#### Clinical Data Work Group Highlights January 18, 2011

**Present:** Jim Martin, Ken Gerlach, Barry Gordon, Isaac Hands, Bruce Riddle, Leon Sun, Roman Tsyvine, Joe Rogers, Chuck May

#### NAACCR Staff: Lori Havener

January 4 minutes approved.

**Plan for the Beta Implementation of NAACCR XML:** The WG reviewed the plan that Leon developed which includes information from the first pilot project, our current work plan (challenges, tasks, registry vocabulary, templates and deliverables), testing phase and then future topics for discussion. Updates were made to the following sections (see bottom of minutes): Final Deliverables 2, 3 and 5; Challenges of data transmission 1, 2, 3, 4; Details of Phase II, Scope; and, Topics for future discussion.

Joe shared some of his thoughts on the process. There is a need for two standards: 1) a CDA that EMR are familiar with that cancer registries can consume and send; and, 2) a NAACCR-specific XML to transmit data among the cancer registry community. Cancer registries will need an import tool for CDA messages to interpret and parse the data to process at the cancer registry. A statement was added to the scope that the WG will work on the CDA message after the XML project is completed.

*Action:* Lori will post the updated document (see bottom of minutes) to SharePoint for review/comment by the WG.

Next Meeting: Tuesday, February 1 at 3:30 pm ET.

# NAACCR-specific XML data transmission standard for cancer registry communities

Version: January 18, 2011

#### **Objectives**

- 1. To improve data transmission interoperability and cost-efficiency
- 2. To provide a set of standard, technique, software tools, and implementation guideline for data transmission in XML format over the web
- 3. ....

## Final deliverables:

- 1. A set of XML format data transmission standards and implementation guideline
- A set of web-based tools to register, maintain and manage NAACCR XML schemas/templates, , and NAACCR-specific vocabulary for data transmission in XML format as the source of data items, templates, checking consistency and conformance with constraints, and validation rules
- 3. Combine to form a technology to assist vendors' software with the function that directly transits data, rather than use flat-file format, over the web among senders and recipients in cancer registry community
- 4. Sets of sample data used for the above bidirectional transmission
- 5. Bidirectional translational software tool
- 6. .....

## Challenges of data transmission and the necessity to use XML format to standardize the process

There is a need to standardize data transmission in cancer registry community to solve the challenging issues below for improvement of data transmission accuracy, flexibility, and cost-effectiveness

- 1. Multiple record formats: At present, five out of six NAACCR record types are still transmitted in flat file format between senders and recipients, while HL7 and XML are at present (and the future) widely used for clinical data transmission and manipulation over the web.
- 2. Losing richness of data: NAACCR record type L data that is converted by eMaRC Plus and E-Path to the format of HL7 files and messages before transmission has a higher rate of data autocoding and autopopulation compared to other NAACCR records types not in HL7 format.
- Outmoded formatting technology: Current data transmission formats, comma-, pipe- or space-delimited, are venders' software-dependent. A universal standard technology such as XML that is the most common tool for data transmissions between all sorts of software tools and platforms would improve data transmission
- 4. Constrained record length: Current NAACCR records are transmitted by entire record blocks and its transmission over the web may be constrained by record length (A-22824, I-3339, C-5564, L- uncertain, U-1543, M-22824). Data transmission in XML format transmission would solve this issue by truncating the individual over-length limit data items without affecting entry block transmission
- 5. Institution-specific data transmission between senders and recipients in cancer registry community must be considered. There is a need to build and maintain a central registry data item repository to standardize and facilitate these data items' transmission

- 6. At present, data transmission in cancer registry community is largely unidirectional, from hospitals/senders to central registries/recipients only. There is a need for bidirectional data transmission.
- 7. Current data transmission flow is that field-driven data retrieved from sender's database is firstly converted to flat files before sending, and the recipient converts the flat files into field-driven format then maps them to the recipient's database. Theoretically and technically, it is possible to eliminate the flat file conversion step by directly transmitting field-driven data between senders and recipients by using XML format.
- 8. ....

# <u>Main tasks</u>

This project will focus on development, test and validation of the following three components through all the phases.

