Connecting Northern Finland Birth Cohorts with Borealis and THL Biobanks (CoCoBi) – Current Progress and Future Plans

Kaisa Silander, THL
What is the CoCoBi project?

An infrastructure project with a scientific aim (s. 2017)

**Infrastructure aims:**
- Integration of data from birth cohorts, population-based cohorts and a hospital biobank
- IT solutions for data harmonization & big data mining
- Ethical & legal issues in dealing with sensitive data

**Scientific aims:**
Studying the determinants and trajectories of healthy aging
Northern Finland Birth Cohorts 1966 (n=12 231) and 1986 (n=9 479)
Oulu1935 and Oulu1945 (n=2000)

NORTHERN FINLAND COHORTS
Longitudinal research program to promote health and well-being of the population

Contact: Minna Männikkö, NFBC Research Director
E-mail: minna.mannikko@oulu.fi
Northern Finland Birth Cohort 1966: Data collection at 45-46 years in 2012 - 2014

**Health, behaviour, work and lifestyle**
- Diet, FFQ
- Fitness/activity (objective)
- Weight, height, bioimpedance

**Cognitive, sensory, psychological/mental:**
- Paired Associative Learning Test (PAL) using iPad
- Vision/eye tests [Hearing]

**Musculoskeletal/dental:**
- Muscle strength, spinal column position/posture (objective measure)
- Measures of osteoporosis
- Dental disorders

**Cardiovascular**
- BP (+central), heart and carotid artery ultrasound, ECG

**Respiratory, lung function, atopy:**
- Spirometry, skin prick tests

**Gastro-intestinal:**
- Kidney and liver function
- Chronic bowel inflammatory diseases

**Urinary tract, reproductive:**
- Menopausal symptoms, Incontinence, prostatic symptoms
- Blood and other samples: Measures of metabolic health, DNA, RNA, cells, urine, stool, saliva

**Data from different national registers**
- Hospital discharge register (HILMO)
- Medications and sick leaves (Social Insurance Institution of Finland, KELA)
- Cancers (Finnish register of cancers)
- Death and death causes (Finnish register of deaths)
- Infection diseases (Finnish register of infections)
- Employment, unemployment and pensions (Register of Finnish Centre for Pensions)

Slide from Minna Männikkö
Oulu1935 and Oulu1945 cohorts
Sirkka Keinänen-Kiukaanniemi et al.

• Studies on aging populations in Oulu were started in 1990 and 2000 for individuals born in 1935 (Oulu35) and 1945 (Oulu45) respectively.

• Aim was to find solutions to maintain individuals’ health, well-being and ability to function.

• Both cohorts comprise of c. 1000 individuals living in Oulu region with the follow-up data from the age of 55 until 80 years (Oulu1935) and until 70 years (Oulu1945).
THL Biobank’s population cohorts: The National FINRISK Study & Health 2000/2011

- Representative, cross-sectional population surveys carried out to assess the health of the population and the risk factors of chronic diseases
- FINRISK 35000 participants, Health 2000/2011 8500 participants
- Data available for each participant:
  - Baseline questionnaire/interview data (sociodemographic and socioeconomic characteristics, medical history, health status, nutrition, exercise, smoking, alcohol consumption, sleep, etc.)
  - Clinical examination
  - Laboratory measurements & omics data
  - Follow up from national health registers
The Health 2000 and 2011 Surveys

• A comprehensive health interview and health examination survey, carried out in 2000-2001, with a follow-up study during 2011-2012.
• >8500 participants aged 30 years and over donated samples
• Data available for research:
  • **Home health interview**: background & sociodemographic information, health & illnesses, use of medicines, use of health services, living habits, environment, functional capacity, work and work capacity, need for help and rehabilitation
  • **Health examination**: symptom interview, anthropometrics, bioimpedance, blood pressures, joint functions, hand grip strength, psychometric tests, resting ECG, spirometry, chest radiography, dental status, clinical examination by field physician
  • **Questionnaires** about life habits and nutrition
  • **Laboratory measurements** from serum and plasma samples (soon:NMR metabolomics)
  • **GWAS** from 7800 participants
Biobank Borealis: Clinical biobank

