Building a Breast Cancer Learning Health System

Jeremy Petch, PhD[1], Christopher Pettengell, BMBCh[2], Jessica Bogach, MD MSc[1], Alexandra Allard-Coutu, MDCM MSc[3], Steven Aviv[2], Gregory R. Pond, PhD[3], Joel Kemppainen[1], Mark N. Levine, MD MSc[3]

1. Hamilton Health Sciences, Hamilton, Ontario, Canada; 2. Pentavere Research Group Inc, Toronto, Ontario, Canada; 3. McMaster University, Hamilton, Ontario, Canada.

Introduction

Day-to-day clinical practice generates large volumes of valuable data that can be used to describe the impact of delivered healthcare on patient outcomes. Learning Health Systems (LHS) have been proposed to leverage this data to its fullest potential, acting as a continuous cycle of scientific evidence informing clinical practice and data captured through clinical practice informing further scientific investigation. However, the latter part of the loop is often missing, and such information is not readily available; data captured in clinical notes are often siloed and/or unstructured, making them unsuitable for analysis with conventional statistical techniques. Manual methods of reviewing these notes to capture patient experience and outcome data are time-consuming and are limited by the scale and quality of data that can be captured.

Objective

Develop a data platform to enable a LHS by applying artificial intelligence (AI) engine, DARWEN™, to clinical documents to characterize the clinical course of breast cancer patients.

Methodology

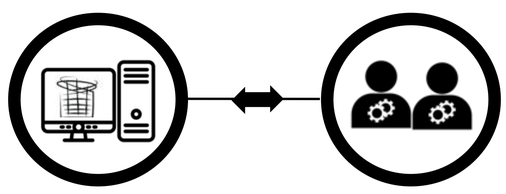
Building the LHS

This study included breast cancer patients seen at the Hamilton Health Sciences (HHS) Juravinski Cancer Centre (JCC) between 2014 and 2018 with at least two years of follow-up. The generated data platform was composed of structured and unstructured data which was extracted using Microsoft SSIS from 6 repositories, or "data silos", shown in **Figure 1**. DARWEN™ AI, developed by Pentavere Research Group Inc, was deployed within the JCC to automate data abstraction from unstructured clinical documents. AI-abstracted data was populated alongside structured data extracts into a longitudinal patient-oriented data warehouse, updated nightly.

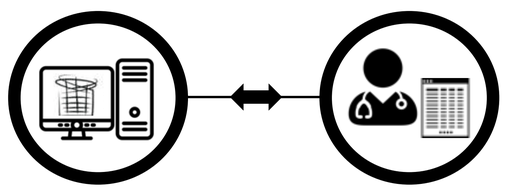
Ongoing Quality Assurance

Two strategies were used for AI quality assurance:

1.
No structured data available: Compare DARWEN™ outputs to manual extraction done by cancer surgeons from the JCC.



2.
Structured data sometimes available: Compare DARWEN™ with structured source (e.g., synoptic reports) *when available*.



Results

Extracted data for this patient cohort is displayed in **Table 1**. AI accuracy varied between extracted features, but remained high, for example: F1=0.95 for ER status (n=1094), 0.92 for PR status (n=1094), and 0.83 for HER2 status (n=946). Manual quality assurance done by cancer surgeons took a mean of 20 minutes, suggesting review of the entire cohort would take ~800 hours. After AI model development, the AI processing of 2339 patient records was completed in ~8 hours running on a 4 core Intel Gold 6248 CPU @ 2.50 GHz server. Upfront time savings using AI were relatively modest due to the time required for model development and tuning, but the ongoing time savings are considerable: **a subsequent data extraction for the 3,464 new patients seen at the JCC between 2019 and June of 2022 was completed in ~12 hours.**

