Building an Organization to Accelerate Data Utilization

Hideki Ninomiya
Datack Inc
Profile

Career
- Tokyo University (Doctor of Medicine)
- Kansai Medical University
  - neurosurgeon
- Medley, Inc.
  - Medical Information
  - Telemedicine
- 3idea, Inc.
  - Data Science
- Datack, Inc.

Skill
- SQL, R, Python
- AWS
- epidemiology, data engineering, machine learning, natural language processing
Service

Real World Data Analysis

- Epidemiology, Data Engineering
- Claims data, Registry, EHR
- Database study

Real World Evidence

Building a clinical database in collaboration with hospitals

Goal: Application to Approval

- Area
  - Rheumatoid Arthritis
Today’s topic

- Why do we use RWD?
- Challenges in utilizing RWD
- Four Keys to Organization Building
- Practices
Why do we use RWD?
Why do we use RWD?

First of all, we need to clarify the purpose of RWD utilization.

Advanced real-world-evidence analytics can play an important role across the pharma value chain.

R&D
- Identify unmet need:
  - Inform research decisions
- Innovate in trial design:
  - Use synthetic control arms
- Improve trial design:
  - Define inclusion/exclusion criteria and end-points
  - Optimize site selection
  - Accelerate recruitment
- Accelerate time to market
- Refine formulations:
  - Determine optimal dosing based on patient response
- Monitor real-world outcomes:
  - Quantify unmet need
  - Understand safety and efficacy profiles

Market access
- Improve evidence of economic value:
  - Demonstrate economic value of treatment to payer
  - Compare trial data with real-world evidence (RWE) to strengthen dossier
  - Enable outcomes-based pricing
- Improve formulary position:
  - Achieve better patient access
  - Show efficacy and safety through head-to-head in silico trials
- Achieve label expansion:
  - Use RWE to eliminate need for new randomized clinical trial

Sales and marketing
- Improve targeting of commercial activities:
  - Target under-diagnosed patients
  - Identify "super responders"
  - Identify patients likely to switch or discontinue
  - Inform design of patient services/solutions
- Refine commercial strategy:
  - Shape product positioning
  - Understand healthcare-provider (HCP) decision making and impact on outcomes
  - Sharpen understanding of influence networks
- Build clinical-decision-support systems:
  - Provide recommendations at point of care based on predictions of outcomes, risk, or disease progression

Medical
- Improve pharmacovigilance:
  - Monitor real-world usage for safety and adverse events
  - Rapidly create granular view on benefits/risks
- Strengthen evidence for differentiation:
  - Analyze efficacy in understudied populations
  - Identify subpopulations for which effect outperforms trials
- Improve effectiveness of medical affairs:
  - Monitor unmet patient need at HCP level
- Improve adherence:
  - Support personal engagement to drive adherence and capture patient-reported outcomes with digital tools
Why do we use RWD?

**green:** start work → **red:** high impact, challenging

---

**R&D**
- Identify unmet need:
  - Inform research decisions
- Innovate in trial design:
  - Use synthetic control arms
- Improve trial design:
  - Define inclusion/exclusion criteria and end-points
  - Optimize site selection
  - Accelerate recruitment
- Accelerate time to market

**Market access**
- Improve evidence of economic value:
  - Demonstrate economic value of treatment to payer
  - Compare trial data with real-world evidence (RWE) to strengthen dossier
  - Enable outcomes-based pricing
- Improve formulary position:
  - Achieve better patient access
  - Show efficacy and safety through head-to-head in silico trials
- Achieve label expansion:
  - Use RWE to eliminate need for new randomized clinical trial

**Sales and marketing**
- Improve targeting of commercial activities:
  - Target underdiagnosed patients
  - Identify “super responders”
  - Identify patients likely to switch or discontinue
  - Inform design of patient services/solutions
- Refine commercial strategy:
  - Shape product positioning
  - Understand healthcare-provider (HCP) decision making and impact on outcomes
  - Sharpen understanding of influence networks
- Build clinical-decision-support systems:
  - Provide recommendations at point of care based on predictions of outcomes, risk, or disease progression

**Medical**
- Improve pharmacovigilance:
  - Monitor real-world usage for safety and adverse events
  - Rapidly create granular view on benefits/risks
- Strengthen evidence for differentiation:
  - Analyze efficacy in under-studied populations
  - Identify subpopulations for which effect outperforms trials
- Improve effectiveness of medical affairs:
  - Monitor unmet patient need at HCP level
- Improve adherence:
  - Support personal engagement to drive adherence and capture patient-reported outcomes with digital tools

---

*How pharma companies are applying advanced analytics to real-world evidence generation | McKinsey*

