EMIF Executive Summary

Project overview 2013-2015

Project rationale and overall objectives of the project

Advances in medical research require an increasing quantity and detail of human health data to answer today's complex and detailed research questions. Important volumes of human health data are available, either through routine clinical care or through research-driven cohort studies. Unfortunately, data sources are typically fragmented in a variety of (isolated) environments and systems and privacy, legal and ethical issues are not always properly covered. To help improve access to and use of patient-level data, the "European Medical Information Framework (EMIF)" was launched beginning of 2013.

The EMIF project aims to create an environment that allows for efficient re-use of existing health data (EMIF-Platform). To ensure immediate applicability, the EMIF project includes two specific research topics that will help guide the development of the Information Framework: the identification and validation of protective and precipitating factors for conversion to Alzheimer's Disease (EMIF-AD) and predictors of metabolic complications of obesity (EMIF-Metabolic).

EMIF-Platform's primary objective is to facilitate the re-use of healthcare data. Given the variety of data sources that may be useful for research, EMIF-Platform will enable identification, assessment and selection of suitable data sources (EMIF data catalogue). Because data come from multiple sources and may have different formats and content, the Platform emphasizes harmonisation of the data according to well established data format and content (semantic) standards to enable answering the research questions. Under a global philosophy of creating data platforms (federation), connecting software (data extraction software) and mechanisms for governance (legal, ethical and privacy) are also included in the work plan. More generally, Platform development, including data access, analysis and visualisation is addressed using an agile paradigm in which user requirements gathering and prototype evaluation are iteratively undertaken. In this approach, both EMIF-AD and EMIF-Metabolic serve as use scenarios and test bed for the Platform activities.

EMIF-AD's overall objective is to improve the design of treatment studies for Alzheimer's disease (AD) in non-demented subjects. To this end, it aims to discover and validate diagnostic markers, prognostic markers and risk factors for AD in non-demented subjects. EMIF-AD will make use of both existing and newly collected data. Key objectives of EMIF-AD are the development of a data platform in collaboration with EMIF-Platform and the use of extreme phenotypes. The data platform (or 'Medical Information Framework') will support data access, data storage, data pooling and analysis across cohorts. Extreme phenotypes will be applied as research methodology for the discovery of new markers. New candidate markers will be validated in cohorts derived from the large population datasets available through EMIF-Platform. Phenotypes can be defined by biomarkers (e.g. beta amyloid load or hippocampal atrophy), cognitive markers (e.g. rate of cognitive decline), or clinical presentation (e.g. resilience to dementia in the oldest-old).

EMIF-Metabolic aims to identify risk markers for metabolic complications of obesity. Obesity is a heterogeneous condition where many obese individuals do not show evidence of the metabolic complications of obesity and, conversely, many non-obese individuals show a dysmetabolic state. The Metabolic topic focuses on identifying biomarkers of risk and mechanisms related to this heterogeneity and to test the biomarkers in small and medium-sized cohorts followed by testing in large clinical populations with outcome data. Identifying useful biomarkers of obesity-related complications (Type 2 diabetes (T2D), cardiovascular disease, cancer and NAFLD/NASH) could facilitate more efficient and focused clinical trials and influence the risk-benefit balance for novel therapeutics by targeting treatments to those at high risk.

Besides the development of the platform and the research activities, possible sustainability models for EMIF are studied in order to ensure adequate post-project continuation of products and services developed.

Overall deliverables of the project

The **EMIF-Platform** aims to build an integrated, efficient framework (technical solutions and governance processes) for consistent re-use of currently available patient-level data to support novel research. EMIF-Platform will provide a means for researchers to 'browse' available data in Europe (EMIF data

This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (EMIF grant n° 115372).

catalogue). This full 'Medical Information Browser' will allow rapid exploration and identification of the wealth of information that at present remains largely 'hidden' in numerous isolated and scattered healthcare environments across Europe. The EMIF-Platform, by using modular functions, advanced search, analysis and visualization functionalities, and navigation interfaces, will allow the data to be available for browsing and re-using in multiple ways by various data researchers under secure governance respectful of data privacy and consent. EMIF-Platform will initially be able to, on its own, leverage data on more than 40 Million European adults and children by means of a network (federation) of healthcare databases and cohorts from 6 different countries (DK, IT, NL, UK, ES, EE), designed to be representative of the different types of existing data sources (primary care datasets, hospital-based databases, claims data, cohorts, national registries and biobanks).

