MRC-HPA Centre for Environment & Health

Imperial College London

www.environment-health.ac.uk

BioSHaRE Environmental Determinants of Health project

Opportunities and challenges of cross-cohort working

CLOSER workshop on cross-cohort research:
Opportunities, challenges and examples
9th Sept 2015

Dr Susan Hodgson
(susan.hodgson@imperial.ac.uk)

Content

- BioSHaRE
  - The Environmental Determinants of Health Project
  - The study cohorts
- Harmonisation
  - Exposures
  - Outcomes
  - Covariates
- Analysis
  - Meta-analysis
  - Pooled individual analysis
- Discussion
  - Opportunities
  - Challenges
  - Cross-cohort research - moving forwards
- References

BioSHaRE

BioSHaRE

Biobank Standardisation and Harmonisation for Research Excellence in the European Union

Mission:

To facilitate data harmonisation and standardisation, data sharing and pooling across multiple biobanks and databases

Why:

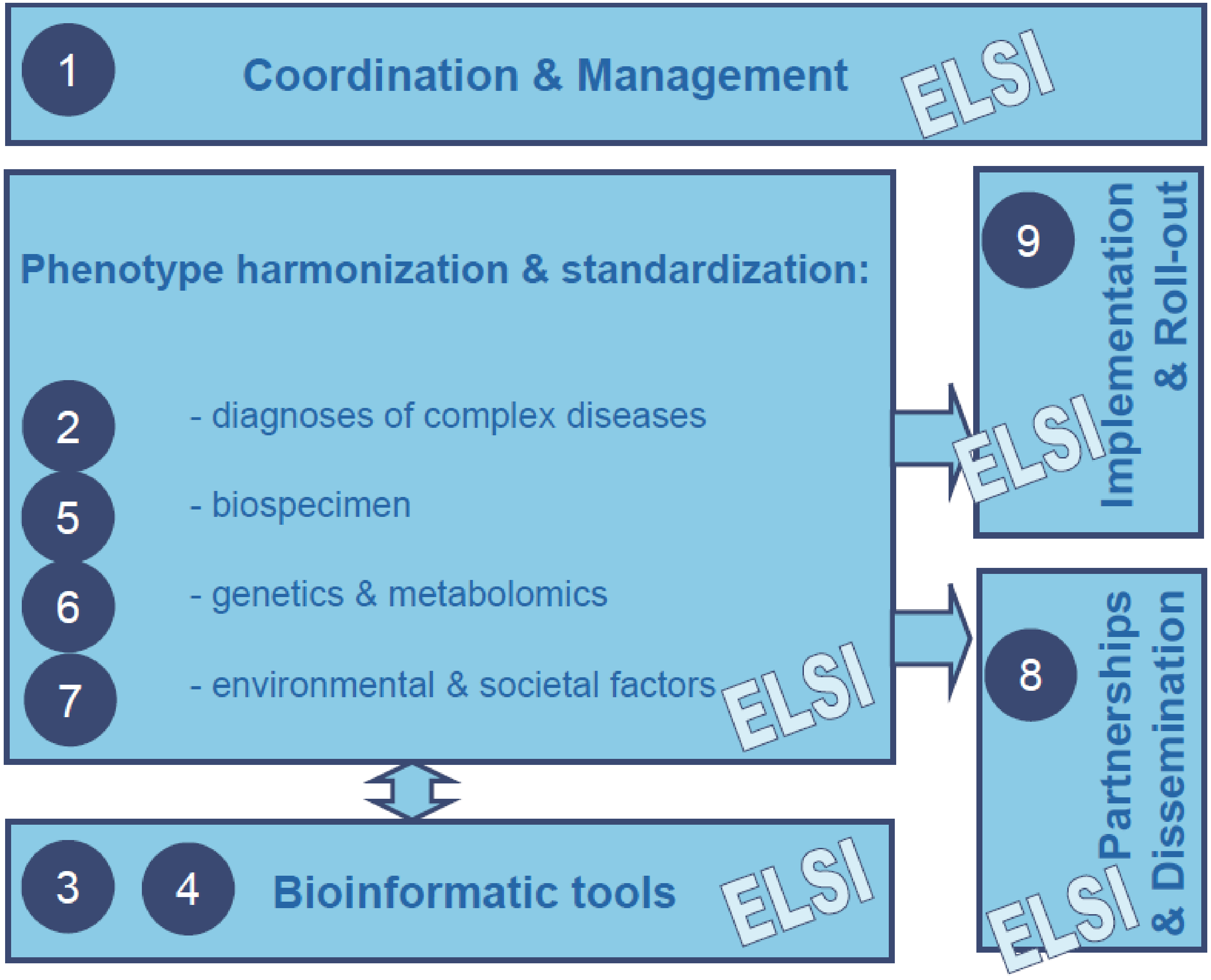
For many scientific questions no single study provides adequate numbers of subjects that are measured/assessed sufficiently well – biobanks must therefore be harmonised and standardised so that studies can pool biobank data in valid and effective ways