- 1. XML schemas (or templates), a standard format guiding the collection of key data elements and providing additional validation
- 2. Clinical statement models, which ensures a common representation of data from the clinical data categories such as medications, family history, allergies, life styles, signs and symptoms, biomarkers, diagnosis, comorbidities, treatments, pathology/laboratory, imagings, recurrences, metastasis, follow-up and death, etc., and fosters data reusability
- 3. NAACCR-specific vocabulary (or terminology) for data transmission in XML format (elements, attributes, definition, usage, etc), which is compatible the industry-standard definition/concepts such as NCI CDE, NCI Thesaurus, CDC's PHIN VARDS, CAP, NAACCR volume I and II, CDA, HL7, LOINC, SNOMED, etc
- 4. Web-based tools that maintain and manage the above three tasks, reflect the need of data items by cancer registry community, are updated timely, and have a long term sustaining plan
- 5. .....

## Strategic process and expected outcomes

This project will be conducted in the following phases.

## 1. Phase I. Concept proof (completed).

- 1) Pilot project of hospital cancer registry data transmission to the central cancer registry using Health Level Seven (HL7) and Clinical Document Architecture (CDA)
- 2) HL7 CDA draft implementation guide to pilot test HL7 CDA transmitting data from a hospital cancer registry to a central cancer registry through a contractual agreement between CDC-NPCR- Alschuler Associates, LLC
- 3) Sources/deliverables derived from this phase include
  - a. HL7 CDA schemas/templates
  - b. Codes and programs for rules processing transmitted NAACCR data items (required, required when available, optional, null, truncation on data items over length limit, etc)
  - c. Transformation software tool that converts NAACCR flat-file to HL7 CDA format or vice versa
  - d. Database that stores and manages XML schemas (HL7 CDA), clinical models, data items transmitted, California state-specific data items
  - e. ....

- 2. Phase II. System development, beta testing, validation and analysis on cost-effectiveness In this phase, the following tasks will be completed (see "Phase II: System development and beta testing" below for details)
  - A web-based tool for management of XML schemas (NAACCRspecific.xsd, NAACCRspecific.DTD)/templates, clinical statement model and NAACCR-specific vocabulary for data transmission in XML format will be developed and populated with previous and needed elements and their attributes.
  - 2) Conducting a beta testing and validation of data transmission in XML format among a few registries and hospitals
  - Analysis of data transmission patterns in flat-file and XML format in terms of data transmission time, accuracy, consistency, cost-effectiveness for software design and development
  - 4) Technique to bidirectionally covert data between the formats of flat files and XML files, and directly transmit field-driven data items between sender's and recipient's databases in XML format
  - 5) NAACCR-specific XML data transmission standard and implementation guide
  - 6) .....

# 3. Phase III. Implementation of NAACCR-specific data transmission in XML format

- 1) To publish the NAACCR-specific XML data transmission standard and implementation guideline
- 2) To allow access of local and regional registries and software vendors to the web-based tools to search, develop, define, register and manage XML schemas/templates, clinical statement models and institution-specific data items
- 3) To open the source-codes, programs, software tools, sample data sets, etc to venders to facilitate software development and implementation
- 4) To synchronize with other software tools (ie, to allow EDITS to read and validate data in XML formatted file, or automatic deploying industry-standard terminology such as CDA, HL7, LOINC, SNOMED, CDC's PHIN VADS, NCI Thesaurus, etc)
- 5) ....

# **Operational (for Leon only)**

- 1. Make it as a national collaborative project (similar but smaller than CS project)
- 2. Get funding from standard setters/sponsors through contacts with one or more vendors
- 3. Develop and deliver a set of standard and implementation guidelines, software tools, sample data for bidirectional data transmission over the web in HL7 XML format

## Details of Phase II. System development, beta testing, validation and analysis on costeffectiveness

Draft Date: January 4, 2011

**Scope**: To create a NAACCR-specific XML data transmission format for use within the cancer registry community; which could be converted or transformed to NAACCR CDA in order to interface with electronic healthcare systems, e.g. electronic health records (EHR), health information exchanges. The NAACCR-specific XML and HL7 CDA are closely tied together; CDA standards are too fluid and will become more standardized in the near future. At that time it would be appropriate at getting an HL7 CDA translator for documents received from EHRs to be able to populate NAACCR cases in the database.