EMIF-AD will provide an overview of the existing AD research cohorts (fingerprinting). This information will be integrated with the EMIF data catalogue and browser under development in EMIF-Platform. For a number of the available research cohorts (currently 6), a "Private Remote Research Environment" is being established. In collaboration with EMIF-Platform, a dedicated tranSMART instance is made available in a secure private cloud. In the course of the project, more European cohorts will be linked to the platform including datasets needed for specific EMIF-AD research questions (consisting of a limited number of variables from multiple cohorts) or complete cohorts. In this context, a "generic taxonomy" for AD research cohorts will be developed which is maximally aligned with the CDISC standards. Besides re-using existing cohort data, a new cohort of 300 cognitively normal subjects aged 60-80 with extensive information on AD biomarkers will be collected. Based on this, validated criteria for preclinical and prodromal AD are being developed and definitions of extreme phenotypes for AD based on cognitive decline, AD biomarkers and resilience to dementia are being identified. This should form the basis for the discovery and validation of novel biomarkers for AD that can be used for diagnosis and/or prognosis, using a wide range of modalities (proteomics, genomics, metabolomics, imaging) in non-demented subjects.

EMIF-Metabolic will identify novel biomarkers and mechanisms for obesity-associated complications using carefully characterized extreme phenotypes and omics technologies. The project will investigate the heterogeneity in the metabolic consequences of obesity in small, medium-sized and large clinical populations with outcome data. Based on this, a descriptive database of

This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (EMIF grant n° 115372).

patients will be generated with available risk factor phenotyping. This will help to evaluate the epidemiology of the conditions of interest. EMIF-Metabolic will deliver real world observational data and data from randomized controlled studies for classification of obesity and related diseases, this to test the identified biomarkers in a real-world setting as well as to test the feasibility of electronic healthcare record (EHR)-driven recruitment to intervention trials.

Summary of progress versus plan since last period

EMIF-Platform has progressed as planned during the third year. The EMIF Catalogue v3 has been released, now including community-based access roles and a new architecture based on the plugins concept. Cooperation with the other two EMIF Topics has been instrumental to the second evaluation round of the Catalogue, which has been updated in this version 3 according to the requirements issued from the use cases defined in year 1 (UCs 2 to 5). Furthermore, five new complex use cases (UCs 9 to 13) have been defined in collaboration with the EMIF-AD and EMIF-Metabolic Topics to answer research questions through specific data extraction exercises. These studies are in different stages of development, and encompass data extraction on endpoints of interest, including e.g. dementia, cardiovascular disease, BMI and NASH/NAFLD. The Jerboa Reloaded software has been improved for this purpose, and optimisation of workflows for data sources has resulted in a prototype workflow management tool to support these processes (TASKA). The OMOP common data model has been chosen to be evaluated in a number of EMIF-Platform data sources, with a view on the re-use of tools that were originally developed within Janssen and that are now made available via the Observational Health Data Sciences and Informatics (OHDSI) project. On the level of raw data handling (cohorts), work has continued on tranSMART, which now allows cross-cohort analysis and has been enriched with pipelines specific for –omics data analysis. In this domain, the methodological research on Knowledge Objects (semantic web) continued and is now gradually being used for data harmonization and preparation of data upload to tranSMART. Governance and data federation activities have concentrated on the development of the Ethical Code of Practice, which analyzes specific ethical and legal issues around the full data cycle in EMIF (discovery – assessment – re-use) and proposes a preliminary governance framework that would also be

This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (EMIF grant n° 115372).

applicable for the long-term sustainability phase. Business modelling studies have progressed and resulted in the first definition of the Business Plan, as well as a more detailed market analysis.

EMIF-AD has progressed as planned. Regarding infrastructure we have greatly improved functionality and content of both the EMIF catalogue and tranSMART instance. As regards novel data creation we included 124 cognitively normal subjects and performed deep-phenotyping. We conducted a large number analysis in existing datasets as part of WP2 leading to some major publications on prevalence and outcome of dementia, amyloid positivity, and prodromal AD. Moreover, we performed the first biomarker discovery studies in WP3 including RNA expression and imaging. We also collated samples for the 1000 cohort with biomarker analysis in this cohort starting in 2016. The three use cases developed in collaboration with the Platform have progressed significantly with regards to protocol definition and are well underway.

