Behind the scenes of a clinical trial, *circa* 2012

**Study question & hypothesis**
- **EAT:** Does 8-week overfeeding impair insulin sensitivity in humans?

**Protocol design & set up**
- Study Coordinator/Postdoc/PhD student/RA...
- Ethics approval
- Inclusion and exclusion criteria
- Protocols: SAP, MOPs, SOPs...

**Screening & Recruitment**

**Intervention & data collection**
- Retention and attrition

**Data management & analyses**
- Data Manager/Postdoc/PhD student/RA
- Master dataset(s)

**Trial end**

The University of Sydney
Incorporating digital technology into clinical trial workflow

Many companies offer products that span across categories. To keep the map simple, the logo is in the “primary” product. Here are the products in the Mind map.

Trial Administration
- Patient Recruitment
- Drug Supply Logistics
- Patient and outcome data management

Software-Enabled Clinical Trials

Virtual Trials
- Protocol design and review
- Site selection & start-up
- Operational management

Patient-Level Data Collection
- Patient and outcome data management

Opportunities in clinical trials

- Identification and recruitment of eligible patients using computable phenotypes
- Improved retention and attrition of patients
- Reduced costs of data collection
- Scalability and reproducibility
- Improved representation from under-represented populations
- Follow up of patients beyond clinical trials

Incorporating digital technology into clinical trial workflow

Paradigm shift is enabled by ever increasing sources of digital health data

Increasing patient recruitment in ANZ

The University of Sydney

https://digital.hbs.edu/innovation-disruption/software-enabled-clinical-trials/
PUBLISHED 30TH SEPTEMBER 2019

Opportunities in clinical trials

- Identification and recruitment of eligible patients using computable phenotypes
- Improved retention and attrition of patients
- Reduced costs of data collection
- Scalability and reproducibility
- Improved representation from under-represented populations
- Follow up of patients beyond clinical trials

Comparing data from clinical trials with other sources of data

<table>
<thead>
<tr>
<th></th>
<th>Clinical trial</th>
<th>Administrative Data Sets (eg. Claims, Admitted, Non-Admitted Patient collections, etc)</th>
<th>eMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original purpose of data collection</td>
<td>Research</td>
<td>Primary data collection</td>
<td>Activity-based funding</td>
</tr>
<tr>
<td>Cross-sectional, prospective cohort</td>
<td>Cross-sectional, prospective cohort</td>
<td>Retrospective, cross-sectional</td>
<td>Cross-sectional, retrospective or prospective cohort</td>
</tr>
<tr>
<td>Cost</td>
<td>$$$$ primarily government funded</td>
<td>$$$ primarily government funded</td>
<td>$ data collection is funded by health care systems; research can be funded by a variety of sources or may not require funding at all</td>
</tr>
<tr>
<td>Access</td>
<td>Researchers with ethics and governance approval</td>
<td>From data warehouse by analysts/data managers in hospitals/industry, researchers with ethics and governance approval</td>
<td>Researchers with ethics and governance approval</td>
</tr>
<tr>
<td>Time frame</td>
<td>Follow-up is restricted by funding, must wait for health outcomes to occur for longitudinal studies</td>
<td>Retrospective with datasets updated daily</td>
<td>Retrospective restricted by date of eMR implementation, additional data collected at low cost</td>
</tr>
<tr>
<td>Study population and follow-up</td>
<td>Based on recruitment, follow-up is scheduled</td>
<td>Curated and aggregated data from eMR systems across a specified population (eg. NSW)</td>
<td>Based on patient use of the specific health system and information system</td>
</tr>
<tr>
<td>Data collection and storage</td>
<td>Established protocols with robust, well-documented approaches to data collection</td>
<td>Established protocols with robust, well-documented approaches to data collection (eg. from eMR)</td>
<td>Little to no established protocols and no data dictionary</td>
</tr>
<tr>
<td>Conditions captured</td>
<td>Any outcomes as specified at the beginning of the study</td>
<td>Captured coded diagnoses and procedures and selected outcomes (eg. length of stay, in-hospital mortality)</td>
<td>Only outcomes requiring care by a clinician, missingness of data due to non-recording</td>
</tr>
<tr>
<td>Summary of pros and cons</td>
<td>High quality data in a pre-specified study population, not-real time, expensive to maintain and extend</td>
<td>State-wide coverage, updated daily and used for reporting, doesn’t capture clinically meaningful information</td>
<td>Potential to be used in real-time applications, no established protocols for data collection or quality, big data - volume, velocity, variety, value, scalable and reproducible and cost-effective</td>
</tr>
</tbody>
</table>

