cancer care ontario | action cancer ontario

# Inter-Rater Reliability Assessment and Electronic Collaborative Stage Data Collection in Ontario, Canada

*Bob Li, Mary Jane King, Gemma Lee, Eric Holowaty, Marta Yurcan, Jillian Ross, Kamini Milnes, Rami Rahal*

*Cancer Care Ontario, Canada*

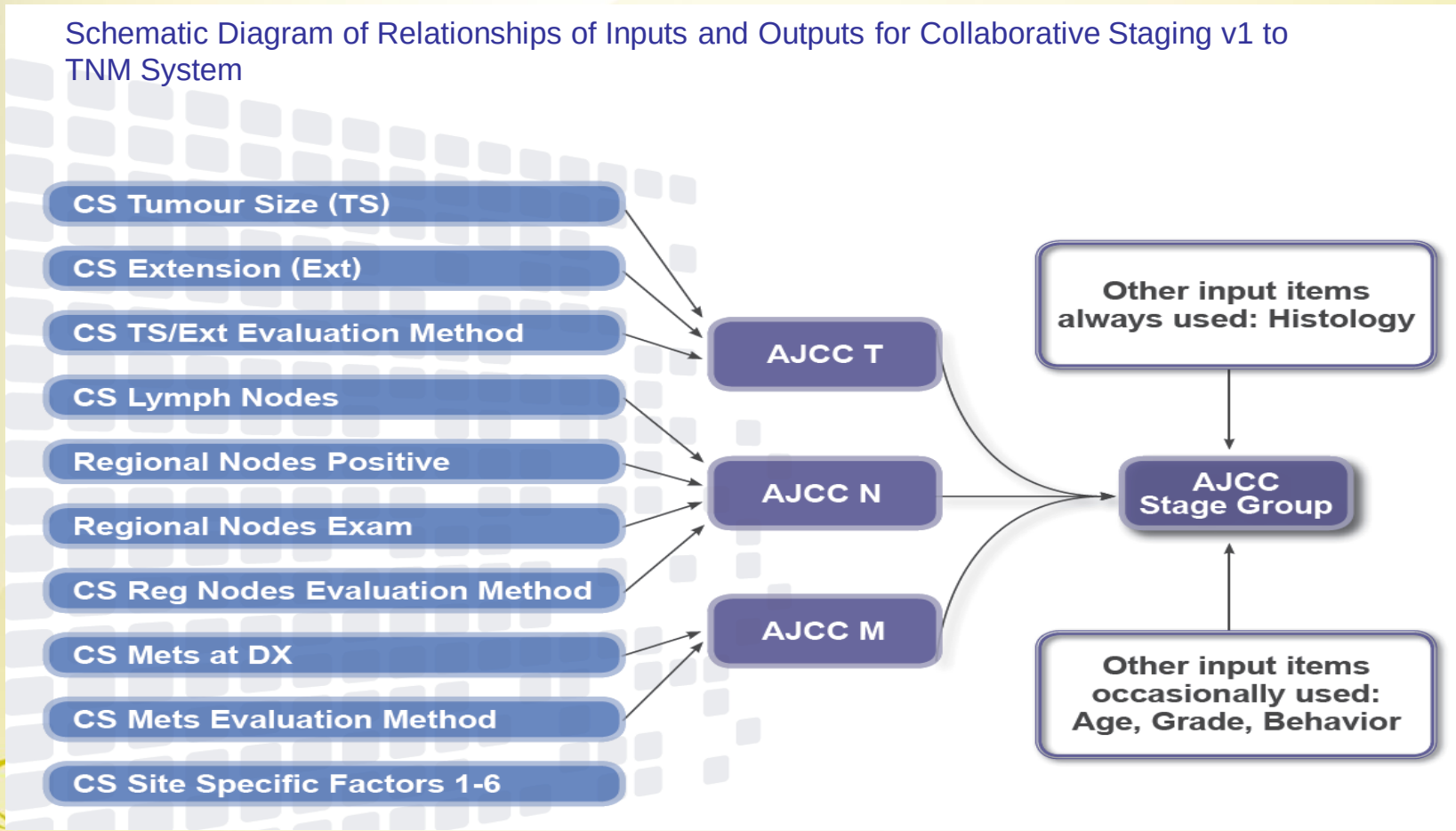
*NAACCR, Quebec City, June 22, 2010*

*Better cancer services every step of the way*

# CS is the new standard in Ontario for cancer staging beginning with 2007 cases

- CS is replacing TNM reporting to align with pan-Canadian guidelines

Schematic Diagram of Relationships of Inputs and Outputs for Collaborative Staging v1 to TNM System



