European Medical Information Framework

Symposium
Friday 22nd September 2017

Instituto de Investigación
Hospital 12 de Octubre
Introductory Welcome

Bart Vannieuwenhuyse
Janssen Pharma R&D
Why is EMIF needed?
Secondary use of health data to enrich research

The “burning platform” for life sciences

Challenge

Today, Pharma doesn’t have ready access to this data, yet insights for safety, CER and other areas are within this clinical domain, which includes medical records, pharmacy, labs, claims, radiology etc.

The value of healthcare data for secondary uses in clinical research and development — Gary K. Mallow, Merck, HIMSS 2012
Project overview

14 European countries combining 57 partners
€56 million worth of resources
3 projects in one
5 year project

ACADEMIC PARTNERS

SME PARTNERS

EFPIA PARTNERS

PATIENT ORGANISATION

EMIF Introduction 22/09/2017
Our vision

To become the trusted European hub for health care data intelligence, enabling new insights into diseases and treatments.

Discover

Assess

Reuse
Project objectives

**EMIF-Platform**

Develop a framework for evaluating, enhancing and providing access to human health data across Europe, support EMIF-Metabolic and EMIF-AD (the specific topics below) as well as support research using human health data in general.

**EMIF-Metabolic**

Identify predictors of metabolic complications in obesity.

**EMIF-AD**

Identify predictors of Alzheimer’s Disease (AD) in the pre-clinical and prodromal phase.
Available data types

Large variety in “types” of data

- Primary care data sets
- Hospital data
- Administrative data
- Regional record-linkage systems
- Registries and cohorts (broad and disease specific)
- Biobanks
- Secondary care data sets
- Paediatric data sets

Data is available from more than 40 million subjects from six EU countries, and in addition:

- 25,000 subjects in AD cohorts
- more than 94,000 subjects in metabolic cohorts
Available data sources

EMIF-Available Data Sources; EXAMPLES

Status Jan 2016

Approximate total (cumulative) number of subjects

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Number of Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>THIN</td>
<td>12M</td>
</tr>
<tr>
<td>PHARMO</td>
<td>10M</td>
</tr>
<tr>
<td>SIDIAP</td>
<td>6M</td>
</tr>
<tr>
<td>ARS</td>
<td>3.6M</td>
</tr>
<tr>
<td>IPCI</td>
<td>2.8M</td>
</tr>
<tr>
<td>AUH</td>
<td>2.3M</td>
</tr>
<tr>
<td>HSD</td>
<td>1.6M</td>
</tr>
<tr>
<td>IMASIS</td>
<td>1M</td>
</tr>
<tr>
<td>SCTS</td>
<td>475K</td>
</tr>
<tr>
<td>PEDIANET</td>
<td>400K</td>
</tr>
<tr>
<td>EG CUT</td>
<td>52K</td>
</tr>
<tr>
<td>SDR</td>
<td>2K</td>
</tr>
<tr>
<td>MAAS</td>
<td>1K</td>
</tr>
</tbody>
</table>

Available data sources

Status Jan 2016

EMIF Introduction
Today at a glance --

• Results from EMIF work to date

• Projecting forward

• Power of data harmonization – OHDSI experience

• Bring it all together – panel discussion
More Information

• EMIF general
  • Bart Vannieuwenhuyse (bvannieuw@its.jnj.com)
  • Simon Lovestone (simon.lovestone@psych.ox.ac.uk)
  • Johan van der Lei (j.vanderlei@erasmusmc.nl)

• EMIF-Platform
  • Johan van der Lei (j.vanderlei@erasmusmc.nl)
  • Nigel Hughes (nhughes@its.jnj.com)

• EMIF-Metabolic
  • Ulf Smith (ulf.smith@medic.gu.se)
  • Dawn Waterworth (dawn.m.waterworth@gsk.com)

• EMIF-AD
  • Pieter Jelle Visser (pj.visser@maastrichtuniversity.nl)
  • Johannes Streffer (jstreffe@its.jnj.com)

EMIF is operating under IMI Grant Agreement nº115372

www.emif.eu
Register for our newsletter!
Research Use Cases – What Have We Learned?

Chair: Prof Simon Lovestone
Oxford University
EMIF Metabolic

Bart Vannieuwenhuyse
Janssen Pharma R&D
EMIF-Metabolic

Bart Vannieuwenhuyse
(on behalf of the EMIF-Metabolic team)

September 2017
Project objectives – EMIF-metabolic

**Project objectives: EMIF-metabolic**

**EMIF-Platform**

Develop a framework for evaluating, enhancing and providing access to human health data across Europe, support EMIF-Metabolic and EMIF-AD (the specific topics below) as well as support research using human health data in general.

**EMIF-Metabolic**

Identify predictors of metabolic complications in obesity

**EMIF-AD**

Identify predictors of Alzheimer’s Disease (AD) in the pre-clinical and prodromal phase
EMIF-Metabolic: objectives

Sample Sources
- EHR/Biobanks
- Large Prospective studies with Endpoints
- Medium Size Cohort Studies
- Small biomarker-rich Cohort or intervention Studies and cell/animal-based studies

WP8
Number of Samples

WP7

WP6

WP5
Number of Biomarkers

Biomarkers
- Clinical
- Genomic
- Proteomic
- Lipidomic
- Metabolomic
- Lipid turnover

Time
WHAT IS NAFLD?
(non-alcoholic fatty liver disease)

❖ Ectopic liver fat from excess consumption of calories arising when safer fat stores are over-filled

❖ NAFLD also risk factor for more severe liver complications
Excess calories (increased intake or reduced energy expenditure)

Subcutaneous stores overwhelmed (genes / sex, FHx, ethnicity, ageing)

FAT
‘Spill over’

Hepatic lipid accumulation
Muscle
Pancreatic beta cell

Perivascular fat ⇒ Endothelial dysfunction

Insulin resistance
Hyperglycaemia

Sattar and Gill (2014) BMC Medicine
NAFLD: risk factor for serious disease

1-2% may progress over 15-20 years.

12% may progress over 8 years [7].

Preiss & Sattar Clinical Science 2008
Some examples of big questions being asked

❖ What is the **prevalence** of documented non-alcoholic fatty liver disease (NAFLD) disease in clinical practice?

  – Does it vary by country?
  – Is it rising over time?

❖ Is NAFLD a (strong) **risk factor** for heart disease?
EMIF- Metabolic – use case

- Use of EHR data – answering the questions
- Findings, Learnings and limitations
Data outputs – 3 quick examples

❖ #1 Loomis et al – Risk of NAFLD by baseline BMI in major US / UK datasets –
  – Higher with rising BMI, in diabetes and potentially in men

❖ #2 prevalence of NAFLD in 4 major EU EHR-datasets
  – Much lower than expected, likely due to under-diagnosis, but prevalence rising

❖ #3 NAFLD is weak predictor of CVD unlikely to be clinically meaningful – goes against some major editorials papers
Hazard ratio NAFLD vs BMI / gender

#1 Loomis, Waterworth, Sattar (2016) JCEM
Hazard ratio NAFLD vs T2DM / BMI

#1 Loomis, Waterworth, Sattar (2016) JCEM
#2 – EMIF – NAFLD prevalence in 4 major EU EHR-datasets – (work near completion)

- Worthwhile question – **yes**
- Data available – **yes**
- Collaboration with all relevant parties – **yes**
Prevalence of NAFLD/NASH
NAFLD Prevalence by gender (on the 1st of Jan 2015)

Higher in men in most datasets and ages – men at higher risk for given BMI
#3 NAFLD & incident MI

- Overall associations modest
- Note findings broadly consistent from 4 major EHRs
- Not able to adjust for more risk factors
- Results important for clinical practice
- NAFLD much stronger risk factor for Diabetes than MI
Experiences gained

❖ Need to work out importance of question first –

❖ Can it be delivered from EHR?
  – Do we have right data / sufficient capture of confounders?
  – Do we have robust assessment of outcomes of interest?
  – Do we have sufficient power?

❖ Works best when data providers, statisticians, scientists / clinicians with relevant epi experience collaborate (need to do this better)
  – Ultra-careful to assure question can be answered with degree of robustness before time and effort expended
  – And, make sure to ask will the answer really take us further?
Limitations

❖ Often missing data of importance
❖ E.g BMI commonly missing or measured only on those with risk factors or disease – potential major biases
❖ Can be overcome but need to be aware
❖ Reverse causality so longer follow-ups help
❖ Coding and understanding or outcome measures can be difficult / vary by EHR
❖ Easy to make simple mistakes, come to potentially wrong / non-robust conclusions
Conclusions

❖ Takes time to come to grips with EHR derived data
❖ Many groups need to come together to make important leaps
❖ Requires time and experience of epi / and understand strengths and limitations of EHRs to make real gains
❖ Lots of richness but need /experience time to realise
EMIF Alzheimer’s

Pieter Jelle Visser
VU Medical Centre, Maastricht
Data sharing in clinical research: the EMIF-AD experience

Pieter Jelle Visser, MD, PhD
Maastricht University
VU University Medical center Amsterdam
The Netherlands
VISION:
- European hub for health care data intelligence, enabling new insights into diseases and treatments

THREE SUBPROJECTS:
- EMIF-Platform
- EMIF-Metabolic
- EMIF-AD

56 PARTNERS FROM 14 EUROPEAN COUNTRIES
Overall aim

– Improve treatment opportunities for predementia AD by:
  • Discovery diagnostic and prognostic markers
  • Increased understanding AD pathophysiology

Approach

– Use existing data
  • Build infrastructure for data access and datasharing
– Use extreme phenotypes as outcome
  • Amyloid positive vs amyloid negative
Alzheimer’s disease

- Most common cause of dementia
- Starts with amyloid aggregation in the brain (plaques)

Plaques in brain
Alzheimer’s disease

- Most common cause of dementia
- Starts with amyloid aggregation in the brain (plaques)

In-vivo amyloid measures

![Plaques in brain](image)

**PET scan**

- Normal
- Alzheimer

**Lumbar puncture**

- Amyloid decreased in CSF
Needs in Alzheimer’s disease

❖ Large clinical datasets for:
   – Studies on etiology
   – Studies on prevalence and course disease
   – Selection of subjects for other studies
   – Monitoring treatment effects

❖ Type of data
   – EHR
   – Research cohorts
     • Clinical based
     • Population-based
     • Speciality groups
Researcher incentives for data sharing

❖ Valid research question
  – Can not be addressed by own data

❖ Acknowledgement in publication

❖ Nice to have
  – Funding
  – Access to pooled data for other analysis
Technical needs

❖ Find data:
   1. EMIF Catalogue
   2. EMIF-AD Participant selection tool (PST)

❖ Harmonise data:
   3. EMIF data model

❖ Access and analyse research data:
   4. TranSMART data platform

❖ Access and analyse EHR data:
   5. Jerboa and Octopus
1. EMIF catalogue (emif-catalogue.eu)

Meta-data of research cohorts and EHR datasets
1. EMIF catalogue data entry

6.01. **Number of subjects with at least one assessment**

6.01.01. Number of subjects

- Normal
  - 266
- Subjective complaints
- MCI
  - 247
- Probable/possible AD (NINCDS-ADRDA)
  - 258
- AD- preclinical stage (IWG/NIA)

