Building business intelligence into health workforce research: a graduate outcome tracking system

Gary Walker and Colleen Cheek

12th April 2018
Ad hoc project data – issues and risk

• Manual data entry and matching prone to error
• Different sources of similar data – what’s the ‘truth’?
• Data on user-level systems inaccessible
• Process is not explicit – reproducible?
• Time taken to re-run same queries for periodic reporting
• Reliance on individuals
• Lack of data security and governance
Strategic Information Systems Planning

- Explicit process
- Capture tacit knowledge
- Automation – data integrity, time, reproducible
- Enterprise level system - protects organisational knowledge, sustainable, secure
- Governance – ethics, privacy and access
- Opportunities – collaboration, innovation, research, decision support,
GradTrack Establishment

- Vision & scope –
  - Initially medical students
  - But with ability to be extensible to allied health and nursing
  - Secure
  - Sustainable
  - Governance

- Key success criteria.
  - Quality data for:
    - Research
    - Reporting
    - Decision support
GradTrack – Information Gathering

• Understand what’s available, what’s already been done
• Tacit knowledge
  • Person dependant roles, head knowledge
  • Legacy macros - bespoke systems
  • Manage risk/fear (we are not replacing their job)
• Go beyond the initial points of contact, they don’t know everything!
• Admin staff driven, but not resourced
• Alumni tracking off the side of desk
• Only time to get bare minimum data
• Manual and sporadic matching
  “It’s quite a process I’d have to say”
• Manual cleaning of data
• Linking RA to postcode
• Stored in Excel or Access Databases

“Data that is loved tends to survive”

Kurt Bollacker, Computer scientist
There is a better way
The GradTrack Data Warehouse

UTAS Student Data

Survey Data

External Data Sources

Clean and Combine (Data Matching)

Central Storage (Data Warehouse)

De-identification

Extracts and Reports

01100
010
001
GradTrack – Source data (raw product)

UTAS Student Data

Survey Data

External Data Sources
GradTrack Data Transformation (processing)

Clean and Combine (Data Matching)
GradTrack Data Warehouse (storage)
GradTrack – Extracts and Reports

De-identification

01100
010
001

Extracts and Reports
GradTrack – Data linkage
GradTrack – Data transformation

After
Example – Medical Students Outcome (MSoD) survey data

- 1 row per student, containing all survey questions from all surveys administered
- Received in CSV format
- Contained 717 columns (fields) of data
- MS Access can only handle 256 fields in a table.

RESULT: Imported into SQL table
Why not Excel/Access?

Example - AHPRA Full National Extract

- 320Mb, per year (Medical only)
- Provided in 12 separate XML files
- Nested data structure
AHPRA Full National Extract – XML map
<table>
<thead>
<tr>
<th>ProfessionNumber</th>
<th>General practice</th>
<th>Emergency medicine</th>
<th>Paediatrics and child health</th>
<th>Surgery</th>
<th>Addiction medicine</th>
<th>Anaesthesia</th>
<th>Physician</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDxxxxxxxx1</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx2</td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx3</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDxxxxxxxx24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
GradTrack - Other features

- Granularity of data
  - Addresses geocoded to SA1 level
  - Accurate MMM and RA coding
  - Mapping of graduate work locations
- QA - provided clean data back to source
- AHPRA match rate: from 68% to 95%
- Used existing IT infrastructure
  - No hardware/software costs
  - Centrally located
  - Centralised security policies
  - Backup, replication and DRP in place
  - The IT people know it exists
Results

- Data from multiple sources (937 fields)
- Automated collection, cleaning and linking
- Data protection
  - Governance, 5 safes
  - Encryption of outputs
  - Ethical approval, consents
- Contribute to research
- Faster reporting
- Foundational work done
- Scale to All health disciplines
- Analytics and decision support
What we’ve learned

• GradTrack uses standard Data Warehouse principles
• Compelling need
• Strong sponsor
• Organisational readiness for a DW
  • With the right mix of skills
• Dedicated personnel
• Defined processes quality data
• Still need to frame the right questions

“The idea is to go from numbers to information to understanding”

Hans Rosling
Opportunities and challenges

Challenge
• Student consent to tracking means we may never get 100% participation
  • onus on us to champion and demonstrate value and integrity

Opportunity
• Collaboration with other disciplines
• Nursing and Allied Health Data
• eg investigate use of statistical techniques to contrast the usefulness of different modelling approaches