CAIR2 Patient Matching: Solving the 25 Million Piece Puzzle

Michael Powell, Immunization Technical Unit Chief, CDPH
Mike Berry, Senior Project Manager, HLN Consulting, LLC
Gary Wheeler, Account Executive and Strategist, DXC Technology

Solving the 25 Million Piece Puzzle

- Overview
- Approach
- RunMatch Analysis
- Conclusions
- Next Steps

California Department of Public Health CDPH

HLN Consulting, LLC

DXC.technology

CAIR2 - *Patients and Doses**

Measure	*0-5 yrs*	*6-18 yrs*	*19+ yrs*	*All Ages*
CA Population	2,629,503	5,733,497	26,745,104	35,108,104
Patients In	3,354,573	5,996,008	17,682,549	27,033,130
% of Pop. In	128%	105%	66%	77%
Patients w/ ≥2 doses	2,077,280	5,373,248	11,700,579	19,151,107
% w/ ≥2 doses	79%	94%	44%	55%
Vaccine Doses	43,216,228	117,866,058	88,645,241	249,727,527

* As of 7/9/2018. CAIR2 only.

Solving the 25 Million Piece Puzzle

- Problem solving
  - Gather information and knowledge
  - Identify the problem
  - Develop Criteria
  - Generate Possible Solutions
  - Analyze Possible Solutions
  - Compare Possible Solutions
  - Make and Implement the Decision

California Department of Public Health CDPH

HLN Consulting LLC

DXC.technology

Solving the 25 Million Piece Puzzle

- Matching Algorithm
  - Designed for UI
  - Majority of CAIR2 doses coming in through DX
- Pendings
  - Bug in Migration
  - Unmanageable
- “Ghost” dups
- Collaborate

California Department of Public Health CDPH

HLN Consulting LLC

DXC.technology

RunMatch Analysis: Introduction

- Objectives:
  - **Examine** CAIR's RunMatch source code and documentation to identify possible inefficiencies, functional shortcomings, or areas for improvement
  - **Experiment** with RunMatch and its capabilities to determine if configuration or functional issues could be causing person-matching issues for CAIR
- Inputs:
  - RunMatch Design document
  - RunMatch Logic and scoring flowcharts
  - RunMatch source code (14,000 lines of C)

RunMatch Analysis: High-level Observations

- Generally: Deterministic, Probabilistic, Machine learning approaches
- Many real-world matching engines are hybrid
- RunMatch has both Deterministic & Probabilistic attributes
- Advantages and disadvantages to each approach
- **Common challenge**: Keeping up with changing data characteristics

California Department of Public Health CDPH

HLN Consulting, LLC

DXC.technology

RunMatch Analysis: Testing Strategy

- Compile RunMatch from source
- Create Oracle database with CAIR tables for RunMatch operation
- Create custom RunMatch client with CSV interface
- Configure Febrl (open source probabilistic matching engine) for comparison
- Run tests against RunMatch and Febrl using:
  - ONC Patient Matching Challenge dataset
  - Custom test cases based on observations from the results

California Department of Public Health CDPH

HLN Consulting LLC

DXC.technology

RunMatch Analysis: Findings

- Strengths
  - Very Fast
  - Relatively low resource requirements (CPU, RAM, etc.)
  - Very good at handling common typos, transpositions, many special cases
  - Good overall match performance compared to Febrl
  - Token configuration can be customized without recompiling
- Weaknesses
  - Complex rule-based model with numerous exceptions / special rules
  - Name string matching algorithm has some specific weak areas compared to edit-distance algorithms such as Jaro-Winkler
  - Lacks built-in deduplication functionality

California Department of Public Health CDPH

HLN Consulting LLC

DXC.technology

RunMatch Analysis: Potential Improvements

- In the CAIR installation:
  - Update names and frequencies in token files
  - Add local cities to token files
  - Use result messages and scores from RunMatch to tweak configuration files
- In the RunMatch software
  - Redirect RunMatch Server output to database to facilitate post-match analysis
  - Human review feature for batch imports
  - Incorporate edit-distance algorithm(s) into RunMatch string-near-matching

California Department of Public Health CDPH

HLN Consulting, LLC

DXC.technology

Moving Forward - Collaboration and Planning

- Review Results, Evolution of RunMatch – Improvement vs. Replacement
- Maximizing Results, Dual Path
- State-Specific Changes, Scoring Adjustments
- RunMatch Enhancement Project Launch

California Department of Public Health CDPH

HLN Consulting LLC

DXC.technology

Project Goals

- Improve access to algorithm results
- Reduce manual intervention (multiple matches)
- Improve algorithm maintainability while
- sustaining performance
- Additional matching criteria

- Working together through joint development

California Department of Public Health CDPH

HLN Consulting LLC

DXC.technology

Project Highlights

- Project commencement March 2018
- DXC funded – client driven
- Replacing C code with Java (>14K lines of code)
- Improvements Include:
  - Configurability – Scoring Adjustments
  - Data Availability, Human Readable Logs
  - Enhanced Ethnic Logic, Calculations based on IIS Population
  - Chart # Logic
  - Matching Test Suite, Test Rules and Scoring Changes

California Department of Public Health CDPH

HLN Consulting LLC

DXC.technology

Next Steps/Conclusions

- Pilot Testing (CA/NE – In Flight)
- Continued Criteria Improvement
- Near name matching
- Addressing address
- Exact match enhancements
- Key Lessons
  - Matching is complex, no easy answers
  - Adjusting for volume of submissions and data patterns is critical
  - Access and understanding key to making informed decisions
  - Better together!!!

California Department of Public Health CDPH

HLN Consulting LLC

DXC.technology

Contact Information

Michael S. Powell, MSc
Michael.Powell@cdph.ca.gov

Mike Berry
berrym@hln.com

Gary Wheeler
Gary.Wheeler@dxc.com

California Department of Public Health CDPH

HLN Consulting LLC

DXC.technology