* data collected from devices (eg. wearables, apps)

**SPEED-EXTRACT (STEMI Patient ElEctronic Data EXTRACTion)**

Proof-of-concept project funded by Ministry of Health July 2018-December 2019

**Primary Aim**

To demonstrate the feasibility of accurately identifying (>90%) patients with ST Elevation Myocardial Infarction (STEMI) from existing suspected Acute Coronary Syndrome data, that reside in electronic medical record systems (EMR) from one quaternary and two feeder hospitals from Northern Sydney LHD within the Sydney Health Partners collaborative.

**Specific aims**

1. To develop a process and method for standardised extraction of eMR data to identify patients with a discharge diagnosis of (STEMI)
2. To determine the extent to which extracted data elements can be used to describe clinical quality measures
3. To share this identified cohort and proposed quality measures with practicing clinicians to ensure ‘face validity’ of the extracted data
**Mapping data extraction to reconstruct patient journeys for patients with chest pain**

**RNSH**
- 57y male presents to ED at RNS by ambulance with chest pain at 07:54
- ECG performed in ambulance
- 75 pathology tests
- 2 haTropin tests performed 7 hours apart
- Patient admitted and discharged from ED
- Discharged at 16:49, 8.9 hours later with discharge diagnosis of chest pain from ED

**RNSH #1**
- 70y male presents to ED at RNS by ambulance with chest pain at 21:37
- Reviewed an ECG in the ambulance
- Presenting information at ED Triage: "CHEST PAIN MEETS ETAMI CRITERIA, ST ELEVATION, ANTERIOR LATERAL ST ELEVATION, BATPHONE FOR SAME"
- Patient arrives to cath lab at 21:57
- Percutaneous coronary intervention performed with door to balloon time <90 minutes
- Cardiology Ward
- LOS = 2.7 days
- ICD10 code: acute transmural myocardial infarction of anterior wall
- Discharge diagnosis: STEMI (JMO)
- Discharge letter - statin, newly diagnosed diabetes

**RNSH #2**
- Present to ED at RNS by ambulance for chest pain
- Readmitted 17.1 days after, spends 22 hours in ED
- 2 ECGs
- 76 pathology tests (2 haTropin tests)
- 70 forms
- LOS = 2.7 days
- ICD10 code: +
- Discharge diagnosis: tachycardia, atrial fibrillation, acute kidney injury (JMO)

**Ryde**
- 45y female presents to ED at Ryde by ambulance for a seizure
- 73 pathology tests
- 2 hsTroponin tests performed 7 hours apart
- 1st hsTroponin performed within 28 minutes of admission
- 174 pathology tests performed, 4 hsTroponins
- 88 forms filled eg (temperature, BP, pain scale, etc)
- Transferred from Ryde to RNS for chest pain
- 1 ECG
- 148 pathology tests
- 204 forms
- LOS = 5.2 days
- ICD10 = +
- Discharge diagnosis: malnutrition, brachy-tachy syndrome, pacemaker

**Overview of data in eMR systems**

- Data extracted from the eMR is structured in transactional data format. Each transaction/entry that occurs in the hospital is recorded as one line item.
- To reconstruct a patient’s journey through hospital, tables need to be linked by: 
  - Encounter ID – unique identifier for a patient’s interaction with the hospital eg. emergency, inpatient, outpatient, community) AND/OR
  - Person ID – unique patient identifier
- For every event of interest, we are examining the most appropriate time-stamp to describe an event.
Identifying all presentations of suspected acute coronary syndrome

- 3 months Test Data Extract: April-June 2017
- Bulk Data Extract: January 2013- June 2018
- Historical and future encounters are extracted