---

Summary

1. Database General Info... 60%
2. Key Publications (1/1) 100%
3. Data Access (47/48) 97%
4. Study Characteristics... 94%
5. Inclusion / Exclusion ... 66%
6. Number of subjects (... 42%
7. Clinical Information (... 100%
8. Dementia rating scal... 76%
9. Subjective Cognitive ...100%
10. Neuropsychiatric S... 91%
11. Quality of Life (10/11) 90%
12. Caregiver (1/1) 100%
13. Health Resource Ut...100%
14. Other scales (2/2) 100%
1. EMIF catalogue search

Boolean Query

Diagnosis: MCI

CSF: A-beta 1-42

PET: Amyloid

APOE genotype
1. EMIF catalogue search

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution name</th>
<th>Location</th>
<th>Last update</th>
<th>Select</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeARN</td>
<td>Leiden – Alzheimer Research Nederland</td>
<td>Netherlands</td>
<td>2017-01-26</td>
<td>□</td>
</tr>
<tr>
<td>EADC-PET</td>
<td>European Alzheimer Disease Consortium PET</td>
<td>Genoa, Provincia di Genova, Liguria, Italy</td>
<td>2017-01-26</td>
<td>□</td>
</tr>
<tr>
<td>IDIBAPS</td>
<td>Institut d'Investigacions Biomèdiques August Pi i Sunyer - Barcelona</td>
<td>Hospital Clinic, University Barcelona, Provincia de Barcelona, Catalunya, Spain</td>
<td>2017-01-26</td>
<td>□</td>
</tr>
<tr>
<td>EADC prodromal</td>
<td>European Alzheimer Disease Consortium prodromal</td>
<td>Maastricht University</td>
<td>2017-01-26</td>
<td>□</td>
</tr>
</tbody>
</table>
### 1. EMIF catalogue cohorts

<table>
<thead>
<tr>
<th>Type</th>
<th>Number of cohorts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical</td>
<td>31</td>
</tr>
<tr>
<td>Population</td>
<td>7</td>
</tr>
<tr>
<td>Trial</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>Number of subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal cognition</td>
<td>49.972</td>
</tr>
<tr>
<td>Subjective cognitive complaints</td>
<td>4.416</td>
</tr>
<tr>
<td>Mild cognitive impairment</td>
<td>10.843</td>
</tr>
<tr>
<td>Alzheimer's disease dementia</td>
<td>9.949</td>
</tr>
<tr>
<td>Other dementia</td>
<td>2.453</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>N=77.633</strong></td>
</tr>
</tbody>
</table>
2. Participant selection tool

MDS=minimal dataset

MDS

PST

MDS

MDS

MDS

MDS

Cohort

Cohort

Cohort

Cohort

Cohort

MDS=minimal dataset
2. Participant selection tool

### AD Cohort Explorer

**Total Subjects:** 477

### Participant Selection

- **All Subjects**
- **Summary Graph**
- **Summary Graph with attribute selection**
- **Attribute Breakdown**

### All Subjects

<table>
<thead>
<tr>
<th>Attribute</th>
<th>IM1</th>
<th>IM2</th>
<th>IM3</th>
<th>IM4</th>
<th>IM5</th>
<th>IM6</th>
<th>IM7</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at Enrollment in Study</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APOE Genotype</td>
<td>35</td>
<td>29</td>
<td>30</td>
<td>311</td>
<td>0</td>
<td>23</td>
<td>49</td>
<td>72</td>
</tr>
<tr>
<td>Consent to Recontact Patient</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>49</td>
<td></td>
<td>49</td>
</tr>
<tr>
<td>CSF Amyloid Beta 42 - Date Last Observation</td>
<td>0</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td>48</td>
<td>98</td>
</tr>
<tr>
<td>CSF Amyloid Beta 42 - Value Last Observation</td>
<td>0</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>49</td>
<td>101</td>
</tr>
<tr>
<td>Current Age</td>
<td>35</td>
<td>29</td>
<td>30</td>
<td>311</td>
<td>0</td>
<td>23</td>
<td>49</td>
<td>477</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>35</td>
<td>24</td>
<td>30</td>
<td>311</td>
<td>0</td>
<td>23</td>
<td>49</td>
<td>101</td>
</tr>
<tr>
<td>Diagnosis - Last Date Recorded</td>
<td>35</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>45</td>
<td></td>
<td>103</td>
</tr>
<tr>
<td>Diagnosis Availability</td>
<td>35</td>
<td>24</td>
<td>30</td>
<td>311</td>
<td>0</td>
<td>23</td>
<td>49</td>
<td>101</td>
</tr>
<tr>
<td>Episodic Memory Value (normalized)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>311</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>311</td>
</tr>
<tr>
<td>Last RAVLT value (normalized)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MMSE - Date Last Observation</td>
<td>35</td>
<td>29</td>
<td>30</td>
<td>0</td>
<td>22</td>
<td>44</td>
<td></td>
<td>160</td>
</tr>
<tr>
<td>MMSE - Value Last Observation</td>
<td>35</td>
<td>29</td>
<td>30</td>
<td>0</td>
<td>23</td>
<td>49</td>
<td></td>
<td>166</td>
</tr>
<tr>
<td>MMSE - Value Last Observation &lt;=+1 yr</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MMSE Observation Count</td>
<td>35</td>
<td>29</td>
<td>30</td>
<td>0</td>
<td>22</td>
<td>49</td>
<td></td>
<td>165</td>
</tr>
</tbody>
</table>

### Subject Selection

- **Summary view**
- **APOE Genotype**
  - [Select all](#)
  - [Clear all](#)
  - e2/e2 [24]
  - e2/e4 [104]
  - e3/e4 [121]
  - e4/e4 [997]
  - e4/e4 [112]

- **Current Age**
  - [Select all](#)
  - 60 - 80
  - [Clear all](#)
  - 60

© Copyright EMIF 2016
2. Participant selection tool

MDS=minimal dataset
2. Participant selection tool

MDS=minimal dataset
3. Data harmonisation

- Research cohorts
  - EMIF-AD common data model
  - CDISC compliant
  - Minimal dataset of 50 variables

- EHR
  - OMOP common data model
4. EMIF-AD tranSMART dataplatform
4. EMIF-AD tranSMART dataplatform
4. TranSMART cohorts

<table>
<thead>
<tr>
<th>14 cohorts</th>
<th>Total N=3423</th>
</tr>
</thead>
<tbody>
<tr>
<td>AddNeuroMed</td>
<td>786</td>
</tr>
<tr>
<td>Amsterdam</td>
<td>172</td>
</tr>
<tr>
<td>Antwerp</td>
<td>150</td>
</tr>
<tr>
<td>CITA</td>
<td>40</td>
</tr>
<tr>
<td>Descripa</td>
<td>881</td>
</tr>
<tr>
<td>EDAR</td>
<td>332</td>
</tr>
<tr>
<td>Genoa</td>
<td>157</td>
</tr>
<tr>
<td>Gothenburg</td>
<td>95</td>
</tr>
<tr>
<td>IDIBAPS</td>
<td>164</td>
</tr>
<tr>
<td>IMAP</td>
<td>144</td>
</tr>
<tr>
<td>Lausanne</td>
<td>40</td>
</tr>
<tr>
<td>Leuven</td>
<td>180</td>
</tr>
<tr>
<td>Pharmacog</td>
<td>147</td>
</tr>
<tr>
<td>Sant Pau</td>
<td>135</td>
</tr>
</tbody>
</table>
## 4. TranSMART minimal dataset

### Demographics
- Age
- Gender
- Years of education

### Clinical baseline information
- Diagnosis
- Functional impairment scale
- Depression scale
- Mini Mental State Examination
- Co-morbidities
- Medication use
- Date of baseline visit

### Baseline Neuropsychological raw scores and z-scores
- Memory test
- Language test
- Attention/Executive functioning test
- Visuoconstruction test
- Date of Neuropsychological examination

### Amyloid measure
- Amyloid measure assessed by CSF or PET
- Date of amyloid assessment
- Cut-off used to define abnormality

### MRI measure
- Measure of hippocampal volume or medial temporal atrophy
- Date of MRI assessment
- Cut-off used to define abnormality

### Clinical follow-up data
- Last diagnosis
- Date of last clinical visit
- MMSE at each follow-up
- Date of MMSE at each follow-up
- Neuropsychological test scores at each follow-up
- Date of Neuropsychological test scores at each follow-up
5. EHR data access
## 5. EMIF-associated EHR datasets

<table>
<thead>
<tr>
<th>Database name</th>
<th>Setting</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>THIN</td>
<td>General practitioner</td>
<td>12 million</td>
</tr>
<tr>
<td>IPCI</td>
<td>General practitioner</td>
<td>2.8 million</td>
</tr>
<tr>
<td>HSD</td>
<td>General practitioner</td>
<td>2.3 million</td>
</tr>
<tr>
<td>AUH</td>
<td>Hospital</td>
<td>2.3 million</td>
</tr>
<tr>
<td>IMASIS</td>
<td>Hospital</td>
<td>&gt; 1.5 million</td>
</tr>
<tr>
<td>GePaRD</td>
<td>Health insurance data</td>
<td>17 million</td>
</tr>
<tr>
<td>ARS</td>
<td>Health insurance data</td>
<td>5 million</td>
</tr>
<tr>
<td>PHARMO</td>
<td>Drug prescriptions</td>
<td>10 million</td>
</tr>
<tr>
<td>EGCUT</td>
<td>Biobank</td>
<td>52,000</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td><strong>52 million</strong></td>
</tr>
</tbody>
</table>
Examples of reuse data in AD

- Prevalence and incidence of dementia in EHR
- Prevalence of predementia AD in research cohorts
- Recruitment from existing cohorts
  - EMIF-AD biomarker discovery study
  - Preclin AD cohort
Incidence AD in EHR

Data from 6 European EHR datasets (n=25 million) with 138,000 dementia cases

Pereira et al Alz Dem 2017
Prevalence predementia AD

Data from 51 research cohorts (n=8000)  Jansen et al JAMA 2015
Example: Prevalence predementia AD

Jansen et al JAMA 2015
Example: Prevalence predementia AD

Jansen et al JAMA 2015
Example: Prevalence predementia AD

Prevalence of Alzheimer disease and amyloid positivity

Jansen et al JAMA 2015
Recruitment existing cohorts: Biomarker discovery study

❖ Aim: find novel diagnostic and prognostic markers for predementia AD using existing data and samples

❖ Steps
  – Identification of cohorts through EMIF catalogue
  – Set-up contracts
  – Data pooling in tranSMART, central sample storage

❖ Status
  – 1200 subjects enrolled
  – Analysis ongoing
Recruitment existing cohorts: PreclinAD and 90+ studies

Recruited from
- Netherlands Twin Registry, Manchester and Newcastle Aging study, hospital settings
Acknowledgements