#### <u>Tasks:</u>

- 1. NAACCR-specific XML file performance goals
  - 1) Optimal XML file size and XML file processing speed
  - 2) Optimal batch data transmission
    - a. Number of records/batch constrained by file size in mode of transmission of record types
    - b. Data item numbers/batch constrained by file size in mode of field-driven data transmission
- 2. NAACCR-specific XML file rules controlling data items
  - 1) Output of optional data items, data items with missing/unknown value
  - 2) Number of repeats (race, medications, labs, comorbidities, etc)
- 3. NAACCR-specific XML schemas (xsd/DTD)/templates
  - 1) A database to store XML schemas/templates and rules
  - 2) One-to-one layout to store detailed schema structure information including name, definition, application for NAACCR record types, versions, header, body, elements and their attributes, nested elements, etc
  - 3) One-to-many layout to store schema/template applications in different clinical courses such as medications, family history, allergies, life styles, signs and symptoms, biomarkers, diagnosis, comorbidities, treatments, pathology/laboratory, imaging, recurrences, metastasis, follow-up and death, etc (clinical statement models)
- 4. NAACCR-specific vocabulary
  - 1) A database to store and map all data items in NAACCR volume II, plus hospital-specific and state-specific data items
  - One-to-one layout to store each of the needed data items including data item ID, name, short description, long description, definition, data type, size, constraints (required, optional, null), default value, element and attributes in XML format
  - Multiple one-to-many layout including data item ID, XML format, clinical statement models, ID in industry-standard terminology (NAACCR volume II, HL7, CDA, LOINC, SNOMED CT, NCI Thesaurus, CDC's PHIN VARDS, CAP, NCI CDE, codes of ICD9, CPT/HCPC, etc)
  - 4) List of clinical statements (medications, family history, allergies, life styles, signs and symptoms, biomarkers, diagnosis, comorbidities, treatments, pathology/laboratory, imagings, recurrences, metastasis, follow-up and death, etc)
  - 5) .....

- 5. Beta implementation test including NAACCR-specific data set retrieval and export from sender's database, XML file creation, transport over the web, transform by XSLT, and import to the recipient's database with appropriate metrics
  - 1) Participants. The central registries and their hospitals will be the ones using the following data collection software:
    - a.Rocky Mountain Cancer Data Systems versus Abstract Plus
    - b. SEER\*DMS vs SEER\*Abs
    - c. Eureka vs C/NExT
  - 2) Tools/scripts needed for data format conversion, encryption and transmission:
    - a.XSLT files to transform XML files to other data format
    - b. Conversion tool to transform data files between the formats of flat files and XML files
    - c.SQL scripts and conversion tool to extract field-driven data items needed for different NAACCR record types from sender's database, convert and send the data items in XML format over the web, then map the XML file by using XSLT and import the field-driven data items to recipient's database
    - d. Tools for data encryption, authenticity, compression, and transport
    - e.Tool for browsing different types of cancer records in XML format
    - f. Tool for validation
  - 3) Document preparation
    - a.NAACCR-specific XML data transmission standards
    - b. NAACCR-specific XML file encryption, authenticity and compression
    - c. NAACCR-specific XML data transmission standards implementation guidelines
  - 4) Web systems used for live data transmission test in XML format FTP, PHINMS, CONNECT, Web Plus, SSH
  - 5) Conduct bidirectional data transmission over the web between senders and recipients
    - a.Data files converted between the formats of flat files and XML files
    - b. Individual data items in the blocks of NAACCR record types are directly transmitted over the web in XML format
    - c. Validation of transmission accuracy, consistency, and efficiency of NAACCR data items, as well as institution-specific data items
- 6. Post-transmission
  - 1) Compile test metrics, analyze, and report
  - 2) Beta implementers submit metrics and experiences to WG
  - 3) List metrics and experiences
  - 4) Steps to go from Beta implementation to full implementation
  - 5) Costs to set-up, maintenance
  - 6) Recommendations
- 7. Topics for future discussion or Phase III of this project
  - I. Allow EDITS to read NAACCR XML (with and without transformation to a column-delimited format)
  - II. Capability to transform to NAACCR CDA
- III. Identify XML attributes to facilitate transformation to CDA
- IV. NAACCR CDA implementation guide
- V. Assess public health and cancer research uses of CDA and related tools
- VI. Assess demand/need for the EHR/HIE community to receive hospital and/or central cancer registry abstract reports

