinate execution
 - Monitor system health
- Visualize results
 - Search
 - Validate

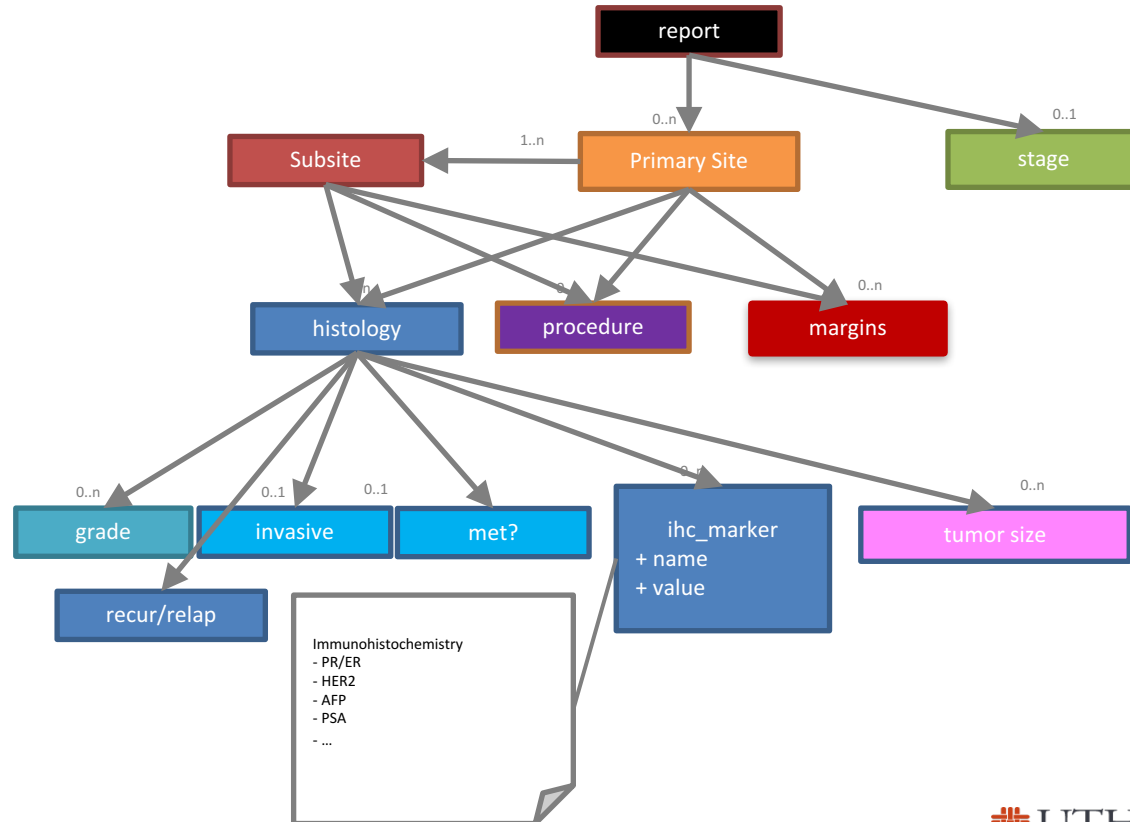
CLAMP-Cancer

- Focus on extracting cancer related information from EHRs
- High-performance default components
- Customizable by end-users
- User-friendly interface

A first attempt – Diagnoses in pathology reports

- Developed an Information Model based on College of American Pathologists (CAP) Cancer Protocols
 - Primary sites, Histology/Grade, Procedure, Margin, Invasion, Metastatic Status, Tumor Size, Biomarkers
- Develop a set of NLP components combining machine learning and rule based approaches to extract entities and relations

CLAMP Cancer Information Model



CLAMP Cancer Components

The screenshot displays the CLAMP Toolkit interface, which is used for managing and editing cancer-related components and pipelines.

Component List (Left Panel):

- Named_entity_recogizer
 - Cancer_Pathology
 - CRF_cancer_entity
 - Biomarker
 - Histology
 - Invasion
 - Margin
 - Metastatic_Status
 - Primary_Site
 - Procedure
 - Specimen
 - SubSite
 - Tumor_Grade
 - Tumor_Size
 - DF_CRF_based_named_entity_rec
 - DF_Dictionary_lookup
 - DF_Regular_expression_NER
- Assertion_classifier
- UMLS_encoder
- Relation_extractor

PathoPipelineFinalML.pipeline (Right Panel):