- Co-PI's: Simon Lovestone, Johannes Streffer
- WP1: Stephanie Vos, Isabelle Bos, Karl Herholz, Stephan Carter, Rainer Hinz, Elles Konijnenberg, Anouk Den Braber, Jori Tomassen
- WP3: Alison Baird, Cristina Legido-Quigley, Richard Dobson, Stephen Newhouse, Sarah Westwood, Anna Myers, Stuart Snowden, Malcom Ward, Lars Bertram, Kristel Sleegers, Katie Lunnor, Henrik Zetterberg, Frederik Barkhof, Mara Ten Kate, Giovanni Frisoni, Alberto Redolfi,
- WP4: Piotr Lewczuk, Hilkka Soininen, Henrik Zetterberg, Kina Höglund, Maria Pikkarainen
- PI's data-cohorts: Philip Scheltens, Sebastiaan Engelborghs, Giovanni Frisoni, Rik Vandenberghe, José Luis Molinuevo, Anders Wallin, Alberto Lléo, Julius Popp, Pablo Martinez-Lage
More Information

• EMIF general
  • Bart Vannieuwenhuyse (bvannieu@its.jnj.com)
  • Simon Lovestone (simon.lovestone@psych.ox.ac.uk)
  • Johan van der Lei (j.vanderlei@erasmusmc.nl)

• EMIF-Platform
  • Johan van der Lei (j.vanderlei@erasmusmc.nl)
  • Nigel Hughes (nhughes@its.jnj.com)

• EMIF-Metabolic
  • Ulf Smith (ulf.smith@medic.gu.se)
  • Dawn Waterworth (dawn.m.waterworth@gsk.com)

• EMIF-AD
  • Pieter Jelle Visser (pj.visser@maastrichtuniversity.nl)
  • Johannes Streffer (jstreffe@its.jnj.com)

EMIF is operating under IMI Grant Agreement n°115372

www.emif.eu
Register for our newsletter!
EMIF Cross Topic

Peter Egger
RWE & Epidemiology, GSK
Research Use Cases – what have we learned?
The EMIF EHR-Platform Perspective

Joint i~HD/EMIF Meeting 21\textsuperscript{st}-22\textsuperscript{nd} September 2017 - Madrid

Peter Egger, Glen James, Myriam Alexander
Epidemiology, GlaxoSmithKline
Overview

- Real world evidence to support healthcare
- The need for real world evidence from Europe
- What EMIF can offer
- Examples of studies conducted so far
- Summary and conclusion
Real world evidence to support healthcare

1. Determine unmet need and the value of intervention
2. Assess impact of health policy and resource allocation
3. Guide clinical development of new molecules
4. Evaluate the real world effects of medications
Real world evidence from across Europe

- Choice of different data sources
- Diversity
  - Geography
  - Healthcare systems and disease management
  - Type of healthcare data
- Large numbers
  - To evaluate rare occurrences
- Need for integrated data
  - Comprehensive patient medical records
What EMIF can offer

Research collaborations based on a wide network of data sources within a **Common Environment**

- Standard formats and tools and consistent ways of working across the different data sources

  - Consistent quality of research
  - More efficient study execution
  - Greater familiarity with study results format
  - More reliable comparisons
A Common Environment for the federated data network

Catalogue

Data source Characteristics
• Size
• Information content

Key dashboards
• Patient demographics
• Key clinical data

Open to all and free

Data Query

• Simple – numbers of patients only
• Fast & low cost
• Pre-approved

Full Study

Study execution – common processes
• contracting
• protocol, rev & approve
• semantic harmonisation
• data extraction
• analysis environment

Standard Modules
• Incid/prev
• Patient profile
• Treatment patterns
• Resource utilisation

Bespoke Studies
• Disease natural history
• Drug effectiveness
• Drug safety
Roadmap for study execution
# Heart failure

<table>
<thead>
<tr>
<th>Concept</th>
<th>ICD10</th>
<th>ICD10CM</th>
<th>ICD9CM</th>
<th>ICPC2ENG</th>
<th>ICPC2P</th>
<th>MTHICD9</th>
<th>RCD2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart failure</td>
<td>Heart failure ISO</td>
<td>Heart failure ISO</td>
<td>Heart failure 428</td>
<td>Heart failure K77</td>
<td>Weakness, heart K20019</td>
<td>Cardiac failure NOS 428.9</td>
<td>Heart failure G584.1</td>
</tr>
<tr>
<td>Cor pulmonale</td>
<td>Pulmonary heart disease, unspecified 127.9</td>
<td>Cor pulmonale NOS 127.91</td>
<td>Pulmonary heart disease K82</td>
<td>Disease, heart, pulmonary K82002</td>
<td>Cor pulmonale NOS 416.9</td>
<td>Cor pulmonale K82003</td>
<td>Cor pulmonale NOS G584.2</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>Congestive heart failure ISO</td>
<td>Congestive heart failure ISO</td>
<td>Congestive heart failure, unspecified 428.0</td>
<td>Failure, congestive cardiac K77002</td>
<td>Congestive heart disease 428.0</td>
<td>Congestive heart disease G580</td>
<td>Congestive heart failure G580</td>
</tr>
<tr>
<td>Left-Sided Heart Failure</td>
<td>Left ventricular failure ISO</td>
<td>Left ventricular failure ISO</td>
<td>Left heart failure 428.1</td>
<td>Failure, ventricular left K77002</td>
<td>Left ventricular failure 428.1</td>
<td>Left ventricular failure G581.1</td>
<td>Left ventricular failure G581.1</td>
</tr>
<tr>
<td>Hypertensive heart and renal disease with renal failure</td>
<td>Hypertensive heart and renal disease with renal failure</td>
<td>Hypertensive heart and chronic kidney disease, unspecified, without heart failure</td>
<td>Hypertensive heart and chronic kidney disease, unspecified, without heart failure</td>
<td>Hypertensive heart and renal disease, unspecified, without heart failure</td>
<td>Hypertensive heart and renal disease, unspecified, without heart failure</td>
<td>Hypertensive heart and renal disease, unspecified, without heart failure</td>
<td>Hypertensive heart and renal disease, unspecified, without heart failure</td>
</tr>
<tr>
<td>Hypertensive heart and renal disease with both</td>
<td>Hypertensive heart and renal disease</td>
<td>Hypertensive heart and chronic kidney</td>
<td>Hypertensive heart and chronic kidney</td>
<td>Hypertensive heart and chronic kidney</td>
<td>Hypertensive heart and chronic kidney</td>
<td>Hypertensive heart and chronic kidney</td>
<td>Hypertensive heart and chronic kidney</td>
</tr>
</tbody>
</table>
Examples of pilot research projects
(UseCases)
## Selected pilot projects

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Title</th>
<th>Progress</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>BMI and the risk of cardiovascular disease and all-cause mortality in European electronic medical records databases.</td>
<td>Analysis ongoing</td>
</tr>
<tr>
<td>10</td>
<td>Association of non-alcoholic fatty liver disease with cardiovascular and liver morbidity in electronic health record databases.</td>
<td>Publication in draft</td>
</tr>
<tr>
<td>11</td>
<td>Dementia: vascular and metabolic risk factors</td>
<td>Publication in draft</td>
</tr>
<tr>
<td>13</td>
<td>Treatment pathway analysis: An evaluation of treatment patterns and drug utilisation amongst cases with incident dementia in Electronic Health Records databases available in the European Medical Information Framework</td>
<td>Analysis ongoing</td>
</tr>
<tr>
<td>14</td>
<td>A nested case-control study of prior history of non-alcoholic fatty liver disease in demented and cognitively impaired individuals matched to healthy controls in European health records data.</td>
<td>Governance Approval</td>
</tr>
<tr>
<td>15</td>
<td>Utilisation of healthcare data to identify sub-types of heart failure patients based on clinical and/or molecular phenotypes</td>
<td>Data Extraction</td>
</tr>
<tr>
<td>16</td>
<td>An exploratory phenome wide association study linking asthma &amp; liver disease single nucleotide polymorphisms and electronic health records from the Estonian Genome Centre at the University of Tartu Database</td>
<td>Governance Approval</td>
</tr>
</tbody>
</table>
## Key data sources

<table>
<thead>
<tr>
<th>Database name</th>
<th>Total number of subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUH - Denmark, hospital (Aarhus) &amp; prescriptions</td>
<td>2.3 million</td>
</tr>
<tr>
<td>THIN - UK, primary care</td>
<td>12 million</td>
</tr>
<tr>
<td>IPCI - Netherlands, primary care</td>
<td>2.8 million</td>
</tr>
<tr>
<td>HSD - Italy, primary care</td>
<td>2.3 million</td>
</tr>
<tr>
<td>IMASIS - Spain, Barcelona, hospital</td>
<td>&gt; 1.5 million</td>
</tr>
<tr>
<td>PEDIANET - Italy, pediatrics</td>
<td>0.4 million</td>
</tr>
<tr>
<td>PHARMO - Netherlands, linked databases</td>
<td>10 million</td>
</tr>
<tr>
<td>SIDIAP - Spain, Catalonia, primary care</td>
<td>6 million</td>
</tr>
<tr>
<td>ARS - Italy, Tuscany, hospital &amp; prescriptions</td>
<td>5 million</td>
</tr>
<tr>
<td>EGCUT - Estonia, total healthcare &amp; biobank</td>
<td>52,000</td>
</tr>
</tbody>
</table>
UC6: Dementia prevalence & incidence in a federation of European EHR databases: The EMIF resource.


Alzheimer's and Dementia (2017), 1-10

- 6 EHR databases analysed (ARS, AUH, IPCI, HSD, IMASIS, THIN)
- Identified 139,000 dementia cases from an overall total of 25 million persons from 2004 to 2012
- Results lower than in the published literature but similar secular trends and patterns over age
Incidence of Dementia

Annual Incidence Rate of Dementia

Rate per 100,000 person years

EHR Database

ARS  AUH  HSD  IMASIS  IPCI  THIN

70-74  75-79  80-84  85-89
Incidence of Dementia

Annual incidence of first dementia diagnosis by age, year and EHR
Prevalence of Dementia

Annual Period Prevalence (%) of Dementia

EHR Database

ARS  AUH  HSD  IMASIS  IPCI  THIN

Prevalence (%) 0  1  2  3  4  5  6  7  8  9  10

70-74  75-79  80-84  85-89
Prevalence of Dementia

One-year period prevalence of dementia by age, year and EHR
UC11: Levels of Blood Pressure, BMI and Total Serum Cholesterol Prior to Dementia Diagnosis


Poster presented at the AAIC in July in London

Background: Research cohorts have suggested changes in vascular risk factor levels prior to dementia onset – to be investigated in large-scale data sources.

❖ 5 EHR databases analysed (AUH, IPCI, HSD, SIDIAP, THIN).
❖ An overall total of 287,000 cases of incident dementia compared to 28,700,000 age- and gender-matched controls on previously measured BMI, blood pressure, and total cholesterol.
❖ BMI and SBP show clear declines prior to dementia diagnosis – although with different patterns. DBP and total cholesterol are less consistent.
BMI Decline Prior to Dementia Diagnosis

Mean differences in BMI (kg/m²) between cases and controls at different intervals prior to dementia diagnosis date.

THIN, SIDIAP, HSD, IPCI, AUH, Meta-analysis.
SBP Decline Prior to Dementia Diagnosis

Mean differences in systolic blood pressure (mmHg) between cases and controls at different intervals prior to dementia diagnosis date.