EMIF-Metabolic progressed as planned with only minor deviations. Cross-topic collaborations were established with joint work packages involving both the EMIF-Platform and the EMIF-AD Topics. The role of insulin resistance as a joint patho-physiological event in promoting the development of both AD and T2D has been addressed in the unique AD-Metabolic topic collaborative project, which is now complete. Extensive metabolomic and transcriptomic studies have been performed in carefully selected individuals with specific liver and diabetes-related phenotypes. Data have been incorporated and analysed in a systems biology approach, leading to the identification of potential therapeutic options in NAFLD and a pilot clinical study. Validation of biomarkers and additional discovery work in large cohorts is ongoing (*n*> 6000 individuals with outcome information). Potential causality of obesity and Alzheimer's as well as endometrial cancer has been analysed using a Mendelian Randomization approach and has now been published. A review of known biomarkers for T2D has been performed and published, which highlights that information on causality is largely absent. Initial epidemiological characterization of obesity-associated disorders (T2D and NAFLD/NASH) in large EFPIA EHR databases (THIN and Humedica) has been performed and a manuscript focused on NAFLD/NASH has now been published. We have continued to have close interactions with EMIF-Platform on two use Cases related to obesity; NAFLD/NASH as well as cardiovascular disease. Data extraction and analysis planning activities supporting the NAFLD/NASH question are now ongoing, with active collaboration from the data owners and Platform

This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (EMIF grant n° 115372).

colleagues. A major achievement has been the generation, approval and dissemination of a standardized protocol to examine the natural history of liver safety biomarkers in EFPIA clinical trial data and the covariates that affect them. This work is important for the interpretation of liver safety signals in clinical trials, as well increasing the understanding of liver biomarkers as endpoints for NAFLD.

Significant achievements since last report

EMIF-Platform

- Progress towards overall vision of Platform architecture: EMIF instances definition ongoing work
- Evaluation and further developments of the EMIF Catalogue: v3 has been released and includes Communities to allow external projects to have their own Catalogues/fingerprints
- Further work on the EMIF Ethics Code of Practice to analyse the specific ethical and legal aspects of data sharing in the context of EMIF
- Data flows improved and prototype developed to support data extraction (workflow management system)
- Mapping of a selection of data sources into the OMOP Common Data Model to evaluate feasibility and further integration of OHDSI tools into the Platform
- TranSMART improvements: cross-cohort analysis, omics data analysis pipelines integration and further analytical tools developed that can interact with it to support cohort work from the Research Topics
- Further developments on the Knowledge Objects Library and development of a plan to integrate into overall architecture
- Definition and start of Complex, full protocol-based Use Cases. Several Data Extraction runs supporting the Research Topics' scientific questions and addressing specific requirements for the Platform:
    - AD: Case-control design in AD (UC11), Inflammation and Dementia (UC12), Treatment Pathways in AD (UC13)
    - Metabolic: Cardiovascular disease and BMI (UC9), Liver disease (UC10)
    - Dedicated, cross-topic (Platform and Research topics) task forces working in collaboration with Data Custodians to design protocols

This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (EMIF grant n° 115372).

- Developments in Jerboa Reloaded: new modules added to support data extraction in Complex Use Cases
- First definition of the Business plan and start of the discussions with EFPIA and data custodians regarding value proposition
- Outreach: inclusion of one new data source into the Platform, with a proposed plan to strategically expand collaborations and increase volume and diversity of data sources

EMIF-AD

- EMIF Catalogue
    - Improved functionalities and interface of the Catalogue so that it can become publicly available for the AD community in the short term, allowing researchers worldwide to select relevant cohorts for scientific research questions and set up collaborations, thus facilitating AD research
    - Leverage of EMIF Catalogue by initiatives outside EMIF including IMI-EPAD, IMI-Prism, DP-UK, JPND-EADB and the Interdem consortium, showing that the Catalogue addresses an important need at least in the AD landscape in Europe
- Private Remote Research database environment
    - Harmonization of 6 cohorts that are now available in tranSMART, with 60 priority variables defined for which cross-cohort analyses can be performed, thus providing means for large-scale pooled analyses
- Preclinical AD cohort
    - The first 124 cognitively normal subjects have been included at VUmc and UNIMAN, with subjects undergoing deep phenotyping of a range of cognitive and clinical biomarkers; subjects from VUmc are monozygotic twins, being the first cohort study on AD biomarkers in twins of which preliminary data showed that AD biomarker traits have a moderately high genetic background (40-70%)
- Scientific analysis
    - Using data from existing cohorts, publication of 30 papers including 2 papers in JAMA. Some key papers:
        - Systematic literature review of the age-stratified prevalence of mild cognitive impairment and dementia in

This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (EMIF grant n° 115372).

