Clinical Cancer Informatics and Natural Language Processing for Research in Canada

Marshall Pitz, MD MHS FRCPC

Medical Oncologist

Chief Medical Information Officer

Head, Clinical Research

CancerCare Manitoba

Associate Professor

Rady Faculty of Medicine

University of Manitoba

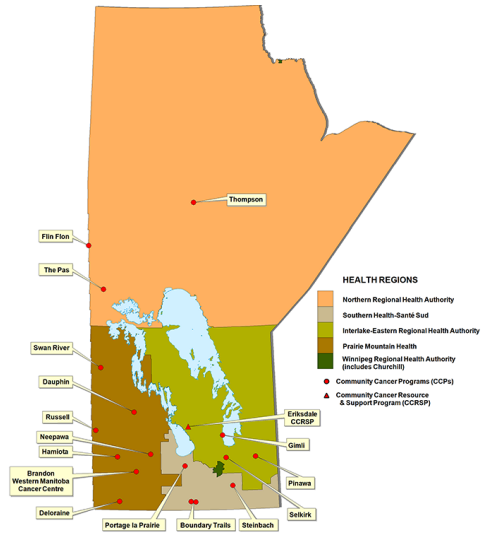
Disclosures

- No financial disclosures

- What I’m presenting is a very narrow view of Cancer Informatics and NLP in Canada

Background

- CancerCare Manitoba
  - Regional Cancer Centre
  - 1.5M people
- Manitoba Cancer Registry
  - 1937 basic reporting
  - 1956 – population based, legislated mandatory reporting
  - NLP for cancer reporting in use since 2009



Clinical Research and the Registry

Case Ascertainment

Excellent survival statistics

Biomarkers

(ER/PR/HER2, EGFR, IDH)

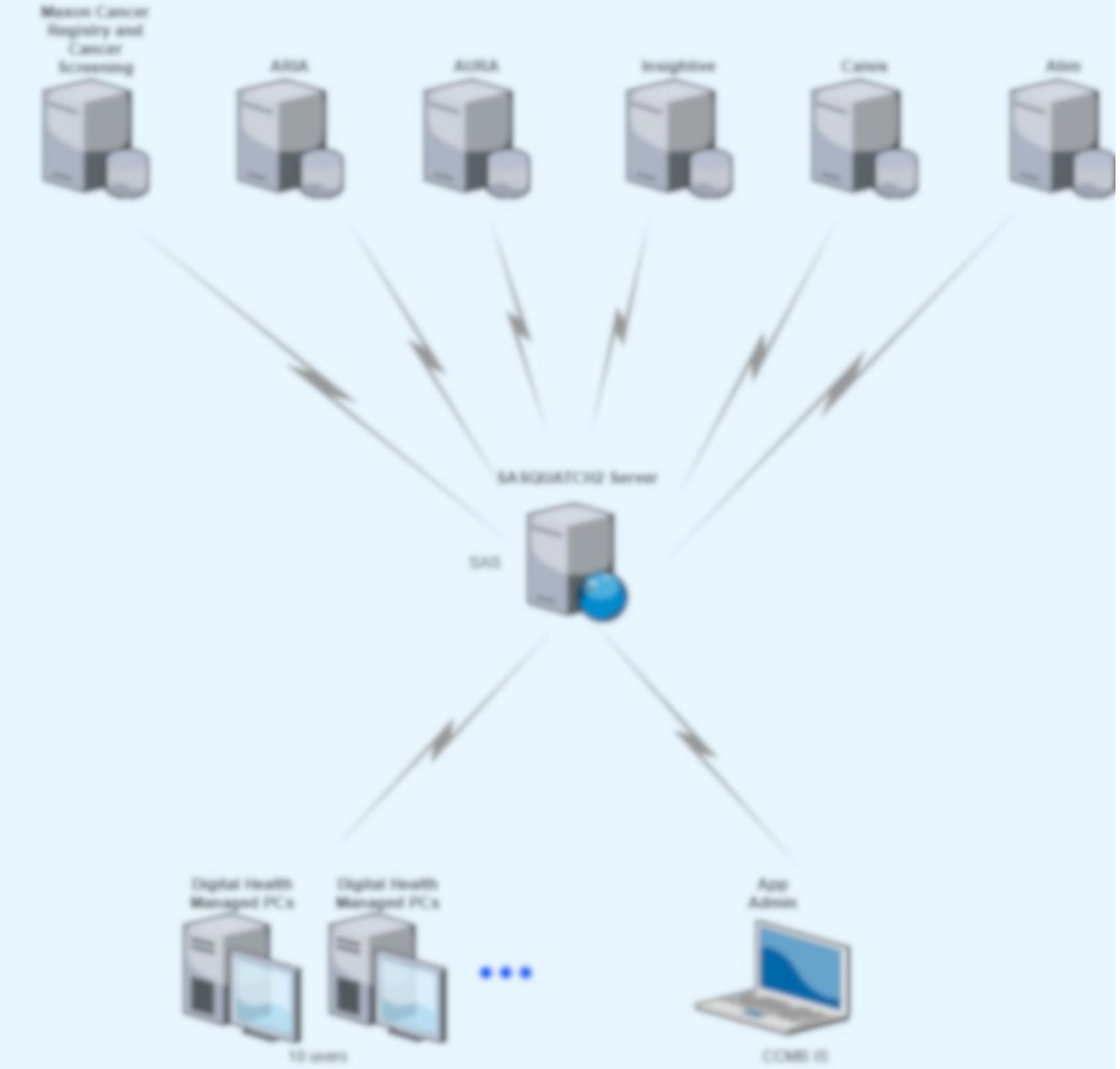
Recurrence and progression not captured well