© 2020 DATACK, inc.
Challenges in utilizing RWD
Challenges in utilizing RWD

1. Insufficient human resources and organization
   ○ Today’s main topic

2. Quality and Quantity of RWD
   ○ Bias, Confounding
   ○ Lack of required clinical data
   ○ Not enough patients

3. Regulation
   ○ Approval
   ○ Pharmacovigilance
## Frequent bias, confounding

<table>
<thead>
<tr>
<th>Category/Subcategory</th>
<th>Percentage (%)</th>
<th>文献数</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confounding</td>
<td>63.2</td>
<td>74</td>
</tr>
<tr>
<td>Confounding by indication</td>
<td>32.5</td>
<td>38</td>
</tr>
<tr>
<td>Time-dependent confounding</td>
<td>6.0</td>
<td>7</td>
</tr>
<tr>
<td>Unmeasured/residual confounding</td>
<td>28.2</td>
<td>33</td>
</tr>
<tr>
<td>Healthy user/adherer effect</td>
<td>5.1</td>
<td>6</td>
</tr>
<tr>
<td><strong>Selection bias</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protopathic bias</td>
<td>3.4</td>
<td>4</td>
</tr>
<tr>
<td>Losses to follow-up (informative censoring)</td>
<td>2.6</td>
<td>3</td>
</tr>
<tr>
<td>Depletion of susceptibles (prevalent user bias)</td>
<td>10.3</td>
<td>12</td>
</tr>
<tr>
<td>Missing data</td>
<td>17.9</td>
<td>21</td>
</tr>
<tr>
<td><strong>Measurement bias</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Misclassification bias</td>
<td>43.6</td>
<td>51</td>
</tr>
<tr>
<td>Misclassification of exposure</td>
<td>23.9</td>
<td>28</td>
</tr>
<tr>
<td>Misclassification of outcome</td>
<td>28.2</td>
<td>33</td>
</tr>
<tr>
<td><strong>Time-related bias</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immortal time bias</td>
<td>25.6</td>
<td>30</td>
</tr>
<tr>
<td>Immeasurable time bias</td>
<td>3.4</td>
<td>4</td>
</tr>
<tr>
<td>Time-window bias</td>
<td>2.6</td>
<td>3</td>
</tr>
<tr>
<td>Time-lag bias</td>
<td>0.9</td>
<td>1</td>
</tr>
</tbody>
</table>

Solutions to the Challenges

1. Insufficient human resources and organization
   - Today’s main topic

2. Quality and Quantity of RWD
   - Bias, Confounding
   - Lack of required clinical data
   - Not enough patients

3. Regulation
   - Approval
   - Pharmacovigilance

   - Epidemiology
     - Understanding of data
   - Building new databases
   - Accumulation by various stakeholders

© 2020 DATACK, inc.
Four Keys to Organization Building
Four Keys to Organization Building

1. A person who makes the strategy
2. Epidemiologists (RWD experts)
3. Each department, each person, gains experience
4. Create a system for using RWD
Required skill of Epidemiologists

● Creating a research design
  ○ prospective research
  ○ database study
● Data handling
  ○ SQL, SAS, R…
  ○ Understanding of database and clinical practice
● Collaborating with clinical experts
  ○ Generating clinical questions
  ○ Making decisions
● Paper writing skills
Practice

- Case: Datack
  - organization
  - workflow
- Extraction Request Form
- Code Set
- Analysis environment
Extraction Request Form

- ○ ○ days

index date

patient definition

end of study period
or
the eligibility period
or
death

follow-up period for the occurrence of a new event

inclusion criteria

exclusion criteria
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>25mg 10×25</td>
<td>2149039F1244</td>
<td>2149039F1244</td>
<td>622144801</td>
<td>4987058892039</td>
<td>04987058143971</td>
<td>14987058892036</td>
<td>1214482010101</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25mg 14×25</td>
<td>2149039F1244</td>
<td>2149039F1244</td>
<td>622144801</td>
<td>4987058892077</td>
<td>04987058143988</td>
<td>14987058892074</td>
<td>1214482010201</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500 10×1</td>
<td>4987058892558</td>
<td>4987058892558</td>
<td>14987058892555</td>
<td>14987058892555</td>
<td>1214482010301</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>25mg 10×25</td>
<td>2149039F1201</td>
<td>2149039F1201</td>
<td>6221565301</td>
<td>4987190032300</td>
<td>04987190662507</td>
<td>14987190032307</td>
<td>1216530010101</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25mg 14×25</td>
<td>2149039F1201</td>
<td>2149039F1201</td>
<td>6221565301</td>
<td>4987190032324</td>
<td>04987190662521</td>
<td>14987190032321</td>
<td>1216530010201</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500 10×1</td>
<td>4987190032348</td>
<td>4987190032348</td>
<td>14987190032345</td>
<td>14987190032345</td>
<td>1216530010301</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>25mg 10×25</td>
<td>2149039F1287</td>
<td>2149039F1287</td>
<td>622177001</td>
<td>4987614408605</td>
<td>04987614408667</td>
<td>14987614408602</td>
<td>1217704010101</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25mg 14×25</td>
<td>2149039F1287</td>
<td>2149039F1287</td>
<td>622177001</td>
<td>4987614408643</td>
<td>04987614408674</td>
<td>14987614408640</td>
<td>1217704010201</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500 10×1</td>
<td>4987614408636</td>
<td>4987614408636</td>
<td>14987614408633</td>
<td>14987614408633</td>
<td>1217704010301</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Code Set for Disease