European populations (Alexander, et al., Journal of Alzheimer's Disease, 2015)

- Prevalence and outcome of prodromal AD (Vos et al Brain 2015)
- Meta-analysis on prevalence of amyloid positivity in non-demented subjects (Jansen et al JAMA 2015)
- Meta-analysis on prevalence of amyloid positivity in demented subjects (Ossenkoppele et al JAMA 2015)
- Trajectories of decline in preclinical and prodromal AD (Bertens et al Alzheimer's and dementia 2015)
- Meta-analysis on uni- and multivariate models to predict the risk of incident dementia in population based datasets (Tang et al Plos One)
- Grey matter correlates of amyloid positivity in non-demented subjects (Tijms et al, Neurobiology of Aging 2015)
- Grey matter correlates of dementia family history and APOE genotype in non-demented subjects (Ten Kate et al, Neurobiology of Aging 2015)
- A novel multi-tissue RNA diagnostic tool (Sood, et al. Genome Biology 2015)

EMIF-Metabolic

- Completion of wet lab work for metabolomics analysis and genomic and transcriptomic profiling in distinct phenotypes
- Published a scientific paper showing that individuals with a high genetic risk for T2D (first-degree relatives to patients with T2D) have an increased susceptibility to the negative consequences of increased body fat including development of T2D
- An in-depth literature review of biomarkers associated with T2D has been completed and a manuscript published
- Completion of joint studies between Metabolic and AD Topics examining the relation between insulin resistance and biomarkers for AD in cerebrospinal fluid
- Potential new therapeutic options in NAFLD/NASH tested in pilot clinical study with promising results

This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (EMIF grant n° 115372).

- A detailed protocol for the analysis of NAFLD/NASH within Platform data sources has been completed, reviewed by the data owners, IRB approval obtained and data extraction activities started; Similar protocol around cardiovascular disease is also close to completion
- Extensive analysis of liver function markers in placebo patients has been performed in GSK and Janssen clinical trials and a manuscript is under preparation; other relevant clinical trials have been identified by EFPIA partners and analysis initiated
- Twelve scientific papers published and two additional submitted; many scientific presentations at international meetings/symposia

Cross topic achievements

- Completion of a joint EMIF-AD/EMIF-Metabolic study on assessing CSF biomarkers in insulin-resistant men started early 2014, including 60 subjects in total
- Establishment of Use Cases between EMIF-Platform and EMIF-Metabolic to assess prevalence of T2D in the EMIF-Platform databases, which is to be extended to exploring several factors influencing these diseases (e.g., lifestyle factors, cardiovascular diseases, medication use)
- Establishment of Use Cases between EMIF-Platform and EMIF-AD to assess prevalence and incidence of dementia and risk factors for dementia in the oldest-old in EHRs
- EMIF Catalogue: Major update of the EMIF data catalogue, both regarding the amount and the visualization of information available
- Data harmonization: 6 AD cohort datasets were harmonized using WebProtégé and uploaded in tranSMART for cross-cohort analysis and development of Knowledge Objects for EMIF-AD to support data harmonization and querying

EMIF outreach

- Presentation of the EMIF project at several scientific conferences and release of 43 peer-reviewed publications on the project's results in the third project year, with 11 additional manuscripts either in final stages of drafting or submitted for review.

This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (EMIF grant n° 115372).

Contacts

Coordinators: Bart Vannieuwenhuyse and Simon Lovestone
bvannieu@its.jnj.com/ simon.lovestone@psych.ox.ac.uk
EMIF-AD: Pieter Jelle Visser and Johannes Streffer
pj.visser@maastrichtuniversity.nl /JSTREFFE@its.jnj.com
EMIF-Metabolic: Dawn Waterworth and Ulf Smith
Dawn.M.Waterworth@gsk.com / ulf.smith@medic.gu.se
EMIF-Platform: Johan van der Lei and Nigel Hughes
j.vanderlei@erasmusmc.nl / nhughes@its.jnj.com

More information

www.emif.eu – info@emif.eu

www.imi.europa.eu

This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (EMIF grant n° 115372).