Buttons: Move up, Move down, Delete, Auto fix, Edit, Edit description

Name	Component	Description
DF_Clamp_sentence_detector	Sentence detector	Rule based sentence detector
DF_Clamp_tokenizer	Tokenizer	Rule based tokenizer
DF_OpenNLP_POS_tagger	POS tagger	OpenNLP based pos tagger
DF_Dictionary_based_section_i...	Section identifier	Dictionary based section header Identifier
Primary_Site	Primary Site	Extract primary body sites
Specimen	Specimen	Extract specimens
Histology	Histology	Extract cancer histologies
Tumor_Grade	Tumor Grade	Extract cancer Grades
Tumor_Size	Tumor Size	Extract cancer tumor size
SubSite	SubSite	Extract body subsites
Procedure	Procedure	Extract cancer procedures
Invasion	Invasion	Extract cancer invasions
Margin	Tumor Margin	Extract cancer margins
Biomarker	Biomarker	Extract biomarkers
DF_NegEx_assertion	Assertion classifier	Assertion info detection using NegEx
Rule_based_cancer_relation_co...	Ruta rule engine	Ruta script

DESCRIPTION:

INPUT:

OUTPUT:

CATEGORY:

Pipeline (Bottom Left Panel):

- MyPipeline
- PipelineLibrary

Console (Bottom Right Panel):

Console

CLAMP-Cancer (Vanderbilt)

Type of information	Exact Matching			Relaxed Matching		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Specimen	0.96	0.96	0.96	0.96	0.96	0.96
Primary Site	0.96	0.95	0.95	0.98	0.97	0.98
Sub-Site	0.91	0.82	0.86	0.94	0.85	0.90
Procedure	0.96	0.97	0.97	0.97	0.98	0.98
Histology	0.91	0.85	0.88	1.00	0.93	0.97
Tumor Grade	0.92	0.94	0.93	0.97	1.00	0.99
Tumor Size	1.00	0.83	0.91	1.00	0.83	0.91
Tumor Margin	0.93	0.93	0.93	1.00	1.00	1.00
Invasion	0.96	0.93	0.95	0.96	0.93	0.95
Biomarker	0.92	0.88	0.90	0.98	0.94	0.96

CLAMP-Cancer and MedKAT (Mayo)

	CLAMP-Cancer			MedKAT		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Tumor size	1.00	0.99	0.99	1.00	1.00	1.00
Dimension Extent	1.00	0.99	0.99	0.99	1.00	0.99
Dimension Unit	1.00	1.00	1.00	1.00	1.00	1.00
Tumor Site	0.94	0.89	0.92	0.96	0.95	0.96
Histology	0.91	0.92	0.92	0.96	0.98	0.97
Grade	1.00	0.88	0.94	0.93	0.97	0.99
Date	1.00	1.00	1.00	1.00	1.00	1.00

Soysal et al. *under review*

Future work

- Advanced methods to facilitate the portable NLP solutions
- New interfaces for cancer researchers
- More cancer specific NLP pipelines for diverse types of clinical notes
- Interoperability with other NLP systems: cTAKES and MetaMap

Acknowledgement

- **Collaborators**

- Josh Denny, MD
- Jeremy Warner, MD
- Cindy Chen, PhD
- Ping Yang, MD
- Hongfang Liu, PhD
- Serguei Pakhomov, PhD
- Xianling Du, PhD

- **Grants**

- NCI U24 CA194215
- CPRIT R1307
- NIGMS R01 GM102282
- NLM R01 LM010681

Team members:

- Jingqi Wang
- Min Jiang
- Ergin Soysal
- Sungrim Moon
- Jun Xu
- Yaoyun Zhang
- Anupama Gururaj
- Yonghui Wu
- Nina Slimi
- Kyle Nguyen
- Tolulola Dawodu
- Yukun Chen
- Qiang Wei
- Saied Pournajati
- Rui Li

