Automatic Extraction of Synoptic Data

George Cernile
Artificial Intelligence in Medicine
AIM
Agenda

- Background
- Technology used
- Demonstration
Partial CAP Checklist definitions for 3 Cancer Sites

**Breast**

**Specimen Type** (Attributes)
- Excision
- Mastectomy
- Other (specify):
- Not specified

**Lymph node sampling**
- No lymph node sampling
- Sentinel lymph node(s) only
- Sentinel lymph node with axillary dissection
- Axillary dissection

**Specimen Size**
- Greatest dimension: _ cm
  *Additional dimensions: _ x _ cm
- Cannot be determined

**Colorectal**

**Tumor Site**
- Cecum
- Right (ascending) colon
- Hepatic flexure
- Transverse colon
- Splenic flexure
- Left (descending) colon
- Sigmoid colon
- Rectum
- Not specified

**Polyp Size**
- Greatest dimension: _ cm
  *Additional dimensions: _ x _ cm
- Cannot be determined

**Brain/Spinal Cord**

**Specimen Type**
- Open biopsy
- Stereotactic needle core biopsy
- Subtotal/partial resection
- Total resection
- Other (specify):
- Not specified

**Specimen Size**
- Greatest dimension: _ cm
  *Additional dimensions: _ x _ cm

**Polyp Configuration**
- Pedunculated with stalk
  - Stalk length: _ cm
- Pedunculated, no stalk
- Sessile
- Fragmented

**Tumor Site (check all that apply)**
- Cerebral meninges
- Cerebrum (specify lobe(s), if known):
  - Basal ganglia
  - Thalamus
  - Hypothalamus
  - Suprasellar
  - Pineal
  - Cerebellum
  - Cerebellopontine angle
  - Ventricles
  - Brain stem
  - Spinal cord
  - Nerve root
  - Other (specify):
Questions

- How often are checklist elements included in a report, and does this vary by type of report/cancer?
- With what degree of confidence can checklist elements be identified and the associated value read?
- Could a system associate the appropriate set of checklist elements with the cancer that is being reported?
- How much variability is there in the completeness of checklist elements between and within types of cancer?
- What type of lexicon and logic system would be required to encode the checklist elements?
- Will the system be able to determine when data are ambiguous and, if so, how should they be dealt with?
# Machine readability criteria

<table>
<thead>
<tr>
<th>Score</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Checklist element present, value clearly defined</td>
</tr>
<tr>
<td>2</td>
<td>Checklist element present, value ambiguous</td>
</tr>
<tr>
<td>3</td>
<td>Checklist element not present, value can be inferred</td>
</tr>
<tr>
<td>4</td>
<td>Cannot determine value</td>
</tr>
<tr>
<td>Checklist Element</td>
<td>Text</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Lymph Node Sampling</td>
<td>“The specimen consists of 3 small lymph nodes ranging in size from 0.5 to 1.2 cm.”</td>
</tr>
<tr>
<td>Tumor Size</td>
<td>“The specimen is serially sectioned to reveal a grayish white firm tumor with ill defined Margins measuring 2 x 2.2 x 2.8 cm in greatest dimensions.”</td>
</tr>
<tr>
<td>Size of Invasive Component</td>
<td>“Invasive Tumour Size  2.0 x 1.7 cm “</td>
</tr>
<tr>
<td>Histologic Type</td>
<td>“Sections showed multiple cores of breast tissue showing fibrosis with scattered areas of small ductal epithelial cells arranged in sheets and cords of cells suspicious for invasive carcinoma.”</td>
</tr>
</tbody>
</table>
Machine readability data