Who:

Consortium of leading population-based cohort studies, with international researchers from diverse domains of biobanking science, including epidemiologists, statisticians, software developers and ELSI experts

https://www.bioshare.eu/

BioSHaRE - Work packages



1
Coordination & Management
ELSI
Phenotype harmonization & standardization:
2
- diagnoses of complex diseases
5
- biospecimen
6
- genetics & metabolomics
7
- environmental & societal factors
ELSI
9
Implementation & Roll-out
ELSI
8
Partnerships & Dissemination
ELSI
3
4
Bioinformatic tools
ELSI

Figure 1.3.1.1: Conceptual overview of the BioSHaRE-EU project

BioSHaRE - Projects

- BioSHaRE-IT
- Environmental Determinants of Health Project
- Healthy Obese Project (HOP)
  - HOP 1 - Associations between smoking, components of the metabolic syndrome and lipid composition in cohorts participating in the BioSHaRE-EU consortium
  - HOP 2 - Lipids and statin use in Healthy Obesity Project – Phase I
- Metabolomics Project
- Social implications of biobanking
- What components drive the metabolic syndrome?

Environmental Determinants of Health Project

Aim:

To study how environmental exposures affect chronic multifactorial diseases

Focus:

Environmental exposure to road-traffic noise and air pollution, with harmonized exposure measures being assigned to participants across multiple cohorts in different countries

MRC-HPA Centre for Environment & Health

Environmental Determinants of Health Project

Wilma Zijlema

(University of Groningen, the Netherlands)

The effect of environmental noise on blood pressure and heart rate

Samuel Cai

(Imperial College London, UK)