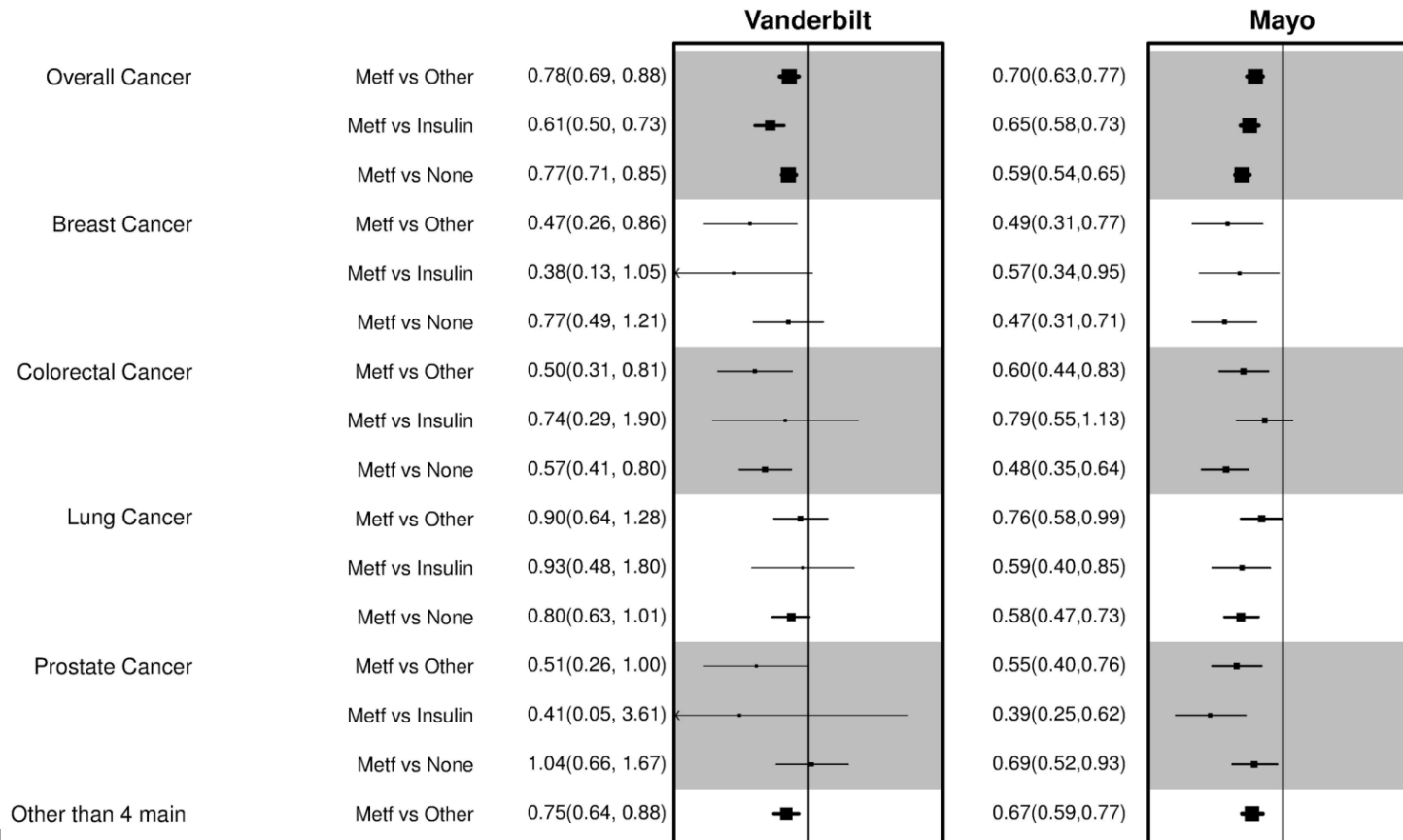


Thank you!

Questions?

hua.xu@uth.tmc.edu

NLP to support clinical research - Metformin and Cancer Survival



Screen 153 drugs for potential cancer therapeutic signals

- Single Drug



153 drugs for chronic diseases



- Carefully selected covariates (30+)

DATA	A	B	C	D	E	F	G
1	DATA	PRODUCT	Price	Months	Year	Units Sold	Revenue
2	047050	Standard Whiglet	\$1.99	February	2010	10226	\$20,450.44
3	041211	Fancy Whiglet	\$1.99	February	2010	3211	\$6,391.89
4	790062	Extravagant Whiglet	\$6.99	February	2010	1001	\$6,996.99
5	041213	Deluxe Whiglet	\$10.99	February	2010	134	\$1,462.76
6	790099	Over-the-Top Whiglet	\$10.99	February	2010	0	\$0.00
7	047050	Standard Whiglet	\$1.99	March	2010	12789	\$25,450.11
8	041211	Fancy Whiglet	\$1.99	March	2010	1288	\$2,564.93
9	790062	Extravagant Whiglet	\$6.99	March	2010	1009	\$6,996.99
10	041213	Deluxe Whiglet	\$10.99	March	2010	201	\$2,192.42
11	790099	Over-the-Top Whiglet	\$10.99	March	2010	0	\$0.00
12	047050	Standard Whiglet	\$1.99	April	2010	9824	\$19,649.26
13	041211	Fancy Whiglet	\$1.99	April	2010	3023	\$6,035.77
14	790062	Extravagant Whiglet	\$6.99	April	2010	1458	\$10,175.42
15	041213	Deluxe Whiglet	\$10.99	April	2010	186	\$2,044.14
16	790099	Over-the-Top Whiglet	\$10.99	April	2010	0	\$0.00