# CS implementation in Ontario led by CCO with all acute care hospitals

- A provincial initiative that aims to improve the quality & completeness of cancer staging data in Ontario
- CS data collection system
  - CCO trained analysts
  - Remote access to hospital records
  - Registry Plus™ software developed by CDC
- CS data collection initiated in 71 non-Cancer Centre hospitals for top 4 disease sites from 2007 forward
  - 14 Cancer Centres to start with 2010 diagnosis year

# Robust data quality program Implemented to ensure CS data is of high quality

- Providing extensive ongoing training of centralized CCO analysts with support from the Public Health Agency of Canada
- Comprehensive array of DQ measures implemented to assess and monitor multiple dimensions of data quality
  - **Reliability**
  - Timeliness
  - Completeness
  - Validity
  - Usability
- Bi-annual reporting of stage data with chart level reporting for all staged cases is being provided to all participating hospitals

# Comprehensive inter-rater reliability assessment undertaken

- Measures the degree to which the same stage result can be obtained through re-abstraction process among CS analysts
- Assesses stage data quality for the dimension of reliability
- High inter-rater reliability assures data users that there has been consistent application of CS data collection system rules by CS analysts

# Process for Assessing Inter-rater Reliability (IRR)

1. Chose reliability indices/measures
2. Established an acceptable level of reliability
3. Tested selected reliability indices with a small subgroup of cases
4. Determined sample size of cases and randomly selected charts for IRR study
5. Calculated reliability results for 5 CS analysts involved in IRR study
6. Analyzed results and identified opportunity for improvement in CS abstraction

# Indices selected to assess inter-rater reliability

| Index | Multiple coders | Metric | Chance agreement | Agreement/ correlation |
|---|---|---|---|---|
| Percentage agreement | N | Nominal | N | Agreement |
| Scott's Pi | N | Nominal | Y | Agreement |
| Cohen's kappa | Y | Nominal | Y | Agreement |
| Krippendorff's Alpha | Y | Nominal,Ordinal ,Interval, Ratio | Y | Both |
| Perreault's Pi | N | Nominal | Y | Agreement |
| Spearman's Rho | Y | Ordinal | N | Correlation |
| Pearson's r | Y | Interval | N | Correlation |

## Two indices used to measure inter-rater reliability

- Modified Percent Agreement
- Krippendorff's alpha

# Modified Percent Agreement

- Percent Agreement= $(\frac{O_a}{O_a + O_d})*100$
  - Where $O_a$ is observed agreement; $O_d$ is observed disagreement
- Percent agreement = Number of agreements of group stage scores divided by the number of total decisions (agreements + disagreements)
  - Modified agreement is assumed when three or more of the five analysts produce the same stage value for a case.
- Advantages - easy to calculate even with multiple analysts; provides information at analyst level
- Limitations – does not consider agreement due to chance; therefore Krippendorf's alpha is also used to mitigate this issue

# Krippendorff Alpha

- Krippendorff's Alpha is expressed as

$$\alpha = 1 - \frac{D_o}{D_e}$$

Krippendorff (2007)

Where

Do = disagreement, observed

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck\ \text{metric}}\, \delta^2_{ck}$$

De = disagreement, expected by chance

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c \cdot n_{k\ \text{metric}}\, \delta^2_{ck}$$

- **Advantages** – can apply to multiple metrics, multiple coders; Considers agreement by chance, and for missing data
- **Limitations** – complex calculation, does not provide information at analyst level, more conservative measure relatively

# Inter Rater Reliability Assessment: Defining acceptable level of reliability

| Value | Level of Agreement |
|---|---|
| <=0 | none |
| 0.01-0.20 | poor |
| 0.21-0.40 | slight |
| 0.41-0.60 | fair |
| 0.61-0.80 | good |
| 0.81-0.92 | very good |
| 0.93-1.00 | excellent |