Collaboration

- Build on to standard registry data capture
- Capture biomarkers
- More than just 1st treatment in the year
- Dose and fraction of RT

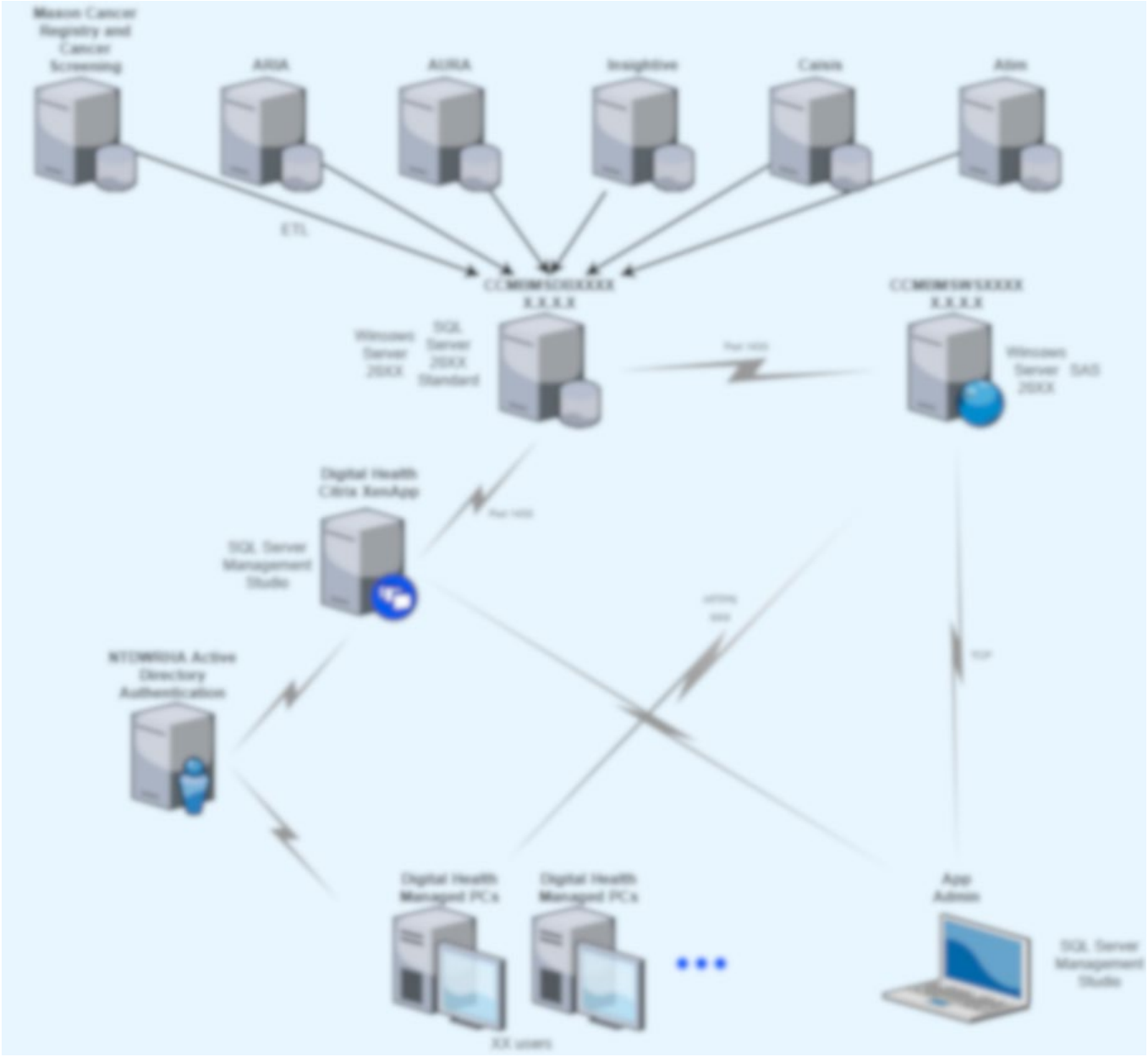
Clinical Research

- Cancer Registry forms the basis of cohort selection/identification
  - Demographics
  - Primary disease
  - Stage
  - Survival
- Link to:
  - EMR
  - Administrative Data
  - Research project data
  - Tumour Bank
  - …



Clinical Research Database Redesign

- Cancer Registry is core to the redesign of our research data structure
- Source of truth for many variables
- Linkage/patient identifier variables
- Highly structured
- Robust data dictionary



Cancer Recurrence as a Clinical Endpoint

Important endpoint for clinical research

Requires chart review (paper/electronic) and multiple systems

- Time consuming
- Expensive
- Reviewer dependent

Could we develop an algorithm that would identify cancer recurrence using:

- Registry
- Administrative Data
- EMR data
- Notes (Natural Language Processing)