all-inclusive covariates (3000+)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N			
1	DATA	PRODUCT	Price	Months	Year	Units Sold	Revenue	Price	Months	Year	Units Sold	Revenue	Price	Months	Year	Units Sold	Revenue
2	047050	Standard Whiglet	\$1.99	February	2010	10226	\$20,450.44	\$1.99	February	2010	10226	\$20,450.44	\$1.99	February	2010	10226	\$20,450.44
3	041211	Fancy Whiglet	\$1.99	February	2010	3211	\$6,391.89	\$1.99	February	2010	3211	\$6,391.89	\$1.99	February	2010	3211	\$6,391.89
4	790062	Extravagant Whiglet	\$6.99	February	2010	1001	\$6,996.99	\$6.99	February	2010	1001	\$6,996.99	\$6.99	February	2010	1001	\$6,996.99
5	041213	Deluxe Whiglet	\$10.99	February	2010	134	\$1,462.76	\$10.99	February	2010	134	\$1,462.76	\$10.99	February	2010	134	\$1,462.76
6	790099	Over-the-Top Whiglet	\$10.99	February	2010	0	\$0.00	\$10.99	February	2010	0	\$0.00	\$10.99	February	2010	0	\$0.00
7	047050	Standard Whiglet	\$1.99	March	2010	12789	\$25,450.11	\$1.99	March	2010	12789	\$25,450.11	\$1.99	March	2010	12789	\$25,450.11
8	041211	Fancy Whiglet	\$1.99	March	2010	1288	\$2,564.93	\$1.99	March	2010	1288	\$2,564.93	\$1.99	March	2010	1288	\$2,564.93
9	790062	Extravagant Whiglet	\$6.99	March	2010	1009	\$6,996.99	\$6.99	March	2010	1009	\$6,996.99	\$6.99	March	2010	1009	\$6,996.99
10	041213	Deluxe Whiglet	\$10.99	March	2010	201	\$2,192.42	\$10.99	March	2010	201	\$2,192.42	\$10.99	March	2010	201	\$2,192.42
11	790099	Over-the-Top Whiglet	\$10.99	March	2010	0	\$0.00	\$10.99	March	2010	0	\$0.00	\$10.99	March	2010	0	\$0.00
12	047050	Standard Whiglet	\$1.99	April	2010	9824	\$19,649.26	\$1.99	April	2010	9824	\$19,649.26	\$1.99	April	2010	9824	\$19,649.26
13	041211	Fancy Whiglet	\$1.99	April	2010	3023	\$6,035.77	\$1.99	April	2010	3023	\$6,035.77	\$1.99	April	2010	3023	\$6,035.77
14	790062	Extravagant Whiglet	\$6.99	April	2010	1458	\$10,175.42	\$6.99	April	2010	1458	\$10,175.42	\$6.99	April	2010	1458	\$10,175.42
15	041213	Deluxe Whiglet	\$10.99	April	2010	186	\$2,044.14	\$10.99	April	2010	186	\$2,044.14	\$10.99	April	2010	186	\$2,044.14
16	790099	Over-the-Top Whiglet	\$10.99	April	2010	0	\$0.00	\$10.99	April	2010	0	\$0.00	\$10.99	April	2010	0	\$0.00

- Hypothesis-driven



Data-driven

Results

Rank	Drug	HR	P-value	Lower	Upper	Adj-p-value
1	sildenafil	0.704	0.000	0.621	0.797	0.000
2	olmesartan	0.714	0.001	0.580	0.879	0.017
3	thyroxine	0.791	0.000	0.734	0.852	0.000
4	carvedilol	0.803	0.003	0.693	0.930	0.033
5	alendronic.acid	0.812	0.000	0.725	0.908	0.004
6	amlodipine	0.839	0.000	0.782	0.901	0.000
7	epoetin.alfa.recom	0.840	0.000	0.772	0.914	0.001
8	ramipril	0.843	0.004	0.750	0.947	0.036
9	simvastatin	0.845	0.000	0.788	0.906	0.000
10	atorvastatin	0.861	0.000	0.797	0.930	0.003
11	metformin	0.862	0.003	0.781	0.952	0.033
12	esomeprazole	0.878	0.000	0.819	0.942	0.004
13	omeprazole	0.879	0.000	0.830	0.931	0.000
14	lisinopril	0.902	0.002	0.846	0.962	0.018
15	lansoprazole	0.908	0.005	0.849	0.971	0.037

4 drugs with strong evidence – ongoing clinical trials

7 drugs with weak evidence from literature