<table>
<thead>
<tr>
<th>Checklist Element Name (CE)</th>
<th>No. of Reports Containing CE Identifier</th>
<th>No. of Reports Containing CE Value</th>
<th>Machine Legibility Score Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Histologic Grade</td>
<td>40</td>
<td>31</td>
<td>40</td>
</tr>
<tr>
<td>Specimen Type</td>
<td>7</td>
<td>55</td>
<td>7</td>
</tr>
<tr>
<td>Lymph Node Sampling</td>
<td>16</td>
<td>39</td>
<td>16</td>
</tr>
<tr>
<td>Specimen size</td>
<td>1</td>
<td>63</td>
<td>1</td>
</tr>
<tr>
<td>Laterality</td>
<td>1</td>
<td>66</td>
<td>1</td>
</tr>
<tr>
<td>Tumor Site</td>
<td>2</td>
<td>58</td>
<td>8</td>
</tr>
<tr>
<td>Size of Invasive Component</td>
<td>9</td>
<td>29</td>
<td>9</td>
</tr>
<tr>
<td>Histologic Type</td>
<td>37</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>Pathologic Staging</td>
<td>4</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>Margins</td>
<td>45</td>
<td>45</td>
<td>45</td>
</tr>
</tbody>
</table>

70 pathology reports of breast cancer

**Checklist element not present, value can be inferred**
Internal Coding system

“SYN  xx  x  xx  xx”

- Site
- Procedure
- Attribute
- Value

e.g.
- Site: Breast  (01)
- Procedure: n/a  (0)
- Attribute: Histologic Type  (07)
- Value: Non invasive Carcinoma  (01)

Code: SYN_0100701
Example from Breast

“Specimen type: Mastectomy, modified radical Orientation”

SYN-0100100
- Site: Breast (01)
- Procedure: n/a (0)
- Value: Attribute Name (00)
- Attribute: Specimen Type (01)

SYN-0100102
- Value: Mastectomy (02)
Synoptic extraction tasks

- Identify the report type
  - Lexicons ICD-O-3, etc return concepts found in text
  - Knowledge base infers report type from concepts

- Extract relevant data elements
  - Lexicon specific report type returns relevant concepts
  - Numeric extraction algorithm
  - Knowledge base matches values and numbers to elements
Site Identification

- **NHL** M-9590/3 – M-9591/3, M-9670/3 – M-9699/3, M-9700/3 – M-9719/3
- **Hodgkin’s Lymphoma** M-9650/3 – M-9667/3
- **Ewing Sarcoma** M-9260/3
- **Uveal Melanoma** C69.* + M-8772/3 – M-8774/3
Inference Engine

- Imbedded AI tool for knowledge base
  - CLIPS (developed by NASA)

- Powerful pattern matching and rule based inference engine

- Can express arbitrarily complex knowledge to provide a system with expert level problem solving
Numeric Extraction

- Algorithm designed to extract numeric values
- Knowledge base compares ranges and units
- Examples
  - “The tumor measures 1.5 cm.”
  - “breast tissue measuring approximately 5 x 2 x 1 cm”
System block diagram

Stage 1: Input Pathology Reports
Stage 2: Identify Report Type (Lexicon and knowledge base)
Stage 3: Extract Checklist Elements (Lexicon and knowledge base)
Stage 4: Output Synoptic Summary