- THIN
- SIDIAP
- HSD
- IPCI
- Meta-analysis
Summary and conclusion

❖ The wealth of data is impressive and the EMIF Platform provides a real opportunity for novel research

❖ Comparing results across data sources provides useful new insights and the basis for further research

❖ Platform tools developed so far work well and system integration is in the process of being tested

❖ Research efficiencies not realised yet as projects are being conducted during Platform development

❖ Useful experiences: Identified specific areas for improvement in study execution

❖ Sustainability of the Platform and its tools is the next goal
New Opportunities for Scaling Up Big Data Research

Chair: Prof Dipak Kalra
Institute for Innovation through Health Data
EMIF Strategic Data Extension Project - Key learnings

Tine Lewi & Omer Saka
Janssen Pharma R&D & Deloitte
European Medical Information Framework

EMIF Strategic Data Extension Initiative Key learnings
Realising the Value from Health Data ~ Improving Care and Research
September 21-22, 2017
Improving access to health data… a key objective in EMIF Platform

Providing data access
- Scale
- Diversity
- Depth

Delivering a working solution
- Privacy enabled solution
- Data harmonisation
- Analytical methods

Conducting relevant research
- Disease insights
- Value analysis
- Pharmaco-epidemiology

EMIF-Platform
How to leverage EMIF methods and solutions potentially for other disease areas

Context

- EMIF demonstrates how to realize the value from health data in two key therapeutic areas: Alzheimer Disease and Metabolics.
- EMIF results are an illustration of how secondary re-use of human health data enables to address research questions which were previously very difficult to answer.

Challenges

- EMIF Platform – Sustainability Workpackage - initiated in 2016 a strategic data extension project for further applying EMIF Platform tools (Catalogue, Workflows management, Harmonization to OMOP Common datamodel) and governance framework and address new research questions.
- A project has been carried out in collaboration with Deloitte to identify potential therapeutic areas being best candidates for EMIF strategic data extension program.
- This initiative aimed at identifying therapeutic areas with unmet needs, high potential for future collaboration being driven by common interests among EMIF members, research communities and data custodians, and promising application domains.
A step-wise approach was developed for identifying relevant disease areas, integrating views from distinct stakeholders.

**Screening of therapeutic areas/diseases**
Screening disease area with rich pipelines, high unmet needs and areas of focus for research community and public sector.

1. Pharma R&D Pipeline
2. Disease Burden
3. Public Health priorities research funding

**Assessing data availability**
Identifying potential data sources to support secondary use of data per selected disease area by high-level screening and characterization of data sources.

**Identifying research needs and potential partners**
Gathering insights through 1:1 interviews with thought leaders on areas of research/application domains with great unmet needs and potential for collaboration.

**Qualitative insights processing**
Providing an evaluation framework to structure the insights and support internal alignment and assess potential opportunities.
Several aspects were gathered per high-level criteria and factored in as part of the short-listing process.

**Pharma perspective**

Assessing pipelines of Pharma – Oncology sales forecasts are the highest of all therapeutic areas, but Endocrine, metabolic is expected to rise at the fastest rate (high/therapeutic level analysis) by 2025.

**Burden of the disease**

Cardiovascular disorders and Oncology as expected are associated with highest mortality rates in countries in scope gathering high number of DALYs.

**Public Health priorities and research funding**

eHealth European projects were mapped according to Therapeutic Areas and Application domains.

**National programmes (EUS) and European Joint Actions in specific therapeutic areas.**

Source: Global health data exchange. (EMIF) Note: Quantities for the European regions. Indicators expressed as total per 100,000 population (DALYs) calculated per some region, WHO, TIR of the EU.
Short-listed diseases were further refined based on strategic priorities

A set of scenario were derived to assess disease ranking per relevant perspective

### Immunology & inflammation
- Osteoarthritis
- Multiple sclerosis
- Non-Alcoholic Steato-hepatitis
- Osteoporosis

### Endocrine, metabolic
- Breast cancer
- NSCLC
- Head & neck cancer
- Gastrointestinal Stromal Tumor (GIST)

### Oncology (solid tumors)
- CLL/Small Cell Lymphocytic Lymphoma (SLL) - NHL
- Mantle Cell Lymphoma

### Hemato-oncology

### Mental health
- Major Depressive Disorder (MDD)

(*): listed unranked
RWE data availability was assessed for the short-listed diseases

1) Step 1: Providing data source synopsis

- Data owners
- Volume of data: #Patients
- Investigator(s) contact details
- Depth of data fields: (epidemiology, clinical, treatment information,...)

2) Step 2: Mapping data fields, access conditions

- Additional data collected (synopsis)
- Year of establishment
- Type of outcomes
- Feasibility to retrieve condition(s) of access and previous collaboration with Pharma

RWD sources | Rheumatoid arthritis

Example for RA

A large number of databases identified contain treatment information as most of them are by essence biologics registers for RA
- A great majority of them have been involved in safety studies and effectiveness studies.
Interviews with disease experts shed light on challenges and research domains with high unmet needs

1:1 interviews conducted in three disease areas and relevant domain expertise…

- Experts with clinical expertise
  - Rheumatoid arthritis
  - Multiple sclerosis
  - Oncology

- Experts with domain expertise
  - Biologics registries
  - Population-registries
  - Biobanking

…brought qualitative insights at several levels

- Inspiring data-sharing initiatives
- Research question(s) / Application domains* with need of increased collaboration
- Past collaborative studies achievements
- Potential relevant stakeholders for partnerships
- Challenges related to collaboration and secondary use of data

(*): application domains: biomarker, Comparative effectiveness, Safety monitoring, Disease understanding, Population segmentation
Experts particularly highlighted the following points…

- There is **value of establishing catalogues/inventories** to report datasets characteristics, data-sharing and partnership rules.
- There is a **need to strengthen researchers and data custodians networking** and data suitability assessment.
- Absence of IT-infrastructure is not a limiting factor by all experts interviewed.

- **Quality management should be ensured**, with further effort warranted at several levels (i.e. patient examination, biological sample level, ab-test, imaging).
- FAIR principles (Findable, Accessible, Inter-operable, Re-usable) should incorporate a *Quality pillar*.

- Importance of promoting a framework that safeguards local data control.
- Risks of loosing data ownership and uncontrolled data dissemination were also highlighted as matter of concerns.
- Personal data protection generates uncertainty in the perspective of the rolling out of the new EU GDPR.
A need of further collaboration supported by enhanced data sharing/re-use in specific research domains

<table>
<thead>
<tr>
<th>Multiple sclerosis…</th>
<th>Rheumatoid arthritis…</th>
<th>Oncology…</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long-term safety studies</td>
<td>Long-term safety studies</td>
<td>Enhanced epidemiology surveillance</td>
</tr>
<tr>
<td>Predictive analytics and patient stratifications</td>
<td>Disease phenotyping Patient segmentation</td>
<td>Monitoring implementation of clinical guidelines</td>
</tr>
<tr>
<td>MS and pregnancy</td>
<td>Biosimilars</td>
<td>Monitoring use of cancer drugs in real life</td>
</tr>
</tbody>
</table>

- Safety issues likely one of the application domains with highest common need to collaborate.
- Need for further identifying of predictive factors of disease onset, progression and treatment response.
- Ongoing MultipleMS and BigMS data collaborative studies.

- Safety: even large registries likely willing to collaborate in to monitor (very) rare adverse events.
- Investigating predictors of disease progression and treatment response will imply data linkage with biobanks and enriched data sets
- In the field of BioS, understanding switching patterns and addressing safety issues.

- Key potential areas of focus beyond disease epidemiology.
- Data harmonization is needed for tumour characterisation, staging and summary treatment.
- Enhanced data linkage between biobanks and oncology population-based registries.
Challenges were also highlighted by experts, with some being reported for both diseases.

Inventory/Catalogue with detailed are still lacking of cancer registries with detailed synopsis is lacking

- Still lacking even in the field of oncology population-based registries.
- Data-linkage, conditions of access and partnerships shall be described in depth.
- A catalogue set up for RA (2006-07) but keeping the information up-to-date in the long run is challenging.

Varying degree of collaboration with private sector and several conditions for successful collaboration to be met

- Data governance represent a major hurdle beyond technical issues.
- Scientific award and co-authorship increasing scientific reputation of study participants and ability to receive research grants and sponsorships.
- Keeping control on the data to mitigate the risk of loss of funding if the governance of the data is not guaranteed.

All fields would benefit from enhanced standardization

- Enhanced standardization in the way data are collected (e.g. exposure being measured and reported).
- A critical preliminary step for the future is to build consensus on standard clinical assessment tools to increase consistency of data collection (exposure, patient outcomes).

A federated network of databases has been pointed out as a possible solution to address needs of further collaboration

- Data governance represent a major hurdle beyond technical issues.
- Willingness to collaborate will depend on the inability to ensure sustainable funding, fair governance framework with no loss of control on the data with mandatory delegation to a third party.
The insights from the research were ultimately structured into a heat-map.

<table>
<thead>
<tr>
<th>Sources of information</th>
<th>1. Insights from EFPIA members</th>
<th>2. Real world data available</th>
<th>3. Experts insights completed by survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEAT-MAP</td>
<td>RA</td>
<td>MS</td>
<td>Oncology</td>
</tr>
<tr>
<td>Alignment with EFPIA members</td>
<td>high</td>
<td>medium</td>
<td>high</td>
</tr>
<tr>
<td>Alignment on focused research questions</td>
<td>high</td>
<td>medium</td>
<td>low</td>
</tr>
<tr>
<td>#datasets identified</td>
<td>high</td>
<td>high</td>
<td>high</td>
</tr>
<tr>
<td>Experts interviews response rate</td>
<td>medium</td>
<td>high</td>
<td>medium</td>
</tr>
<tr>
<td>Need for an inventory in the field</td>
<td>high</td>
<td>high</td>
<td>high</td>
</tr>
<tr>
<td>Need for networking</td>
<td>high</td>
<td>medium</td>
<td>high</td>
</tr>
<tr>
<td>Need for an IT-infrastructure for data-sharing and re-use</td>
<td>medium</td>
<td>medium</td>
<td>medium</td>
</tr>
<tr>
<td>Existing collaborations with data-sharing</td>
<td>high</td>
<td>high</td>
<td>high</td>
</tr>
</tbody>
</table>

Distinct color-grading (indications due to number of responses in IBD and oncology (n=7 respondents each).
Several drivers for enhanced collaboration within the scope of EMIF have been identified.

**Drivers for success and opportunities**

**Rheumatoid arthritis…**

A **strong alignment within EFPIA members** with members likely willing to steer next steps.

Relevant data sources **(biologics registries)** were identified with precedent in **collaborative studies** implying data-sharing in the field of **safety and effectiveness**.

Existing **EULAR Taskforce on Biologics Registers** to enhance collaboration: share best practices, maximize quality, and provide infrastructure to enable methods development.

**No existing large-scale collaborative study** funded yet by EU (e.g. Horizon 2020) or likely planned in the near future.

**Oncology**

Strong **alignment within EFPIA members with leadership** committee foreseen within EMIF.

In absence of publicly available catalogue, deep diving into all cancer registries and cancer specific research data sources is a very resource-intensive task.

**First contact initiated with chairman of the European Network of Cancer Registries**, willing to build follow-on discussion with EMIF.