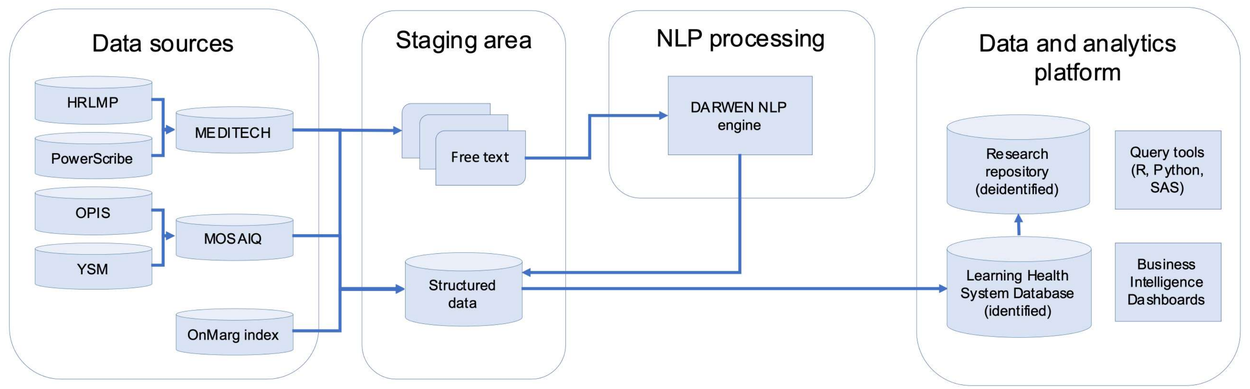


Figure 1. The LHS platform. Above is an outline of the data flow from the raw source (left) to a structured, longitudinal, patient-oriented database (right) which is fit for use.

Table 1. Patient characteristics (N=2339).

Characteristic	Result N = 2339
Age (years)	
Median (Range)	61 (24-97)
Sex	
Female	2320
Male	19
Tumour Histology*	
Invasive ductal (with DCIS/LCIS)	1872 (895)
Invasive lobular (with DCIS/LCIS)	185 (82)
Mixed (with DCIS/LCIS)	77 (32)
Other (with DCIS/LCIS)	62 (13)
DCIS alone	90
LCIS alone	4
Unknown	49
Biomarker*,‡	
ER (+/-/unknown)	1828 / 328 / 183
PR (+/-/unknown)	1452 / 582 / 305
Her2 neu (+/- / indeterminant / unknown)	393 / 1498 / 8 / 440
Triple negative	179
Surgery†,‡	
Breast Conservation	1260
Mastectomy	617
Modified Radical Mastectomy	467
Axillary Node Dissection or Sentinel Node Biopsy	1976
Positive	604
Negative	1372
Comorbidities†,‡	
Atrial Fibrillation	187
Stroke	131
Hypertension	1047
Chronic Obstructive Pulmonary Disease	153
Coronary Artery Disease	204
Diabetes	475
Patients with ESAS at Baseline	2092

DCIS, Ductal carcinoma in situ; ER, Estrogen receptor; ESAS, Edmonton Symptom Assessment Scale; LCIS, Lobular carcinoma in situ; PR, Progesterone receptor.

* Data was derived to minimize missingness or unknown. Used structured pathology data and then missing or unknown elements supplemented using DARWEN™ from the clinical notes.
† Data elements extracted from unstructured clinical notes using DARWEN™.
‡ Patients could have multiple values and may be included in each category more than once.

Conclusion

Automated integration of AI-extracted clinical data from documents across patient records is possible and is a preferred method of supporting a functional LHS. This comprehensive system will empower clinicians to leverage high-quality real-world data to supplement clinical decision-making and research efforts.

Acknowledgements

This work was supported by a grant from Roche Canada. We would like to thank the following individuals who provided support during the conceptualization and development stages of this work: Bill Butler, Paul Brown, Julie Bosworth, Duane Bender, Tim Dietrich, Ashirbani Saha, Peter Sztur, Jonathan Ranisau, Ted Scott.