- VII. Identify NAACCR data items needed for transmission to the EHR/HIE communities
- VIII. Assess the need/demand for the EHR/HIE communities to received state-specific data items
- IX. Could test as an IHE Profile, would need EHR vendor buy-in
- X. Assess the NAACCR-specific vocabulary for data items and OIDs
- XI. Test and implementation of data transmission in XML format between central tumor registry and physician offices

# Plan for the -Implementation of NAACCR XML with Tasks List

Draft of January 4, 2011

#### Steps:

- I. Requirement Gathering
  - A. Establish targets to validate optimal benchmark NAACCR XML file size and processing speed
  - B. Batch capability
    - 1. Establish an upper limit on the number of records that should be processed in one batch operation, the lower limit will be 1. How many cases can be handled at one time?
    - 2. Define batch operations as use cases, such as "Export cancer abstracts to file" and "Import cancer abstracts into database"
    - 3. Explore the possibility of including metadata in some part of the XML file that describes the file's contents
  - C. Repeating functionality for data items, e.g. Race, and section, e.g. Treatment [NOTE: Revisit because supplemental treatment information will change business model.]
  - D. Decision for transmission of data items with values unknown
  - E. Decision for structure of XML elements and attributes
    - 1. Review Pilot (Alschuler) CDA IG and the associated XML format used to transform from CDA to NAACCR Flat
  - F. NAACCR XML to NAACCR Flat and vice versa software transforms
  - G. NAACCR XML implementation guide
  - H. Establish NAACCR XML Transmission Guidelines such as encryption, authenticity, compression, and transport
  - I. Assess structure for EDITS to read NAACCR XML (with and without transformation to a column-delimited format)
- II. Specification Development
  - A. Estimate size of NAACCR XML compared to NAACCR Flat (base-line metrics)
  - B. Develop NAACCR XML Implementation Guide
  - C. Decision/Develop DTD (document type definitions) or Schema for NAACCR XML (DTD to defines document structure with a list of legal elements and attributes).
  - D. Develop validation tools
  - E. Develop XSLT to transform XML to other forms (proof of concept)
  - F. Write and test NAACCR XML to NAACCR Flat and vice versa software transforms
  - G. Beta implementers to develop export and import software for use in hospital or central cancer registry software.
  - H. Develop or identify a tool(s) for browsing XML Cancer Abstracts
- III. Transmission Test
  - A. Identify data sources to test: simulated and active
  - B. Beta implementers test NAACCR XML creation/import/export/validate/transform capability with appropriate metrics
  - C. Conduct live transmission test (FTP, PHINMS, CONNECT, Web Plus, SSH)

- IV. Compile Test Metrics, Analyze, and Report
  - A. Beta implementers submit metrics and experiences to WG
  - B. Metrics and experiences into a Report
    - 1. List metrics and experiences
    - 2. Steps to go from Beta implementation to full implementation
    - 3. Costs to set-up, maintenance
    - 4. Recommendations

Topics for future discussion or Phase II of this project

- XII. Capability to transform to NAACCR CDA
  - A. Identify XML attributes to facilitate transformation to CDA
  - B. Assess demand/need for the EHR/HIE community to receive hospital and/or central cancer registry abstract reports
  - C. Identify NAACCR data items needed for transmission to the EHR/HIE communities
  - D. Assess the need/demand for the EHR/HIE communities to received statespecific data items
  - E. Could test as an IHE Profile, would need EHR vendor buy-in
  - F. Assess need for OIDs and establishment of NAACCR OIDs Registry (Web-based tool)
  - G. NAACCR CDA implementation guide
  - H. Assess public health and cancer research uses of CDA and related tools

A set of web-based tools to register, maintain and manage clinical statement models