A new Health Information System planned to be launched in 2017 by ENCR as a new platform to disseminate relevant information to the research community and general public.
Biosimilars are an area of opportunity for research

Building a cross-country collaborative study of biosimilars in Rheumatoid arthritis and IBD:

Treatment patterns – Safety – Effectiveness

1. **High attractiveness** with likely **endorsement by multiple stakeholders** given implications of providing real-world evidence based assessment in the field of biosimilars to foster successful adoption of most appropriate treatments by mitigating patient acceptance and building prescribers confidence (in collaboration with payers)

2. **A cross-application domains** topic with feasibility to move towards **Predictive analytics**

3. **Feasibility to expand beyond RA to other rheumatic diseases** to broaden the reach of the research and maximize impact.
In this changing landscape EMIF and further programs have a significant role to play

1. Maturity of private, curated data through large-scale investments
   - HLI gains $220 million investment through investors including Illumina, Celgene and GE Ventures
   - PatientsLikeMe secured a $100+ million investment through partnership with iCarbonX
   - Merck, has joined the Oncology Research Information Exchange Network® (ORIEN)
   - Verily – Alphabet baseline, BGI/BC

2. Governments are investing in generating patient data sets
   - Qatar is establishing a genome map of the local population (6000 genomes sequenced by mid 2017)
   - China confirmed precision medicine to be part of its Five Year Plan for 2016-2020
   - 21st Century Cures – 1M pt. cohort+, IMI, UK 1000 genomes
   - Small geographies: The Human Project is studying the lives of 4,000 NYC households over the span of decades across domains

3. HIT companies looking to monetize data
   - Cerner Healthintert – collects data from disparate sources for pop health and precision medicine
   - Phillips using AWS to analyze and store 15 PB of patient data gathered from 390 million imaging studies, medical records, and patient inputs
   - Higi – kiosks in pharmacies that tracks trends and changes in body data available in partnership with pharmacies

Drivers
- Land grab and competitive move by private ventures and nations to achieve market ownership of channels in HCLS consuming key data sets and obtaining critical mass of data
- Costs lowering of high content testing creating
- Businesses with infrastructure today must seek subsidy in a world where capital/infrastructure costs are penalized vs. pure virtualized cloud operations
- Large healthcare spend entices data investments

Considerations
- Expect any software business with a cloud model to have a data approach. Assume it is available or coming.
- Disposition to private data aggregators and federal initiatives are ‘big bets’ – opportunity from being a scale buyer to drive agenda
- Platform, tools, and talent to make full use of data
Patient/Citizen Generated Health Data: The Next Real World Data Frontier

Alison Bourke
QuintilesIMS
Patient/Citizen Generated Health Data: The Next Real World Data Frontier

Alison Bourke
Scientific Director, RWI, QuintilesIMS
Madrid, 22 September 2017
Why now?

• Digital social communication opens new channels to/from patients/citizens

• Factors outside the formal clinical environment (eg social deprivation, exercise, diet) have a huge health impact and E/M-health facilitates access to more routine and granular data outside of clinical setting eg Fitbit

• Patient-centric view
Patient Generated Data

Much RWD is EMR, and some limitations of EMR alone:-

- May not be dispensed prescriptions so compliance unknown
- Limited information on OTC medications
- Limited data on non-routine care, lifestyles, diet
- Limited information on how patients feel
- No information on patients’ health/life priorities
- Little data on environment eg climate, pollution

Summary – Snapshot data from the Healthcare team view
Patient/Citizen Generated Data
(Also known as Patient Generated Health Data - PGHD)

Social media

Apps

Bio-sensors

IOT

Environmental sensors

Wearables

QuintilesIMS Confidential
Location of Wearables

- Wrist: 55%
- Chest: 23%
- Purse/pocket/shoe: 17%
- Arm: 8%
- Head: 7%
- Leg: 5%
- Clothing: 6%
- Ear: 5%
- Ankle: 3%
- Necklace: 3%
- Finger: 1%
Types of Patient Generated Data

• Biometrics outside of clinical setting

• Environmental factors
  Example: MyAirCoach - the use of home-monitoring and mHealth systems to predict deterioration in asthma control and the occurrence of asthma exacerbations: Honkoop et al. *BMJ* 2017

• Lifestyle - fitness, diet, sleep
  > 160 Fitbit ClinicalTrials.gov studies – including obesity, cancer, post surgery

• Adherence & compliance of treatment

• Qualitative data - QOL & values - PROMs
  Example: Cloudy with a Chance of Pain
• Volume
• Variety
• Velocity
Issues/Challenges of PGHD (1 of 2)

• Cost
• Sample Bias
• Patient Recruitment
• Patient Retention
• Data Access
  o Technology
  o Ownership
  o Consent models
• Confidentiality
• Identifiability
• Device Standardisation
Issues/Challenges of PGHD (2 of 2)

• Keeping up with fast development
• Data Standardisation
• Data Quality
  o Subjectivity
  o Completeness
  o Accuracy
• Cyber security
• Workflow
• Analytics, eg NLP, ML, data visualisation
• Liability of actionable insights.
Direct Patient data + EMR: Example 1 - AFLOAT

Atrial Fibrillation Longitudinal Outcomes Assessment Study

To assess the symptom burden of AF in newly diagnosed patients identified within one week of symptom recording.

516 case – control pairs were identified as soon as data was received from THIN practices.

⇒ 82% GP response rate and 50% patient response rate to questionnaires forwarded.

Assessment of the safety of LABAs in asthma in routine care by combining health-care databases and direct patient follow-up.

Text messages sent monthly to the patient.

Phone interviews with the patient every 4 months and if any severe asthma exacerbation is detected by text message.
An exploratory study of self-reported medication use in pregnant women and pregnancy outcomes with validation of self-reported data through electronic health records & national prescription data

- Recruitment via leaflets in pharmacies, pregnancy websites, advertising
- Internet v phone
- Self-reported medication use (including non prescription eg herbal, illicit) compared with data from electronic health records, national prescription data, and regional prescribing practices

The PROTECT project received support from the Innovative Medicines Initiative Joint Undertaking (www.imi.europa.eu)
Direct Digital Patient Involvement

Offers fantastic research and clinical care advantages:

• Improved recruitment/retention to research
• Supplement traditional RCT and EMR data
• Increase the frequency and accuracy of data capture
• Assist early detection and diagnosis
• Inform clinical pathways, drug usage and utilisation
• Support precision medicine – targeted therapy
• Improve & inform medicines adherence
• Prioritise care in line with patient view

New projects….?
Crossing the new RWD frontier

More than just new data.....
Improved research & clinical care

Data fuelled apps managing disease

New carer – patient dynamic
Any questions?

Thank you

alison.bourke@quintilesims.com
Ethical Considerations Within Federated Data Use

Prof Dipak Kalra
Institute for Innovation through Health Data
The trustworthy scaling of big data research

Dipak Kalra
Bart Vannieuwenhuyse, Janet Addison, Nige Hughes, Caroline Sage, Nathan Lea, Louis Schilders, Kathleen Fadden
If we are to scale up big health data research, across data sources, across countries

- Trust is needed to protect the interests of
  - Data subjects
  - Data sources
  - Research users
  - Society as a whole
Components enabling the trustworthy reuse of health data for research

- Bona fide (societally acceptable) purposes
- Bone fide research organisations
- Transparently defining the source data: FAIR principles
- Precisely specifying the intended research
- Complying with research ethics and consent
- Protecting the identity of data subjects
- Agreeing terms for recognition and reward
- Compliance and audit
- A social contract?
The EMIF Code of Practice (ECoP)

 développed in order to help ensure:

- that the EMIF Platform and Services are used in ways that comply with legislation and policies on data protection
- that EMIF upholds best practices in the protection of personal privacy and information governance
- that EMIF promotes best practices in the conduct of clinical research using health data, for the public good

We expect to contribute this into a wider European governance landscape for research using big health data.
The key characteristic of bona fide research is that its objective is to discover new knowledge intended for the public good and to be made publicly accessible (i.e. published). A bona fide research organisation is one that is appointed or accredited or funded to undertake bona fide research, and/or has made public its commitment to adhere to recognised research governance principles. It is not a requirement that such research is the primary business of that organisation, or that all of the research undertaken by that organisation is published. It is not a requirement that the organisation be publicly funded. New knowledge includes the corroboration of, or the challenge to, existing knowledge as well as completely new discoveries. Intermediary stages of the research life cycle might not be made publicly accessible.

EMIF research users seeking health data access will be verified to be members of bona fide research organisations who have legitimate purpose in conducting research queries on health data.
The EMIF Charter - principles shaping the ECoP

- The EMIF platform can only be used, for assessing the feasibility of a study and for conducting research, by bona fide research organisations and for the objective of discovering new knowledge intended for the public good and to be made publicly accessible (i.e. published)

- Data sources
  - will always have autonomy over which data are made accessible and for which types of research
  - will always determine ethical acceptability and scientific validity
  - must be transparent about their data

- Data users
  - must adhere to the ethical rules and privacy protection policies of each data source
  - may only use the data for the specific agreed research purposes
  - must acknowledge the sources of the data they have used, and EMIF
The FAIR Guiding Principles

- To be Findable
- To be Accessible
- To be Interoperable
- To be Reusable

 can be applied to access to health data, for research
EMIF Code of Practice

- Data source agreement
  - Provides trustworthy data access
  - Specifies pre-conditions for remote querying
  - Agrees governance and terms with EMIF

- Federated feasibility service
  - Permissions specified through term sheets
  - Provides dataset or query interface
  - Governs datasets and audits

- Feasibility query tools
  - Undertakes distributed feasibility studies
  - Specifies the purpose of feasibility studies

- Data user agreement
  - Meets bona fide criteria
  - Complies with governance and audit
Feasibility: data source obligations

- Provides trustworthy data access
  - Specifies the access to fine grained data items
  - Specifies the data subject population
  - Assures that data are appropriately de-identified
  - Provides data updates regularly
Feasibility: data source obligations

- Specifies pre-conditions for remote querying
  - Kinds of acceptable bona fide research organisation
  - Kinds of acceptable research uses (e.g. for protocol optimisation)
  - Subject area domains (e.g. therapeutic areas, research domains)
Term list for specifying kinds of research organisation

- Pharma company
- Medical device manufacturer
- ICT company
- Regulatory body
- Academic research organisation
- Payer
- Government department
- Patient associations and charities
Term list for specifying types of research study

- Observational/non-interventional
- Interventional
- Comparative effectiveness
- Health economic studies
- Market research
- Post-authorization Safety Studies
- Post-authorization Efficacy Studies
- Pharmacogivilance
Feasibility: data user obligations

- Meets bona fide criteria
- Conducts bona fide research
- Intends to use the data for bona fide feasibility purposes
These must match the pre-conditions of each data source on which the query may be executed.
Feasibility: data user obligations

- Complies with governance and audit
  - Limits query access to staff working on each feasibility study
  - Ensures query results are only used for the specified purposes
  - Maintains audit logs of queries and result set handling
  - Provides EMIF with independently verified audit trails
Governing EMIF Analysis Services

EMIF Code of Practice

Code of practice for data sources

co-defines and signs

Data sharing agreement

co-defines and signs

Code of practice for data users

- Research purposes and protocol
- Data specification and access
- Measures to protect privacy
- Data enrichment services
- Fees to be paid
- Handling of analysis results
- Demonstrating compliance
Putting the ECoP into a bigger context...