Cancer Recurrence Study

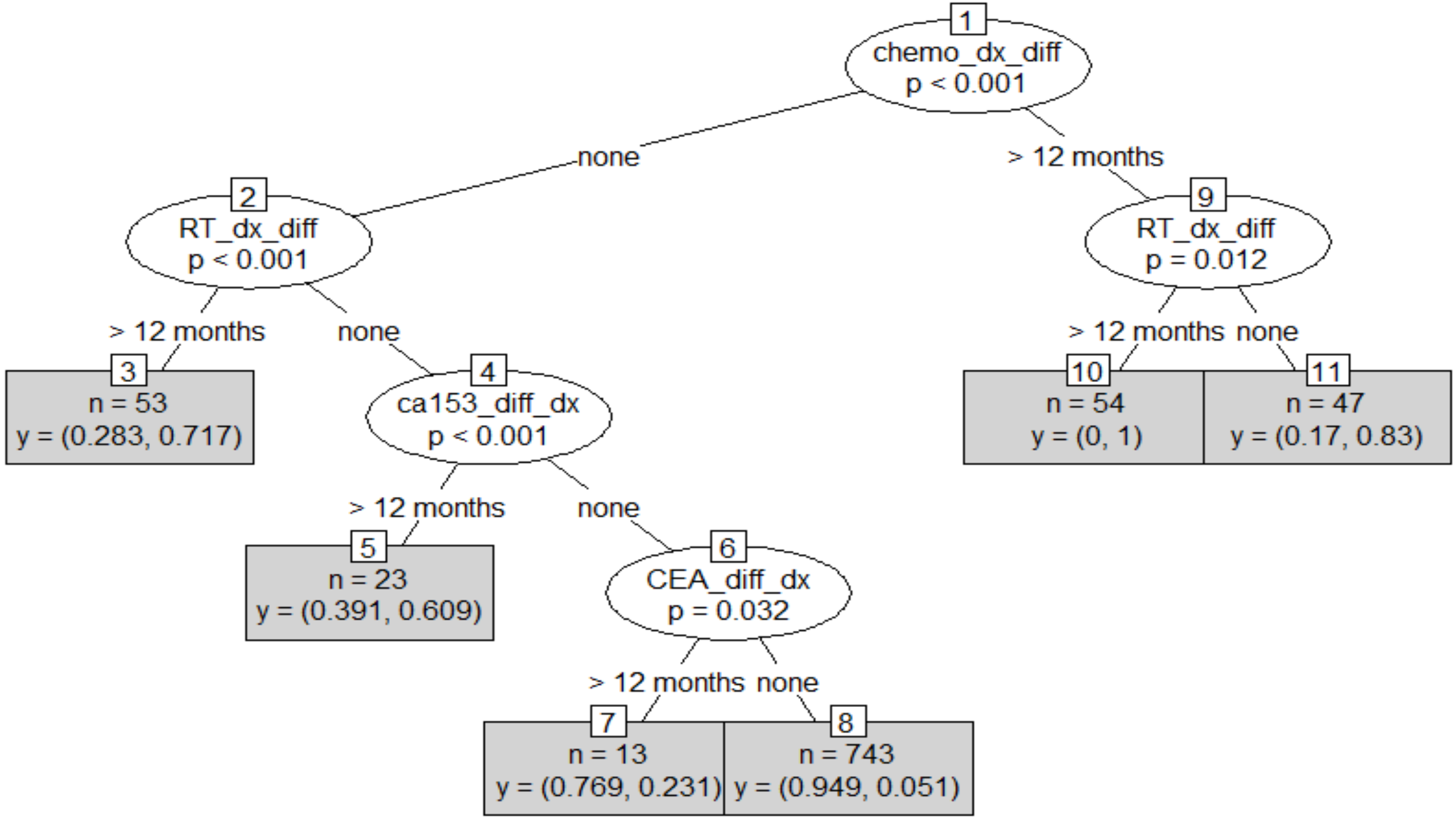
- Breast & Colorectal Cancer
- Stage I - III
- 2004 – 2007 (Algorithm development)
- 2008 – 2012 (Validation)
- Cancer Registrar chart review (gold standard)