Cardiac keywords and symptoms
- Chest pain, chest tightness, shortness of breath, dyspnoea, weakness, nausea, vomiting, palpitations, syncope, presyncope, unwell, cardiac arrest, indigestion, sweaty, diaphoresis, dizziness, light-headedness, fatigue, clamminess, pale, ashen, loss of consciousness, SAlAMI, ETAlM, STEMI, NSTEMI, out of hospital cardiac arrest, ventricular tachycardia, ventricular fibrillation, failed thrombolysis, cath, cath lab, coronary bypass graft, ami, stent, angiogram, angi, epigastric pain, arm heaviness, chest heaviness

Included abbreviations, misspellings and additional keywords

UpSet plot showing the numbers of encounters meeting individual (left hand side) and multiple inclusion criteria (right hand side)

Diversity of episodes of care

Data in the EMR is captured as encounters. To enable patient-centric analyses, this data was converted into episodes of care. This is especially important for transferred patients where data is captured as ≥2 encounters.

Single Episodes of Care

Adjacent Episodes of Care

Overlapping Episodes of Care

Each row depicts a single patient. The pink shading represents the new journeyID that’s been created. Blue boxes represent encounters. • is a proxy for cath lab (admission to or procedure performed in cath lab)
Validation study of ICD10 coded STEMI

- **Rationale:** ICD10 codes can be used to identify STEMI but are not entirely reliable and are only available after the episode of care.

- Designed and built a user interface where cardiologists can easily sight all relevant aspects of a patient record (one at a time) and select a diagnosis. Data includes:
  - ECGs
  - First medical note
  - Blood tests (including hsTroponin)
  - Angiogram report
  - Discharge letter

- **Population for Validation Study:**
  - The starting population is 1144 episodes of care from admitted patients in NSCCLHD with hsTroponin changes.
  - Of these we will select 750 unique episodes of care for validation which will include cases with and without ICD10=STEMI.

  * a. $\Delta 30\%$ between initial and subsequent hsTroponin measurements
  * b. at least one hsTroponin measured during the encounter is $>99^{th}$ percentile for normal reference population OR
  * If hsTroponin $>1000\text{ng/L}$

**Outcome**

Labelled dataset that can be used to train algorithm(s) to identify "real" STEMI.

---

**Clinical Care Standards**

**Acute Coronary Syndromes Clinical Care Standard**

1. A patient presenting with acute chest pain or other symptoms suggestive of an acute coronary syndrome should be cared for by a documented acute care assessment pathway.
2. A patient with acute chest pain or other symptoms suggestive of an acute coronary syndrome should receive a 12-lead electrocardiogram (ECG) and the results are analysed by a clinician experienced in interpreting an ECG within 10 minutes of the first emergency contact.
3. A patient with acute ST-segment-elevation myocardial infarction (STEMI), for whom emergency percutaneous coronary intervention (PCI) or fibrinolysis is appropriate, is offered timely percutaneous coronary intervention (PCI) or fibrinolysis in accordance with the time frames recommended in the current National Heart Foundation of Australia/Cardiac Society of Australia and New Zealand Guidelines for the Management of Acute Coronary Syndromes.
4. A patient with a non-ST-segment-elevation acute coronary syndrome (NSTEMI) who is assessed to be at intermediate or high risk of an adverse event.
5. The role of coronary angiography, with a view to timely and appropriate coronary revascularisation, is discussed with a patient with a non-ST-segment-elevation acute coronary syndrome (NSTEMI) who is assessed to be at intermediate or high risk of an adverse event.
6. Before a patient with an acute coronary syndrome leaves the hospital, they are involved in the development of an individualised care plan. This plan identifies the lifestyle modifications and medicines needed to manage their risk factors, addresses any medication needs and includes a referral to an appropriate cardiac rehabilitation service where appropriate.