- There is a need to:
  - champion and govern a **trustworthy** health data driven ecosystem including EHRs and clinical research platforms
  - promote to society the importance of using health data for research, to increase the scale, efficiency and societal benefits from clinical research, to improve health and health care
  - engage with society on governance standards that can be jointly upheld by data providers and users, and which are deemed by all to be **trustworthy**
LUNCH & DEMONSTRATIONS
Data Harmonisation & Novel Data Reuse

Chair: Assistant Prof Peter Rijnbeek
Erasmus University Medical Center, Rotterdam
OMOP CDM & OHDSI

Assistant Prof Peter Rijnbeek
Erasmus University Medical Center, Rotterdam
EMIF and the Observational Health Data Sciences and Informatics (OHDSI) initiative

Realising the Value of Health Data ~ Improving Care and Research
September 22th, 2017 Spain
Background

Massive numbers of electronic health records (EHR) are currently being collected globally in observational databases, including structured data in the form of diagnoses, medications, laboratory test results, and unstructured data contained in clinical narratives. This opens unprecedented possibilities for research and ultimately patient care.

Challenges

Observational databases differ in both purpose and design. Each has different logical organizations and physical formats, and the terminologies used to describe the medicinal products and clinical conditions vary from source to source.

We need to standardize
Translation to a common data model and standard vocabularies

Any common data model aims to achieve both syntactic and semantic operability.

**syntactic operability:**
common underlying data structure
(standard grammar)

**semantic operability:**
common understanding required to interchange information
(standard vocabulary)
Observational Health Data Sciences and Informatics (OHDSI) has been established as a multi-stakeholder, interdisciplinary collaborative to create open-source solutions for large-scale analytics using the OMOP CDM. [http://ohdsi.org](http://ohdsi.org)

OHDSI has established an international network of researchers and observational health databases with a central coordinating center housed at Columbia University

OHDSI community in action

OHDSI Collaborators:
• >140 researchers in academia, industry, government, health systems
• >20 countries
• Multi-disciplinary expertise: epidemiology, statistics, medical informatics, computer science, machine learning, clinical sciences

Databases converted to OMOP CDM within OHDSI Community:
• >50 databases
• >660 million patients
A caricature of the patient journey
Each observational database is just an (incomplete) compilation of patient journeys.
Questions asked across the patient journey

Which treatment did patients choose after diagnosis?

Which patients chose which treatments?

How many patients experienced the outcome after treatment?

Does one treatment cause the outcome more than an alternative?

Does treatment cause outcome?

What is the probability I will develop the disease?

What is the probability I will experience the outcome?
Complementary evidence to inform the patient journey

Clinical characterization: What happened to them?

Observation

Patient-level prediction: What will happen to me?

Inference

Population-level effect estimation: What are the causal effects?

Causal inference
What is OHDSI’s strategy to deliver reliable evidence?

• **Methodological research**
  - Develop new approaches to observational data analysis
  - Evaluate the performance of new and existing methods
  - Establish empirically-based scientific best practices

• **Open-source analytics development**
  - Design tools for data transformation and standardization
  - Implement statistical methods for large-scale analytics
  - Build interactive visualization for evidence exploration

• **Clinical evidence generation**
  - Identify clinically-relevant questions that require real-world evidence
  - Execute research studies by applying scientific best practices through open-source tools across the OHDSI international data network
  - Promote open-science strategies for transparent study design and evidence dissemination
Collaboration EMIF and OHDSI

- EMIF has adopted the OMOP-CDM and is actively mapping European databases (see next talk);
- Is incorporating the OHDSI tools in the EMIF Platform;
- Is contributing to the tool development;
- Has supported the addition of security layer on top of the toolset;
- Has evaluated OHDSI tools in the EMIF community
ATLAS is a free, publicly available, web based, open source software tool for researchers to conduct scientific analyses on standardized observational data.

http://www.ohdsi.org/web/atlas (use Chrome)
Welcome to ATLAS.
ATLAS is an open source application developed as a part of OHDSI intended to provide a unified interface to patient level data and analytics.

Documentation
The ATLAS user guide can be found here.

Getting Started
- Define a New Cohort
- Search the Vocabulary

Search the different ontologies used to describe patient level data around the world

Release Notes

<table>
<thead>
<tr>
<th>ATLAS Version 2.0.0 Release Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>WebAPI Version 2.0.0 Release Notes</td>
</tr>
</tbody>
</table>

This latest release contains **30** feature enhancements and issue resolutions:

- Release 2.0.0
- Handle missing evidence
- Inconsistent concept name link/text background colors/color key info for classification concepts
- Vocabulary search results view "search results for" display field sometimes missing search text
- Concept set optimization hangs
- Initial startup of app hangs on splash with an error 'No component name specified'.
- Going to Concept Sets from Home results in two requests to WebAPI/conceptset
- Initialization failure does not display error on interface
- Configure Vocabulary does not switch source that is queried.
- Concept set comparison
- Bug in clicking explore tab
Platform Integration

Integration of Atlas in the EMIF Catalogue -> installation of OHDSI toolset on top of database on central EMIF server

[Diagram showing the integration process with EMIF, Atlas, Data, Vocabulary, Result sets, and Import/Export interfaces.]
Example: Large-Scale Patient-Level Prediction
Example:
OHDSI Network Study Treatment Pathways

George Hripcsak et al. Characterizing treatment pathways at scale using the OHDSI network PNAS, 2016 doi:10.1073/pnas.1510502113
Next Steps

- Continue integration in EMIF platform
- Test runs with feasibility approach
- Treatment Pathways Study in more databases in OHDSI including our EMIF databases with a focus on T2DM
- Workshop with all DCs on the use of the OHDSI tools
- Evaluation of the translation of the European databases to the OMOP-CDM (next talk)
EMIF & Data Custodians
Experience with OMP CDM mapping in Europe

Michel Van Speybroeck
Janssen Pharma Data Sciences
EMIF and Data Custodians experience with Observational Medical Outcomes Partnership (OMOP), Common Data Model (CDM) mapping in Europe

Michel Van Speybroeck - Janssen
September 22nd, 2017
The challenge of health data harmonisation

I.f.o research question

Prior to any research question

Point of Harmonisation

Physical

Keep data local

Bring data central

Level of Harmonisation

Structural (syntactic)

Meaning (semantic)

Model

Bespoke

HL7 based

OMOP

i2b2

Sentinel

i~HD Annual Conference 2017 | 17
6
# Data sources in scope

<table>
<thead>
<tr>
<th>Database</th>
<th>Country / Region</th>
<th>Population Size</th>
<th>Type</th>
<th>Mapping Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agenzia regionale di sanità della Toscana (ARS)</td>
<td>Italy / Tuscany</td>
<td>5.10^6</td>
<td>Administrativ e</td>
<td>Completed</td>
</tr>
<tr>
<td>Aarhus University Hospital Database</td>
<td>Denmark</td>
<td>2.310^6</td>
<td>Administrativ e</td>
<td>Completed</td>
</tr>
<tr>
<td>Health Search IMS Health LPD</td>
<td>Italy</td>
<td>1.610^6</td>
<td>Primary care</td>
<td>Completed</td>
</tr>
<tr>
<td>Integrated Primary Care Information (IPCI)</td>
<td>Netherlands</td>
<td>2.810^6</td>
<td>Primary care</td>
<td>Completed</td>
</tr>
<tr>
<td>Pedianet</td>
<td>Italy</td>
<td>0.410^6</td>
<td>Pediatric data</td>
<td>In Progress</td>
</tr>
<tr>
<td>Pharmo</td>
<td>Netherlands</td>
<td>8.410^6</td>
<td>Primary care</td>
<td>Completed for cohort</td>
</tr>
<tr>
<td>Information System of Parc de Salut Mar (IMASIS)</td>
<td>Spain</td>
<td>1.410^6</td>
<td>Hospital data</td>
<td>In Progress</td>
</tr>
<tr>
<td>The Information System for the Development of Research in Primary Care (SIDIAP)</td>
<td>Spain / Catalonia</td>
<td>6.410^6</td>
<td>Primary care</td>
<td>In Progress</td>
</tr>
<tr>
<td>The Health Informatics Network (THIN)</td>
<td>United Kingdom</td>
<td>1210^6</td>
<td>Primary care</td>
<td>Completed</td>
</tr>
<tr>
<td>Estonian Genome Center at the University of Tartu (EGCUT)</td>
<td>Estonia</td>
<td>5210^3</td>
<td>Biobank</td>
<td>Completed</td>
</tr>
</tbody>
</table>
The process that was followed

- Analyze Current Data Source
- Define Business Logic
- Develop ETL
- Test ETL
- Deploy ETL
- Acceptance Testing

Vocabulary Mapping

Evaluate Infrastructure Technology

White Rabbit
- Profiling of data
- Generating fake data sets

Rabbit in a Hat
- Specification

Usagi
- Vocabulary Mappings
Critical Success Factors

❖ Bringing the right expertise together:
  – Deep understanding of the source database
  – Understanding of the OMOP CDM structure and vocabularies
  – Technical expertise:
    • Database(s)
    • Extract - Transform - Load (ETL) development – programming language irrelevant
    • Tool installation (OHDSI tools are predominantly based on Java)

❖ Development of the vocabulary mappings is the most resource intensive activity.

❖ Focused effort – importance of project management and proper resource allocation

❖ Quick assessment of results
Effort of mapping to OMOP CDM

THROUGHPUT TIME:
Total Turnaround time:

EFFORT:
Preparation (profiling / spec) 5 – 10 days
ETL Development 20 – 50 days
Vocabulary Mapping 40 – 100 days
Tool installation 2 – 10 days
Data Quality and Harmonisation

- Data quality: the degree to which data represent physical reality for a person at a point in time
  - Data accuracy recording
  - Annotation (coding, description, method)
  - Time representation

![Data Quality Chart](chart.png)
Verifying Results - Achilles
### Example: Drug Level Mappings

<table>
<thead>
<tr>
<th>Database</th>
<th>Ingredient</th>
<th>Clinical Drug Comp</th>
<th>Clinical Drug Form</th>
<th>Quant Clinical Drug</th>
<th>Clinical Drug</th>
<th>Unmapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Source 1</td>
<td>5%</td>
<td>11%</td>
<td>12%</td>
<td></td>
<td></td>
<td>72%</td>
</tr>
<tr>
<td>Data Source 2</td>
<td>81%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19%</td>
</tr>
<tr>
<td>Data Source 3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Source 4</td>
<td>35%</td>
<td>4%</td>
<td>1%</td>
<td>56%</td>
<td></td>
<td>4%</td>
</tr>
<tr>
<td>Data Source 5</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Source 6</td>
<td>8%</td>
<td></td>
<td></td>
<td></td>
<td>2%</td>
<td>70%</td>
</tr>
<tr>
<td>Data Source 7</td>
<td>12%</td>
<td>7%</td>
<td>3%</td>
<td>65%</td>
<td></td>
<td>14%</td>
</tr>
<tr>
<td>Data Source 8</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- All relevant source records should be mapped
- Depending on the source between 80 and 100% of codes can be mapped
- Level of mapping might not correspond to the level of source data
Key take-aways