Report type or Site:
- Breast
- Lung
- Skin
- Etc.
Live Demonstration

**Synoptex Reporting Tool**

**Elements found:** 15/26
- BREAST Synoptic Report
  - **Specimen Type:** Lymph node
  - **Lymph Node Sampling:** Axillary dissection
  - **Lymph Nodes Examined:** 15/26
  - **Specimen Size (for specimens larger than 2.5 cm):** 2.10 x 0.6 x 1.5 cm
  - **Laterality:** Left
  - **Tumor Site:** Upper outer quadrant
  - **Size of Invasive Component:** 1.9 cm
  - **Histologic Type:** Ductal carcinoma in situ (DCIS)
  - **Invasive carcinoma (NOS):** 15/26
  - **Invasive ductal carcinoma:** 15/26
- **Grading System:** Bloom Richardson
  - **Grade:** 3
- **Tubule Formation:** 3
- **Nuclear Morphism:** 3
- **Neurot:** Greater than 20 mitoses per 10 HPF (scores 3)
- **Malign Count:** Greater than 20 mitoses per 10 HPF (scores 3)
- **Pathologic Staging (pTNM):**
  - **T Stage:** T3
  - **N Stage:** N0
  - **M Stage:** M0
- **ER Status:** Negative
- **PR Status:** Negative
- **Margins:** Negative
  - **Margins uninvolved by invasive carcinoma:** Negative
- **Extent of Margin Involvement for Invasive Component:** Negative
- **Extent of Intratumoral Component:** Negative
- **Lymphatic/Vascular Invasion:** Negative
  - **Pathologic:** Vascular invasion absent
  - **Histologic:** Vascular invasion absent
- **Microcalifications:** Histologic califications present
- **Additional Pathologic Findings:**

**Values found:** 13/26

- **BREAST Synoptic Report**
  - **Lymph node:** Axillary dissection
  - **Lymph Nodes Examined:** 13/26
  - **Specimen Size (for specimens larger than 2.5 cm):** 2.10 x 0.6 x 1.5 cm
  - **Laterality:** Left
  - **Tumor Site:** Upper outer quadrant
  - **Invasive carcinoma (NOS):** 13/26
  - **Invasive ductal carcinoma:** 13/26

**Concordance Code Tissue:** 1.11; **Concordance Code Tissue:** 2.21

**Diagnosis:** Breast, NOS Lymph node: left breast. Lymph node, NOS: left axillary lymph nodes.

**Specimen:** Breast lymph node (left) - Invasive ductal carcinoma. Modified Bloom-Richardson

**Grade:** 3. Invasive tumour measures 1.9 cm in greatest dimension. Resection margins negative. Invasive tumour measures less than 1 mm from the anterior resection margin. Lymphovascular invasion absent. See synoptic report.

**Axillary lymph nodes dissection:** 13 lymph nodes identified and all negative for malignancy.

**Specimen:** Consists of a round piece of fibrous tissue. The specimen is oriented with a short suture marking the superior margin and a long suture marking the inferior margin. No skin ellipse. Resection margins intact. Serial sectioning reveals a nodular, solid mass with relatively ill-defined margins, measuring approximately 2.2 cm. Superficially, 2.0 cm, anteriorly, 2.0 cm, and posteriorly, 2.5 cm, medial to the node. The node closely borders on the anterior margin, and measures approximately 1 cm. The superior margin is 1.5 cm from the inferior margin, 1.5 cm from the lateral margin, 3 cm from the medial margin, and 0.2 cm from the posterior margin.

**Sections:** axe A, lateral margin, axe B, medial margin, axe C, superior margin, axe D, inferior margin, axe E, anterior margin, axe F, posterior margin, axe G, and axe H, posterior margin. axe Q, adjacent breast tissue.

**Specimen:** Consists of a piece measuring 13.6 x 1.5 cm. In maximal dimension. Lymph nodes are retrieved and submitted as follows. axe A, axe D, and axe C contain one lymph node divided between two cassettes. axe D, axe E, and axe F contain one lymph node divided between two cassettes. axe Q, adjacent breast tissue.

**Histologic findings:** are summarized in a synoptic report for invasive ductal carcinoma.

**Procedure:** 1. Biopsy of left breast, histologic type: invasive ductal carcinoma.

**Site:** Left breast.

**Histologic:** Invasive ductal carcinoma.

**Tumour size:** Invasive tumour measures 1.9 x 1.4 cm, in greatest dimension (microscopic measurements, block 1-4).
Shows how value was derived
Uses for Synoptex

- API for embedded synoptic extraction capability
- Add-on to E-path for automatic synoptic report generation
- Quality control tool for reporting institutions
- Automated data extraction to populate database for clinical trial matching system
- Other uses where synoptic summaries are needed (ICU report scanning)