**Indicators that are potentially measurable using eMR data**

**Quality statement 2 – Early Assessment**

- **Indicator 2b:** ECG performed within 10 minutes of arrival of ambulance
- **Indicator 2c:** ECG performed and interpreted within 10 minutes of arrival to ED

**Quality statement 3 – Timely Reperfusion**

- **Indicator 3a:** STEMI patients receiving fibrinolysis or PCI
- **Indicator 3b:** STEMI patients receiving fibrinolysis within 30 minutes of hospital arrival
- **Indicator 3c:** PCI patients with STEMI with door-to-device within 90 minutes

**Quality statement 6 – Individualised Care Plan**

- **Indicator 6b:** Patients discharged on aspirin or dual antiplatelet therapy
**Indicator 2c:** Proportions of patients with ECGs received within 10 minutes

Self-presentations  
Ambulance arrivals

**Indicator 6b:** Median door to balloon time

PCI-capable facility  
Transferred to a PCI-capable facility

---

**Indicator 6b: Patients with ACS being discharged on aspirin or dual antiplatelet therapy**

**Step 1: Identify Medications**

Extract 12666 unique medication names (~350K medication results) in SPEED-EXTRACT

**Step 2: Group Medications**

Match them to the medication list and groups used in SNAPSHOT (~100/350K exactly matches)

**Step 3: Review and update medication list and groups**

Manual review of unmatched medications and partial text matching (~124/350K medication results)

Example from a discharge letter

Proportion of ICD10=STEMIs discharged on aspirin and dual antiplatelets

---

The University of Sydney
What symptoms do patients with ACS present with?

**ED Triage Assessment Interface**

**Aim**
1) Identify patients presenting with acute chest pain or other symptoms associated with ACS
2) Classify encounters into those with high, intermediate, low or no likelihood of cardiac-related chest pain

**Source**
43000 ED Triage forms from 1st April - 30th June 2017

**Features of cardiac-related chest pain**

i. **Nature** (crushing, heavy, weight, pressure)

ii. **Location of the pain and radiation** (central, typically left sided, up to the jaw, retrosternal, epigastric, radiating to the throat)

iii. **Associated features** (e.g. sweating, nausea, shortness of breath)

iv. **Exacerbating and relieving factors** (e.g. chest pain that is worse with respiration is less likely to be cardiac-related)

v. **Timing** (prioritising current symptoms, also taking into account resolving symptoms)

**Step 1: Text Mining – Identifying Symptoms And Keywords**

**Fuzzy Text Matching**

- Counts minimum number of insertions, deletions and substitution it takes to turn one word to another
- Deletion: Chest Pain → Chest Pain – requires 1 deletion (r)
- Insertion: Ches Pain → Chest Pain – requires 1 insertion (t)
- Substitution: Chast Pain → Chest Pain – requires 1 substitution (a -> e)
- Allow up to 2 of any deletions, insertions, substitution

- 32 different ways "chest pain" has been spelt
Step 2: Text Mining Contextual Analysis Examples

## Symptom/Keyword Grouping and Risk Category

<table>
<thead>
<tr>
<th>Symptom/Keyword Group</th>
<th>Risk Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>high_positional_chest_pain</td>
<td>High</td>
</tr>
<tr>
<td>chest_tightness/heaviness</td>
<td>High</td>
</tr>
<tr>
<td>check_coronary_artery_symptom</td>
<td>High</td>
</tr>
<tr>
<td>high_description_chest_pain</td>
<td>High</td>
</tr>
<tr>
<td>check_squeezing_constricted_banding</td>
<td>High</td>
</tr>
<tr>
<td>check_ecg_changes</td>
<td>High</td>
</tr>
<tr>
<td>check_eteve</td>
<td>High</td>
</tr>
<tr>
<td>check_l_axilla_pain</td>
<td>High</td>
</tr>
<tr>
<td>check_retoros_heaviness</td>
<td>High</td>
</tr>
<tr>
<td>intermediate_description_chest_pain</td>
<td>Intermediate</td>
</tr>
<tr>
<td>check_clapidigrel</td>
<td>Intermediate</td>
</tr>
<tr>
<td>check_pain_on_exertion</td>
<td>Intermediate</td>
</tr>
<tr>
<td>shortness_of_breath</td>
<td>Low</td>
</tr>
<tr>
<td>dizzy</td>
<td>Low</td>
</tr>
<tr>
<td>vomiting</td>
<td>Low</td>
</tr>
<tr>
<td>nausea</td>
<td>Low</td>
</tr>
<tr>
<td>pale</td>
<td>Low</td>
</tr>
<tr>
<td>palpitation</td>
<td>Low</td>
</tr>
<tr>
<td>no chest pain</td>
<td>No</td>
</tr>
</tbody>
</table>