- Participants recognize the benefit of mapping to a common data model
  - Makes the knowledge of the source more explicit
  - Enables scalable research
  - More transparency in protocols
- But even with a CDM you need to have the direct interaction with data custodians to understand elements that are not captured in a data model.
- Mapping to an OMOP CDM is only the first step in a process
- Performing the mappings is a significant effort: dedicated resources, time-boxed and with the right expertise is critical
- A (more) formal process for evaluating the mapping results is required
Working with Cohorts: Switchboxes & Knowledge Objects

Rudi Verbeeck
Janssen Pharma IT
Deep semantic harmonization of clinical cohort data

Rudi Verbeeck
i~HD and EMIF joint event
22 September 2017 – Madrid, Spain
Research cohorts – the supply side

- Deep phenotyping based on research protocol
- Informed consent
- Cohort datasets look similar, but are not the same

Brøndby Haveby, Denmark
Researchers – the demand side

Data discovery & suitability: cohort selection
- Source metadata
- Aggregated, precomputed statistics & profiles

Feasibility: participant selection
- Aggregated results
- Combinations of variables

Feasibility: variable selection
- Aggregated results
- Descriptive stats

Analysis: PRRE
- Harmonized subject level data
- Data sharing agreements, ethical approval
- Limited availability

User

i~HD and EMIF joint event – September 2017 – Madrid, Spain
Data harmonization

Data custodians
- Identify local concepts
- Specify mappings
- Define security

Community
- Specify global and derived concepts
- Define research groups
Guiding principles

Generalization
- Compatibility = “inherent quality” + protocol
- Treat mappings, metadata and data equally
- Allow complex mappings

Efficiency
- Distribute ownership knowledge = responsibility
- Build library of reusable objects
- Technical ≠ semantic harmonization

Security
- Set local, propagate to global
- Fine grained
  - Use groupings
  - Use reasoner
- Allow traceback to source data

Implementation: semantic web
- Ontology describes application domain
- Specify minimum required information
- Use inferencing (rules) to populate with data

i~HD and EMIF joint event – September 2017 – Madrid, Spain
Dependency graph knowledge objects

LKO Year of birth
- GKO Age@BL
- GKO Avg Age@BL

LKO BL visit year
- GKO Sex
- GKO Count Male
- GKO Count Female

LKO Gender
- GKO Rey AVLT
- GKO Rey AVLT Z-score

LKO Rey AVLT
- GKO ApoE Allele 1
- GKO ApoE Allele 2

LKO ApoE Allele 1
- GKO ApoE Genotype
- GKO ApoE4 Genotype
- GKO ApoE4 Carrier

LKO ApoE Allele 2

LKO Height
- GKO Height

LKO Weight
- GKO Weight
- GKO BMI
Switchbox

EPAD

DPUK

ROADMAP

EPAD

Pharma Cog

DPUK

Cohort 1

Cohort 2

Cohort n

Global library

i~HD and EMIF joint event –September 2017 – Madrid, Spain
Switchbox

- Present uniform data interface to different projects
- Modelled on the OMOP CDM
- Switchbox contains predefined list (extendible) of harmonized variables
- Automatic connector from Switchbox to knowledge objects (global library)
  - Downstream knowledge objects come for free
  - No mappings from local knowledge objects
Architecture

Cohort Selection Tool (CST; Catalogue) → Participant Selection Tool (PST) → Variable Selection Tool (VST) → Private Remote Research Environment (PRRE)

Governance & Security/Integration Layer

Source data / SB
Semantic ETL
Harmonized triples
Data source 1

Source data / SB
Semantic ETL
Harmonized triples
Data source n
Catalogue - Communities

Welcome to the EMIF Catalogue
Below you can see existing communities

- **EMIF-EHR**
  - **Electronic Health Record data**
  - **Total:** 14
  - EMIF-EHR intends to explore how the massive data available in pan-European EHR systems can improve biomedical research.

- **EMIF-AD**
  - **Alzheimer's Disease cohorts**
  - **Total:** 45
  - The overall aim of EMIF-AD is to build an IF for studies on neurodegeneration in order to discover and validate AD biomarkers.

- **EPAD**
  - **Total:** 18
  - EPAD aims to develop an infrastructure to undertake adaptive studies for early and accurate decisions on the development of drug candidates or combinations.

- **ADVANCE**
  - **Total:** 2
  - Accelerated development of vaccine benefit-risk collaboration in Europe

- **DEMO**
  - **Total:** 3
  - This community is only used for Catalogue experiences

**Related publications**

- 2016/Jan
  - Tolerability and pharmacokinetics of oxaloacetate 100 mg capsules in Alzheimer’s subjects. has been published on BBA clinical 5 (120-3) [EMIF AD]

- 2015/Nov
  - The association between PGC-1a and Alzheimer’s disease. has been published on Anatomy & cell biology 45 (1-6) [EMIF AD]

- 2015/Oct-Dec
  - A New Look at Glaucoma, has been published on Journal of ophthalmic & vision research 10 (502-3) [EMIF AD]

- 2015/Apr/S
  - Evidence that electronic health records can promote physician counseling for healthy behaviors, has been published
## Catalogue – AD cohorts

The image shows a webpage from the EMIF Catalogue, which includes a table listing various AD cohorts. The table includes columns for Acronym, Name, Institution name, Location, Last update, and a checkbox for each entry. The cohorts listed are:

- **AddNeuroMed**
  - AddNeuroMed, Innovative Medicines for Europe (Innomed)
  - Institute of Psychiatry, King’s College London
  - City of London, Greater London, England, United Kingdom
  - 2016-04-04

- **ADGEN**
  - Kuopio-ADGEN
  - University of Eastern Finland
  - Kuopio, Kuopio, Pohjois-Savo, Finland
  - 2015-11-06

- **ADNI-1**
  - Alzheimer’s Disease Neuroimaging Initiative
  - University of California
  - San Francisco County, California, United States
  - 2015-11-06

- **ADNI 2**
  - Alzheimer’s Disease Neuroimaging Initiative
  - University of California
  - San Francisco County, California, United States
  - 2015-11-06

- **ADNI-GO**
  - Alzheimer’s Disease Neuroimaging Initiative
  - University of California
  - San Diego County, California, United States
  - 2015-11-05
Participant selection tool
Variable selection tool

## Variable Selection Tool

Total Subjects: 7580

<table>
<thead>
<tr>
<th>Name</th>
<th>IM1</th>
<th>IM2</th>
<th>IM3</th>
<th>IM4</th>
<th>IM5</th>
<th>IM6</th>
<th>IM7</th>
<th>IM8</th>
<th>IM9</th>
</tr>
</thead>
<tbody>
<tr>
<td>APOE Genotype</td>
<td>244</td>
<td>119</td>
<td>161</td>
<td>264</td>
<td>0</td>
<td>234</td>
<td>509</td>
<td>84</td>
<td>49</td>
</tr>
<tr>
<td>RAVLT</td>
<td>250</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MMSE</td>
<td>250</td>
<td>119</td>
<td>163</td>
<td>0</td>
<td>3507</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CERAD Word List Delayed Recall</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3530</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Amyloid Beta 42 in CSF</td>
<td>0</td>
<td>119</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Logical Memory Delayed - Norm</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>264</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Logical Memory Immediate - Norm</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>264</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gender</td>
<td>250</td>
<td>0</td>
<td>164</td>
<td>264</td>
<td>3530</td>
<td>234</td>
<td>509</td>
<td>84</td>
<td>49</td>
</tr>
</tbody>
</table>

**Selected:**
- MMSE
- Gender
- Amyloid Beta 42 in CSF

Only show selected variables
tranSMART cross-trial analysis
## Data in tranSMART

<table>
<thead>
<tr>
<th>Cohort name</th>
<th># Subjects</th>
<th># in cross trial</th>
<th># in 1000 samples cohort</th>
<th># expected in 1000 samples</th>
<th># unique variables</th>
<th># visits</th>
</tr>
</thead>
<tbody>
<tr>
<td>AddNeuroMed</td>
<td>774</td>
<td>46</td>
<td>14</td>
<td></td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Amsterdam</td>
<td>172</td>
<td>172</td>
<td>172</td>
<td>170</td>
<td>67</td>
<td>14</td>
</tr>
<tr>
<td>Antwerp</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>147</td>
<td>14</td>
</tr>
<tr>
<td>Cità – GAP</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>120</td>
<td>1</td>
</tr>
<tr>
<td>Descripa</td>
<td>881</td>
<td>881</td>
<td>29</td>
<td>228</td>
<td>215</td>
<td>12</td>
</tr>
<tr>
<td>EDAR</td>
<td>332</td>
<td>332</td>
<td>203</td>
<td>173</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Genoa</td>
<td>157</td>
<td>157</td>
<td>157</td>
<td>157</td>
<td>35</td>
<td>2</td>
</tr>
<tr>
<td>Gothenburg</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>101</td>
<td>127</td>
<td>7</td>
</tr>
<tr>
<td>IDIBAPS</td>
<td>164</td>
<td>164</td>
<td>120</td>
<td>120</td>
<td>92</td>
<td>11</td>
</tr>
<tr>
<td>IMAP</td>
<td>144</td>
<td>144</td>
<td>144</td>
<td>85</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Lausanne</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>112</td>
<td>8</td>
</tr>
<tr>
<td>Leuven</td>
<td>180</td>
<td>180</td>
<td>180</td>
<td>180</td>
<td>53</td>
<td>1</td>
</tr>
<tr>
<td>Pharmacog</td>
<td>147</td>
<td>147</td>
<td>147</td>
<td>147</td>
<td>59</td>
<td>5</td>
</tr>
<tr>
<td>Sant Pau</td>
<td>135</td>
<td>135</td>
<td>45</td>
<td>45</td>
<td>150</td>
<td>8</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3411</strong></td>
<td><strong>2637</strong></td>
<td><strong>1221</strong></td>
<td><strong>1221</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

EMIF key distinguishing features:

Data custodians - Supply
- Empower data owners and SMEs to distribute the workload for deep harmonization
  - Metadata
  - Mappings
- Specify minimum info for unambiguous interpretation of metadata & mappings, generate the data
- Build library to encourage re-use
- Control access

Researchers - Demand
- Progressive protocol refinement and drill down to the data
  - Data discovery
  - Suitability
  - Feasibility
  - Data analysis
- Tools
  - Cohort selection tool
  - Patient selection tool
  - Variable selection tool
  - PRRE
Acknowledgements

Data harmonization framework

- Janssen: Luiza Gabriel, Alvaro Cortes-Callabuig, Michel van Speybroeck
- U. Manchester: James Cunningham

Tools and scripts

- Know.Bi: Bart Maertens

Library

- ITTM: Serge Eifes, Adriano Barbosa, Kavita Rege

TranSMART data load

- Maastricht U.: Isabelle Bos, Stephanie Vos
- The Hyve: Janneke Schoots – van der Ploeg, Olaf Meuwese, Andy Sewgobind, Stefan Payralbe
Placebo as a Surrogate for RWD

Prof Derek Nunez, Gurparkash Singh & Peter Egger
Duke University, US,
Janssen Pharma R&D & RWE & Epidemiology GSK
Re-Use of Clinical Trial Data

A Case Study Sponsored by EMIF-Metabolic

Derek Nunez MD FRCP (presented by Peter Egger PhD, GSK) & Gurparkash Singh PhD, Janssen