- Does adding NLP of clinical notes improve detection of breast or colorectal cancer recurrence compared to existing discrete data?

Administrative Definition of Recurrence

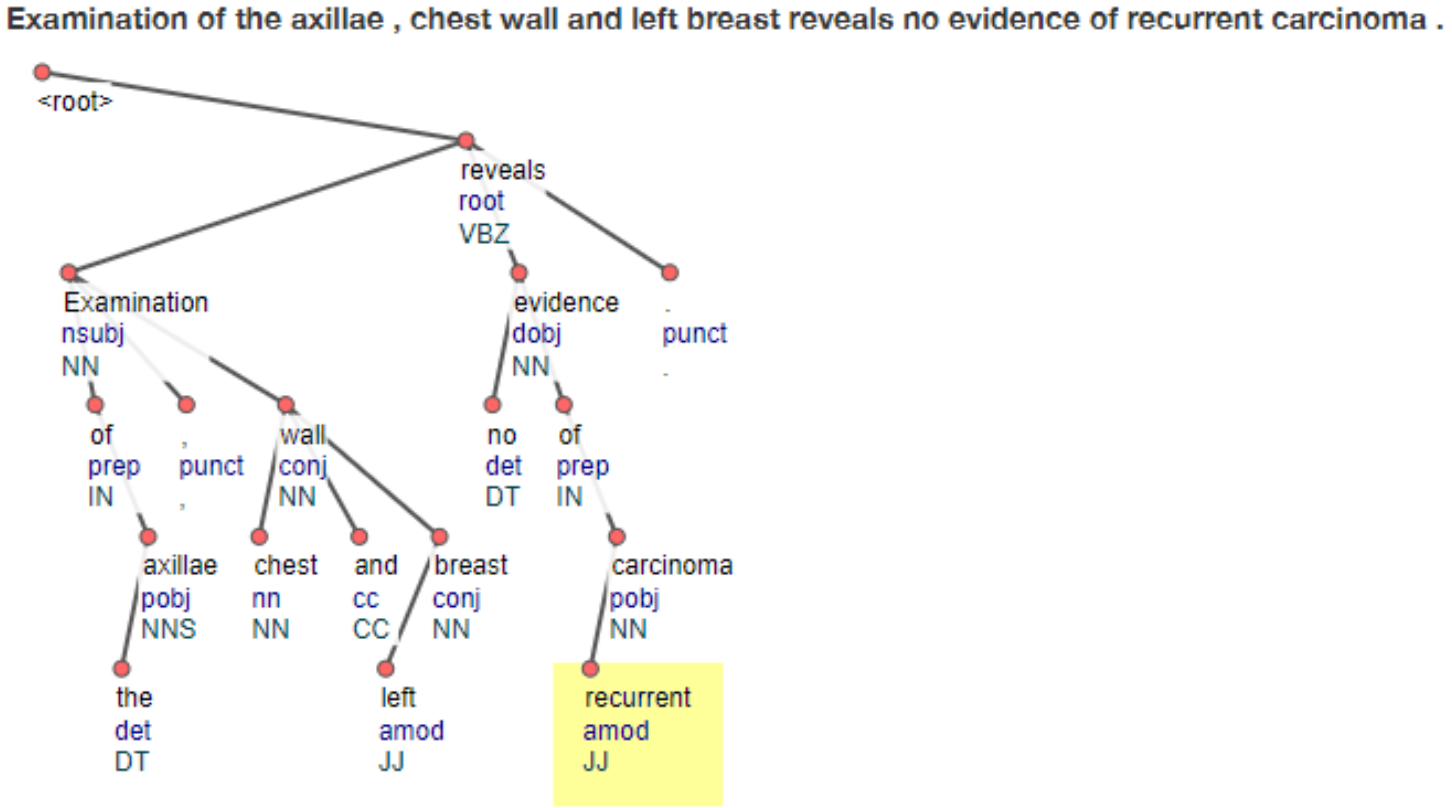
- Index date: original cancer diagnosis; initiation of treatment