### DMARDs

**Biological DMARDs (boDMARDS+bsDMARDS)**
- TNF
- インフリキシマブ (ATCコード: L04AB02)
- エタネルセプト (L04AB01)
- アダリムマブ (L04AB04)
- ゴリムマブ (L04AB06)
- セルトリズマブベゴル (L04AB05)
- IL6
- トシリズマブ (L04AC07)
- サリルマブ (L04AC14)
- T cell
- アパタセプト (L04AA24)

**Synthetic DMARDs**

**Targeted (tsDMARDS) = JAK阻害剤**
- トファシチニブ (L04AA29)
- オルミエント (L04AA37)

**Conventional (csDMARDS)**
- メトトレキサート (L04AX03)
- レフルノミド (L04AA13)
- ブシラミン (M01CC02)
- ペキシラミン (M01CC01)
- サラゾスルファビリジン (A07EC01)
- ミゾリビン (ATCコードなし)
- タクロリムス (L04AD02)
- 金チオリン酸ナトリウム (M01CB01)
- オーラノフィン (M01CB03)
- メタリプターゼ (M01CC01)
- アクタリット (ATCコードなし)
- イグナチモド (ATCコードなし)
<Disease: ICD-10 code>
- MI: I21
- Angina pectoris: I20
- Stroke: I60-I64
  - Ischemic stroke: I63-I64
- Atrial fibrillation: I48
- Heart failure: I50

<CABG or PCI (K code)>
- K546 紅皮的冠動脈形成術
- K547 紅皮的冠動脈粥薬切除術
- K548 紅皮的冠動脈形成術（特殊カテーテルによるもの）
- K549 紅皮的冠動脈ステント留置術
- K550 冠動脈内血栓溶解療法
- K550-2 紅皮的冠動脈血栓吸引術
- K551 冠動脈形成術（血栓内膜摘除）
- K552 冠動脈、大動脈バイパス移植術
- K552-2 冠動脈、大動脈バイパス移植術（人工心肺を使用しないもの）
Analysis environment requirements

- Reproducibility of analysis
- Easy to review code
- Centralized data management and pre-processing
- Version control
- Storage, Memory, Process speed
- Security

Reproducibility & Replicability

<table>
<thead>
<tr>
<th></th>
<th>dataset</th>
<th>study design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reproducibility</td>
<td>Same</td>
<td>Same</td>
</tr>
<tr>
<td>Replicability</td>
<td>Different</td>
<td>Same</td>
</tr>
<tr>
<td></td>
<td>Same</td>
<td>Different</td>
</tr>
<tr>
<td></td>
<td>Different</td>
<td>Different</td>
</tr>
</tbody>
</table>
## Analysis Environment

<table>
<thead>
<tr>
<th>Component</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>data storage</td>
<td>Amazon Simple Storage Service (Amazon S3)</td>
</tr>
<tr>
<td>database</td>
<td>Amazon Athena, Amazon Redshift</td>
</tr>
<tr>
<td>analysis server</td>
<td>Amazon Elastic Compute Cloud (Amazon EC2)</td>
</tr>
<tr>
<td>language</td>
<td>MySQL, R, Python</td>
</tr>
<tr>
<td>version control system</td>
<td>Bitbucket</td>
</tr>
<tr>
<td>execution environment</td>
<td>Jupyter</td>
</tr>
</tbody>
</table>
Test of results

Based on clinicians and previous studies, it is important to determine indicators to check the validity of the tabulation and analysis.