- Developed text mining algorithm in the entire cohort of patients with suspected acute coronary syndrome
- 36 different symptom/keyword searches which accommodates for abbreviations and misspellings
- The encounter is assigned the highest risk category associated with the symptom/keyword group identified in the text

Our text mining tool called “CLACK” can take any input text source and examine strings of interest eg. risk factors, evidence of thrombolysis, past medical history, etc
Proof-of-concept project to automate extraction of clinical data from eMR to enhance discovery of new biomarkers of coronary artery disease

**Aim:** To test the feasibility of using eMR data from SPEED-EXTRACT to populate BioHEART, an ongoing clinical trial to identify new biomarkers and new mechanisms of coronary artery disease and myocardial infarction in patients presenting with STEMI.

**Mapping data variables between clinical trials and eMR data sources**

**Example of risk factors**

<table>
<thead>
<tr>
<th>BioHEART Clinical Trial</th>
<th>SPEED-EXTRACT eMR</th>
</tr>
</thead>
</table>
| Hypertension?           | 1. Prior medical history of hypertension recorded in ED notes (any occurrence of htn, hypertension, excluding negations)  
                          | 2. New record of hypertension recorded in progress notes (as above)  
                          | 3. Existing medications for hypertension (1077 types)  
                          | 4. Mean of 10 highest BP measurements over the episode of care |
| Current Smoker?         | 1. Record of ex- or current smoker recorded in structured field  
                          | 2. Record of ex- or current smoker in ED notes |
| Previous Smoker?        | 1. Record of ex- or current smoker recorded in structured field  
                          | 2. Record of ex- or current smoker in ED notes |
| Pack Years (20/day)     | Record of “pack years” recorded in ED notes |
| Dyslipidaemia?          | 1. Prior medical history of hyperlipidaemia recorded in ED notes (any occurrence of hcl, hypercholesterolaemia, excluding negations)  
                          | 2. New record of hyperlipidaemia in progress notes (as above)  
                          | 3. Prior prescription for statin or lipid-lowering medications (518 types)  
                          | 4. abnormal cholesterol test results (pathology) |
| Diabetes?               | 1. A HbA1c test result greater than 6.5 percent  
                          | 2. Prior medical history of diabetes recorded in ED notes (any occurrence of diabetes, diabetics, T2DM, IDDM, NIDDM, excluding negations)  
                          | 3. New occurrence of diabetes recorded in progress notes (as above)  
                          | 4. Any record of medications for insulin or antidiabetics (83 types) |

*presenting symptoms, family history, medical history, complications*
The ABC of embedded trials (Prof Neena Modi)

Accurate data
- Clear oversight and transparency across the data pipeline from data extraction from information systems, processing and analysis
- Ensuring data quality by following the data journey from ED through to medical coding

Benefits and buy-in
- Fortnightly meetings with stakeholders spanning Ministry of Health, eHealth, cardiologists, professors in nursing and digital health, software engineers, data analysts and data scientists

Confidence and trust
- Embedded in the health system closest to the data source
- Communication and iteration of results to check face validity

Work in progress

Stage 3:
NSW Health Facilities

Stage 2:
Expansion across Sydney Health Partner facilities (WSLHD, SLHD)

Stage 1:
NSLHD & CCLHD

PoC (eHealth/NSLHD)
Rapid Data Ingestion
Goal:
Demonstrate the feasibility and value of a rapid extraction and ingestion cloud platform

Ongoing work
- Data pipeline
- Assessment and management of diabetes in patients with acute MI
- Health care utilisation in patients presenting with low risk of cardiac related chest pain
- Risk factors in patients with STEMI
- BNP levels in heart failure...
It takes a herd!!