- First surgery/chemo/RT >”X” months afterwards
- Second surgery/chemo/RT >12 months after first surgery/chemo/RT
- Palliative care consults >6 months after initial treatment


1
chemo_dx_diff
p < 0.001
none
> 12 months
2
RT_dx_diff
p < 0.001
9
RT_dx_diff
p = 0.012
> 12 months
none
> 12 months
none
3
n = 53
y = (0.283, 0.717)
4
ca153_diff_dx
p < 0.001
10
n = 54
y = (0, 1)
11
n = 47
y = (0.17, 0.83)
> 12 months
none
5
n = 23
y = (0.391, 0.609)
6
CEA_diff_dx
p = 0.032
> 12 months
none
7
n = 13
y = (0.769, 0.231)
8
n = 743
y = (0.949, 0.051)

Natural Language Processing to detect recurrence

- What is NLP?
  - Statistical algorithms that ‘read’ text (clinical notes) to assign values to structured variables
    - “recurrent breast cancer” labels that note as a “recurrence”
  - Different types of algorithms, machine learning and deep learning techniques
  - Natural language is a significant challenge:
    - Indirect language
    - Spelling mistakes
    - Unclear subjects for recurrence (eg. recurrence of cancer vs. recurrence of pain)



Iterative Process

- Tagging / Labelling features of notes
- Identify patients that have recurrence
- Identify the notes (dates) recurrence
- Improve the algorithm

Results

- 2114 Breast cancer patients; 353 recurrences

- Mean age 61
- Stage I: 39%
- Stage II: 43%
- Stage III: 18%

Results

Model	Sensitivity	Specificity	Scaled Brier
Administrative	78.0%	95.7%	0.51
NLP	90%	90%	0.42

*unpublished; preliminary

Conclusions

The use of NLP for clinical research is challenging

- Model building continues to improve
- Not yet a replacement for chart review

Cancer registry data can be foundational for clinical research

Clinician-Registry collaboration can greatly improve the utility

Acknowledgements

Manitoba Team

- Dr. Harminder Singh
- Dr. Kathleen Decker
- Pascal Lambert
- Liz Harland

Waterloo Team

- Dr. Helen Chen
- Dr. George Michaelopoulos
- Sujan Subendran
- Hussam Kaka
- Alex Wong