- Percentage of patients using bDMARDs
  - Clinician's prediction: 20%
Easy to review

- Check the work of others
  - data
  - script
  - code
  - period

```
In [41]:
import yaml
import pandas as pd
from dataack.path import user_home_path
from dataack.sql import athena_quire_base

def load_athena_config():
yaml_file = user_home_path("dataack/config/athena_mirai.yml")
with open(yaml_file) as f:
    config = yaml.load(f.read(), Loader=yaml.SafeLoader)("mirai")
return config

def athena_quire(sql_query, dtype=None):
    """Execute sql_query to select in Athena.
    Args:
        sql_query(str): SQL you want to execute.
    Return:
        pd.DataFrame: Result of sql_query
    Examples:
        >>> df = athena_quire("SELECT * FROM sample_db LIMIT 1000")
        >>>
        config = load_athena_config()
        return athena_quire_base(sql_query=sql_query, config=config, dtype=dtype)
```

```
In [9]:
athena_quire("show tables")
```

Out[9]:
```
+---------+-----------------+-----------------+----------+
<table>
<thead>
<tr>
<th>m_drug</th>
<th>EngineExecutionTime</th>
<th>DataScannedInBytes</th>
<th>Data:</th>
</tr>
</thead>
<tbody>
<tr>
<td>m_patient</td>
<td>0.365 seconds</td>
<td>0.0 MB</td>
<td></td>
</tr>
<tr>
<td>t_disease</td>
<td>0.0 seconds</td>
<td>0.0 MB</td>
<td></td>
</tr>
<tr>
<td>t_hisdrug</td>
<td>0.0 seconds</td>
<td>0.0 MB</td>
<td></td>
</tr>
<tr>
<td>t_hisxamlab</td>
<td>0.0 seconds</td>
<td>0.0 MB</td>
<td></td>
</tr>
<tr>
<td>t_hisradio</td>
<td>0.0 seconds</td>
<td>0.0 MB</td>
<td></td>
</tr>
<tr>
<td>tope</td>
<td>0.0 seconds</td>
<td>0.0 MB</td>
<td></td>
</tr>
<tr>
<td>t_ranote</td>
<td>0.0 seconds</td>
<td>0.0 MB</td>
<td></td>
</tr>
<tr>
<td>t_rapanel</td>
<td>0.0 seconds</td>
<td>0.0 MB</td>
<td></td>
</tr>
</tbody>
</table>
```

Is the data correct?
Easy to review

Is there a disease other than RA?

```python
In [12]:
dis.disease.value_counts()
```

<table>
<thead>
<tr>
<th>Disease</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>RA</td>
<td>2327</td>
</tr>
<tr>
<td>PMR</td>
<td>37</td>
</tr>
<tr>
<td>SLE</td>
<td>9</td>
</tr>
<tr>
<td>SjS</td>
<td>8</td>
</tr>
<tr>
<td>SSc</td>
<td>5</td>
</tr>
<tr>
<td>PM</td>
<td>3</td>
</tr>
<tr>
<td>MCTD</td>
<td>3</td>
</tr>
<tr>
<td>Behcet</td>
<td>2</td>
</tr>
<tr>
<td>DMy</td>
<td>2</td>
</tr>
<tr>
<td>IgG4RD</td>
<td>1</td>
</tr>
<tr>
<td>潰瘍性大腸炎関連関節炎</td>
<td>1</td>
</tr>
</tbody>
</table>
Name: disease, dtype: int64

In [66]:

```python
# 2020年について、各患者で一番古い処方日を選ぶ
def first_odrymd_2020 = drug_2020.groupby('pid', as_index=False).odrymd.min()
def first_odrymd_2020.head(3)
```

<table>
<thead>
<tr>
<th>pid</th>
<th>odrymd</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2020-06-18</td>
</tr>
<tr>
<td>1</td>
<td>2020-01-22</td>
</tr>
<tr>
<td>2</td>
<td>2020-03-25</td>
</tr>
</tbody>
</table>

Is the period correct?

Is the condition correct?

```python
In [72]:

# 同一患者で2020年で一番古い日の処方データに、複数種類の薬剤が含まれている場合の取り扱いについて
# 00(Bio), 10(MTX), 20(DMARDs), 30(ステロイド), 40(NSAIDs), 99(その他)の順に優先する
# JAKは分類していない。
# 例えばBioとMTXが同時に処方されていた場合、その患者はBioとして集計する
# 下記は分類ごとの処方数
def first_drug_2020.groupby('pid').bunrui.min().value_counts()
```

<table>
<thead>
<tr>
<th>Value</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1248</td>
</tr>
<tr>
<td>20</td>
<td>535</td>
</tr>
<tr>
<td>30</td>
<td>325</td>
</tr>
<tr>
<td>0</td>
<td>314</td>
</tr>
<tr>
<td>1</td>
<td>65</td>
</tr>
<tr>
<td>99</td>
<td>29</td>
</tr>
</tbody>
</table>
Name: bunrui, dtype: int64
Today’s summary

● We need to clarify the purpose of RWD utilization.

● Challenges in utilizing RWD & Solution
  ○ Insufficient human resources and organization
  ○ Quality and Quantity of RWD
  ○ Regulation

● Four Keys to Organization Building

● Practices
  ○ Case: Datack
    ■ organization
    ■ workflow
  ○ Extraction Request Form
  ○ Code Set
  ○ Analysis environment
My beliefs

Data can lead society in the right direction.