# Inter Rater Reliability Assessment: Sample Size Determination

An approximation to the sample size will be determined by following formula (Douglas G. Bonett, 2002)

$$n = 8z_{\alpha/2}^2\{(1-\tilde{\rho}_I)^2(1+(k-1)\tilde{\rho}_I)^2\}/\{k(k-1)w^2\}+1$$

- **Where**
- $Z_{a/2}^{\ 2}$ **is the point on a standard normal distribution exceeded with probability α/2**
- $\rho_i$ **is Intra-class correlation coefficient**
- **W is desired width obtained by setting** $w = 2z_{\alpha/2}(\text{var}\,\hat{\rho}_I)^{1/2}$
- **K is number of CS analysts**

# Inter Rater Reliability Assessment: Sample Size

- **Assumptions:**
  - Two-way random effects model was used in the analysis.
  - α=0.05, two-sided;
  - power=0.80;
  - $\rho_i$ is Intra-class correlation coefficient, it was from pilot study and is 0.70
  - W=0.20;
  - K=5
- Final total sample size for four disease sites is 198
- Random cases selected from hospitals and coded by five raters: 49 colorectal cancers (CRC), 49 lung cancers, 48 breast cancers and 52 prostate cancers.

# Results: Percent Agreement on Derived AJCC Group Stage

| Diagnosis Site | Coder 1 | Coder 2 | Coder 3 | Coder 4 | Coder 5 | All |
|---|---|---|---|---|---|---|
| BREAST | 91.3 | **65.2** | 97.8 | 84.8 | 91.3 | 86.1 |
| CRC | 98 | 92.2 | 96.1 | 88.2 | 92.2 | 93.3 |
| LUNG | 86.5 | 88.5 | 92.3 | 86.5 | 88.5 | 88.5 |
| PROSTATE | 94.2 | 88.5 | 94.2 | 96.2 | 94.2 | 93.5 |
| Total | 92.5 | **84.1** | 95 | 89.1 | 91.5 | 90.4 |

# Results: Krippendoff Alpha

| Scenario | Diagnosis Site | Alpha | CI 95% |
|---|---|---|---|
| **Scenario one:** no any exclusion, data as nominal | **CRC** | 0.85 | 0.82--0.89 |
| | **LUNG** | 0.69 | 0.63--0.74 |
| | **BREAST** | 0.69 | 0.64--0.73 |
| | **PROSTATE** | 0.79 | 0.74--0.84 |
| | **All four diagnosis** | 0.8 | 0.78--0.82 |
| **Scenario two:** stage unknown as missing, data as nominal | **CRC** | 0.92 | 0.89--0.94 |
| | **LUNG** | 0.76 | 0.71--0.80 |
| | **BREAST** | 0.89 | 0.85--0.93 |
| | **PROSTATE** | 0.91 | 0.86--0.95 |
| | **All four diagnosis** | 0.89 | 0.88--0.91 |
| **Scenario three:** stage unknown as missing, stage value was truncated as ordinal (0,1,2,3,4) | **CRC** | 0.98 | 0.96--0.99 |
| | **LUNG** | 0.88 | 0.84--0.91 |
| | **BREAST** | 0.95 | 0.94--0.97 |
| | **PROSTATE** | 0.94 | 0.90--0.98 |
| | **All four diagnosis** | 0.96 | 0.95--0.97 |

# Findings

- "Good" or "Very Good" level of reliability was found in the abstraction of CS data
- Opportunities to improve CS data collection processes were identified and mplemented
  - Provided additional education on application of rules to assign Stage Unknown for breast and lung cancer cases (esp for one CS analyst)
  - Introduced tools to monitor participation of CS analysts in education and training (and processes for follow-up if sessions are not attended)
  - Created better access to training materials and resources for CS analysts to apply in daily practice

# Next steps:

- Establishing process for performing inter-rater reliability (IRR) assessment on regular basis
- Reviewing whether future assessments should consider options for a gold standard based audit i.e., comparison of physician staging

# Questions?

bob.li@cancercare.on.ca