**Clinical data from hospital records and registries**

**Raw analysis data**

**Combination and harmonization of the data**

**Digital images & data analysis tools**

**Prospective samples:** annually 40,000 samples

**Pathology archives:** 1.8 M samples from 0.5 M individuals

**Finnish Maternity Cohort:** 2M serum samples from >0.9M pregnant women (1983-)

**Sample management**

**Hospital data**

**Sample collections**

Slide from Pia Nyberg
Project overview

Integration of clinical data from a hospital biobank with cohort data

Big data mining and analytics

Biobank Borealis
Other Finnish hospital biobanks

Novel predictive features for healthy/unhealthy aging
Early detection of persons in risk of age-related illnesses
Possibility for intervention

Oulu 1935 & 1945
NFBC 1966
NFBC 1986
FINRISK THL Biobank
Health 2000 THL Biobank

Variable harmonization
Regulatory, legal & ethical issues

• **Data must be combined using personal identification numbers:**
  • What agreements and permissions are needed?
  • Which organization can combine the personal identifiers?

• **Integraged individual-level data:**
  • Where should the integrated data be stored?
  • How can researchers obtain access to the data?
  • How is pseudonymization handled (e.g. large genomic data sets)?

→ Need to develop processes and tools that enable the sharing and combining of sensitive data, while respecting privacy regulations and the consent of the participants.
Regulatory, legal & ethical issues

- **Data must be combined using personal identification numbers:**
  - What agreements and permissions are needed?
  - Which organization can combine the personal identifiers?

- **Integraged individual-level data:**
  - Where should the integrated data be stored?
  - How can researchers obtain access to the data?
  - How is pseudonymization handled (e.g. large genomic data sets)?

→ Need to develop processes and tools that enable the sharing and combining of sensitive data, while respecting privacy regulations and the consent of the participants.
Overlap analyzed between NFBC cohorts (1966, 1986, 1935 & 1945) and THL’s population cohorts FINRISK and Health 2000

A total of 1047 individuals have participated in NFBC and THL studies
- FINRISK 2002: 315 ind.
- FINRISK 2007: 151 ind.
- Health 2011: 87 ind.

Of these, approximately 800 are unique individuals, while 200 have participated in more than one THL study
Age distribution of overlapping sample donors

# of FINRISK and H2000 participants

- NFBC 1986
- NFBC 1966
- Oulu 1945
- Oulu 1935

Age at 2017

# Men | # Women
Development of the reference dataset

Attributes were selected from research projects focusing on aging and lifelong health

- 17 attributes originate from DynaHEALTH project (a H2020-project)
  - Focus on factors of metabolic syndrome

- Complementing attributes from the research of aging from the University of Oulu projects
  - Focus on frailty

- The finalized reference dataset consist of 116 attributes, from the following categories:
  - Background information, physiological measurements, cognitive factors, diseases, functional capacity, general health, psychological and socioeconomical factors & work capacity

More details: See poster by Niina Eklund
BiobankUniverse for mapping attributes

- Open source tool created by Chao Pang & Morris Swertz
- Maps attributes using ontology, semantic and lexical matching
- Pang et al. BIOINF, 2017
Next steps: Overlaps and data integration

• Finish testing overlap between THL and Borealis Biobaks and NFBC and Borealis Biobank
• Compare the reference data set with the data available for each of the sample donors
• Prepare harmonized data sets for the selected ’healthy aging’ variables for all sample collections
Big data challenges

• **Omics data handling:**
  • GWAS for >50000 participants
  • Sequence data (WES & WGS) for >5000 participants
  • Serum NMR metabolomics (134 metabolic measures) for >45000 participants

• **Hospital data - extraction from various hospital databases:**
  • Pathology SNOMED-coded diagnoses and organs
  • Radiology imaging
  • Laboratory values
  • Diagnoses in ICD-10 codes, surgeries, etc.
  • In-house patient record system

• Data collected from cohort study participants (500-10000 variables) & linkage to national registries

→ **A system that supports**
  • robust data extraction and merge procedures
  • metadata management
  • storing the of the data in secure, high-performance and analytics-ready manner
  • Service for efficient data usage
Consortium members

- NFBC project center, University of Oulu
  - PI: Minna Männikkö
  - Heli Lehtiniemi
  - Samppa Rohkimainen

- Biobank Borealis, Oulu University Hospital
  - Pia Nyberg

- THL Biobank, THL, Helsinki
  - Kaisa Silander
  - Niina Eklund
  - Salla-Maaria Pätsi

- The Center for Advanced System Studies along with Computer science group, University of Oulu
  - Tapio Seppänen
  - Jaakko Sauvola

Funding:

[Academy of Finland logo]