Madrid September 2017
Disease insights from a variety of data sources

- Observational patient health data sources
  - Administrative databases for health insurance purposes
  - EHR data for patient management purposes
  - Disease / treatment registries
  - Biobanks

- Patient health data from Clinical Trials
  - Clinical Trials are conducted to evaluate the safety and efficacy of a new treatment
  - Can Clinical Trial data be re-used to evaluate disease?
Clinical Trial Data from Placebo Arms

Pros

– Trials can be very large (10,000 +) and long (3+ years)
– Placebo arms
  • No Investigational Drug(s) to complicate interpretation
  • Subjects often on ‘Standard of Care’ medications
– Subjects are observed periodically using standardized reporting tools (physical exams, laboratory measurements, ECGs etc)
– Longitudinal trends may be discernable
– May include novel data collections – digital data directly from patients, such as from wearables, real-time recording by patients
Clinical Trial Data from Placebo Arms

Challenges

❖ Providing ‘real world’ insights
  – Inclusion/Exclusion criteria may skew subjects away from “Real World” patients
  – More intense disease monitoring and management
  – Close observation of subjects may alter behaviour
  – Subjects may drop-out during a trial

❖ Access to data
  – Trial consent forms must allow the re-use of data
  – May be difficult to collaborate across companies
Case study

Background
❖ Nonalcoholic Fatty Liver Disease (NAFLD) is commonly associated with obesity and/or type 2 diabetes
❖ NAFLD is common (10-30% of adults), but progression to more severe liver disease is uncommon and predictors are not well understood

Key objectives
1. How well do BMI and liver endpoints track together?
2. NAFLD progression and baseline predictors

Can Clinical Trial data be re-used to address these objectives?
- Can use data from trials not designed to investigate specifically NAFLD objectives but where NAFLD measures such as Alanine aminotransferase (ALT) and aspartate aminotransferase (AST) are measured as ‘safety biomarkers’
The STABILITY trial (The STabilisation of Atherosclerotic plaque By Initiation of darapLadlb TherapY)

- Tested darapladib (LpPLA₂ inhibitor) vs Placebo
- 15,828 subjects enrolled (663 centers in 39 countries)
  - At high cardiovascular risk (chronic coronary artery disease or risk factors [one of age ≥60 years, diabetes, smoker, low HDL-C, polyvascular arterial disease, renal dysfunction])
- High background use of standard-of-care treatment (eg. statin therapy)
- Randomization to darapladib or placebo
- Median duration of follow up: 3.7 years

<table>
<thead>
<tr>
<th>Cardiovascular risk factors</th>
<th>2687 (34.0)</th>
<th>2664 (33.6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes requiring pharmacotherapy — no. (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-density lipoprotein cholesterol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (IQR) — mg/dl</td>
<td>44.4 (38.6–52.9)</td>
<td>44.8 (38.6–53.7)</td>
</tr>
<tr>
<td>&lt;40 mg/dl — no. (%)</td>
<td>2786 (35.2)</td>
<td>2646 (33.4)</td>
</tr>
<tr>
<td>Smoker — no. (%)</td>
<td>1656 (21.0)</td>
<td>1572 (19.8)</td>
</tr>
<tr>
<td>Renal dysfunction — no. (%)§</td>
<td>2374 (30.0)</td>
<td>2410 (30.4)</td>
</tr>
<tr>
<td>Polyvascular disease — no. (%)§</td>
<td>1187 (15.0)</td>
<td>1185 (15.0)</td>
</tr>
</tbody>
</table>
Number of subjects on Placebo (all subjects)

Focused on subjects with complete data over 36 months

![Graph showing the number of subjects on Placebo over 42 months.](image-url)
Subject characteristics

<table>
<thead>
<tr>
<th>Subject characteristics</th>
<th>STABILITY</th>
<th>DIA Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>36 months</td>
</tr>
<tr>
<td><strong>n=4264</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>64.2 (9.1)</td>
<td>67.2 (9.1)</td>
</tr>
<tr>
<td>BMI (kg/m^2), mean (SD)</td>
<td>28.8 (4.9)</td>
<td>28.9 (5.0)</td>
</tr>
<tr>
<td>Males, n (%)</td>
<td>3,525 (83)</td>
<td>3,525 (83)</td>
</tr>
<tr>
<td>T2D, n (%)</td>
<td>1,605 (38)</td>
<td>1,605 (38)</td>
</tr>
<tr>
<td>HbA_1c, %, mean (SD)</td>
<td>7.3 (1.4)</td>
<td>7.4 (1.5)</td>
</tr>
<tr>
<td>eGFR &lt;60 mL/ min/1.73m^2, n (%)</td>
<td>566 (13)</td>
<td>611 (14)</td>
</tr>
<tr>
<td>Current smoker, n (%)</td>
<td>1,257 (29)</td>
<td>630 (14.8)</td>
</tr>
</tbody>
</table>
GSK Results
STABILITY: ALT, AST & AST/ALT ratio

Baseline BMI Category

A  ALT

B  AST

C  AST/ALT ratio
STABILITY: Modeled data of the association of ALT, AST and AST/ALT ratio to visit BMI

Estimates adjusted for baseline age, gender, smoking, T2D status and eGFR category and incident cardiovascular events

mean ± 95% CI
STABILITY: Effect of Change in BMI (type 2 diabetes status)
STABILITY: Modeled data of the association of ALT, AST and AST/ALT ratio in the ‘BMI Gainer’, ‘BMI Loser’ and ‘Stable BMI’ tertiles

A  ALT

B  AST

C  AST/ ALT ratio

Estimates adjusted for baseline age, gender, smoking, T2D status and eGFR category and incident cardiovascular events

mean ± 95% CI
SUMMARY

❖ Clinical trials can be a rich source of longitudinal data for analysis of ‘Natural History’ of a disease or condition.

❖ Need to control for the impact of subject selection criteria and subject drop-outs (important when performing meta-analyses across trials).

❖ ‘Normalisation’ procedures may need to be implemented for laboratory endpoints to correct for variations in analytical procedures and reference ranges.
Janssen Results
## Janssen Placebo Data: 3 Completed Phase 3 Trials

<table>
<thead>
<tr>
<th>Trial</th>
<th>NCTID* (Janssen Identifier)</th>
<th>Phase 3 Clinical Trial Population</th>
<th>Duration</th>
<th>Eligibility Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NCT01106625 (DIA3002)</td>
<td>Subjects with T2DM on metformin and sulphonylurea</td>
<td>52 weeks</td>
<td>Age, years: 18-80, HbA1c, %: ≥7.0 - ≤ 10.5, FPG, mg/dL: &lt;270 (15 mmol/L), eGFR (mL/min/1.73m²): ≥ 55</td>
</tr>
<tr>
<td>2</td>
<td>NCT01064414 (DIA3004)</td>
<td>Subjects with T2DM with moderate renal impairment</td>
<td>52 weeks</td>
<td>≥ 25, HbA1c, %: ≥7.0 - ≤ 10.5, FPG, mg/dL: &lt;270 (15 mmol/L at Week-2), eGFR (mL/min/1.73m²): ≥30 - &lt;50</td>
</tr>
<tr>
<td>3</td>
<td>NCT01106651 (DIA3010)</td>
<td>Older Subjects with T2D</td>
<td>104 weeks</td>
<td>55-80, HbA1c, %: ≥7.0 - ≤ 10.0, FPG, mg/dL: &lt;270 (15 mmol/L at Week-2), eGFR (mL/min/1.73m²): ≥ 50</td>
</tr>
</tbody>
</table>
Number of Subjects on Placebo in 3 DIA Trials Combined

Focused on subjects with complete data over 12 months

- Baseline: 483
- Month 1: 472
- Month 3: 442
- Month 6: 366
- Month 12: 325
- Month 18: 164
- Month 24: 154
## Janssen: Subject Characteristics

<table>
<thead>
<tr>
<th></th>
<th>STABILITY</th>
<th>DIA Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>36 months</td>
</tr>
<tr>
<td><strong>n=4264</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>64.2 (9.1)</td>
<td>67.2 (9.1)</td>
</tr>
<tr>
<td>BMI (kg/m(^2)), mean (SD)</td>
<td>28.8 (4.9)</td>
<td>28.9 (5.0)</td>
</tr>
<tr>
<td>Males, n (%)</td>
<td>3,525 (83)</td>
<td>3,525 (83)</td>
</tr>
<tr>
<td>T2D, n (%)</td>
<td>1,605 (38)</td>
<td>1,605 (38)</td>
</tr>
<tr>
<td>HbA(_{1c}) %, mean (SD)</td>
<td>7.3 (1.4)</td>
<td>7.4 (1.5)</td>
</tr>
<tr>
<td>eGFR &lt;60 mL/ min/1.73m(^2), n (%)</td>
<td>566 (13)</td>
<td>611 (14)</td>
</tr>
<tr>
<td>Current smoker, n (%)</td>
<td>1,257 (29)</td>
<td>630 (14.8)</td>
</tr>
</tbody>
</table>
STABILITY versus DIA Trials : ALT
Baseline BMI Category

STABILITY

DIA Trials
STABILITY versus DIA Trials: AST

Baseline BMI Category

STABILITY  

DIA Trials

BMI = Body Mass Index
STABILITY versus DIA Trials: AST/ALT Ratio

Baseline BMI Category

STABILITY

DIA Trials

BMI = Body Mass Index
STABILITY versus DIA Trials : GGT
Baseline BMI Category

STABILITY

DIA Trials

BMI = Body Mass Index
DIA Trials: Modeled data of the association of ALT, AST and AST/ALT ratio to visit BMI

- Linear mixed models were also applied on 3 the DIA trials correcting for baseline age, gender, smoking and eGFR category but not for T2D status (since no non-T2D subjects in DIA Trials) nor for incident cardiovascular events.

- Results are less informative because of:
  - Limited combined longitudinal follow-up time (only 12 months)
  - Relatively limited number of subjects in different sub-groups
  - Wide 95% confidence intervals of estimates
SUMMARY

- Janssen DIA placebo trial data show similar pattern in ALT, GGT, AST and the AST/ALT ratio as noted in the GSK STABILITY trial placebo data.

- Precompetitive sharing of data and analyses is feasible.

- The use of liver biomarkers in this pilot provided insights that need to be confirmed by amalgamation of further datasets.
Other Initiatives of data sharing within Pharma

- Janssen has teamed up with the Yale University Open Data Access (YODA) Project for the responsible sharing of clinical research data to researchers.

- TransCelerate has set up the Placebo and Standard of Care (PSoC) Initiative to enable the sharing of de-identified data – from subjects either on placebo or the active ingredient.
Acknowledgements

**Janssen**: Geert Byttebier, Elisa Fabbrini, Gary Meininger, Barry Schwab, and Bart Vannieuwenhuyse

**GSK**: Myriam Alexander, Nick Galwey, Derek Nunez, Dawn Waterworth and Laura Yerges-Armstrong

**University of Glasgow**: Naveed Sattar

**University of Pisa School of Medicine**: Ele Ferrannini
COFFEE BREAK
Panel Discussion

Chair: Nigel Hughes
Janssen Pharma R&D
Closing Statement

Prof Simon Lovestone
Oxford University