The effect of road traffic noise and air pollution and cardiorespiratory health

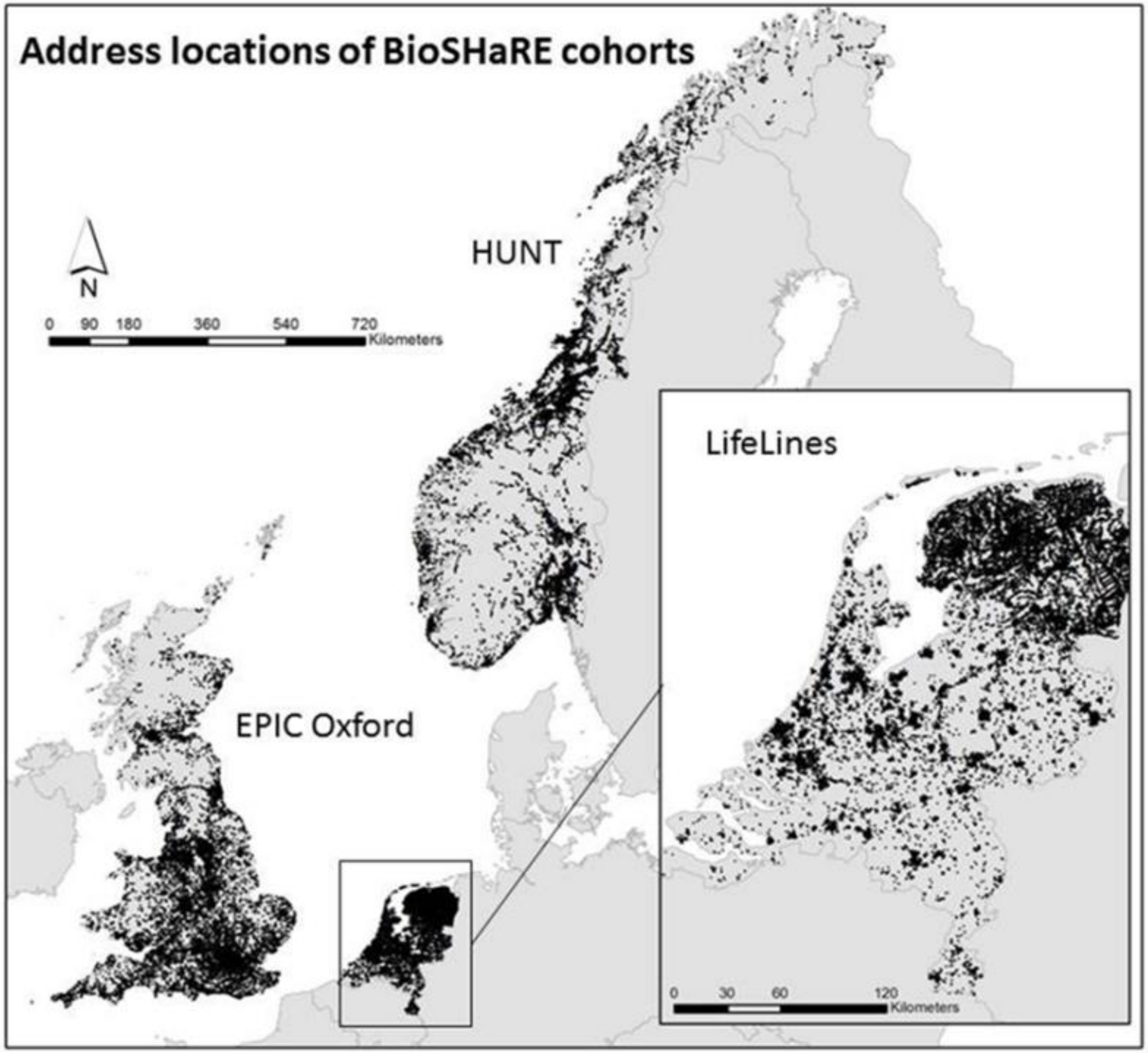
Cohorts

EPIC-Oxford - 57,000 participants recruited in 1993-1999 from across the UK

HUNT - 50,000 participants recruited in 1984-1986 from the Nord-Trøndelag County, Norway

Lifelines - 95,000 participants recruited in 2007-2013 from the Groningen, Friesland, Drenthe regions of the Netherlands

UK Biobank - 500,000 participants recruited in 2006-2010 from across the UK


Address locations of BioSHaRE cohorts
N
0 90 180 360 540 720 Kilometers
HUNT
EPIC Oxford
LifeLines
0 30 60 120 Kilometers

Harmonisation

Harmonisation vs standardisation

- **Standardisation** implies that two or more studies have access to, and adopt and apply, precisely the same protocols, e.g.
  - the same questionnaires
  - the same sample collection and processing protocols
  - the same IT/database structures
- **Harmonisation** implies that although the protocols of two studies may be different, the information generated carries a similar enough meaning to allow the two studies to be pooled…

Harmonisation - Exposures

Aim: develop harmonised air pollution and noise exposure models

- Able to be assigned at the individual address-level
- And to cohort participants across the EU

Process:

- Secure permission from each cohort to access participant address data
- Geocode each participant address
- Run the exposure models, to assign exposure measures to each cohort participant
- Return these exposure data to each cohort

Air pollution

ESCAPE Land Use Regression model

Land Use Regression model developed within FP7 funded European Study of Cohorts for Air Pollution Effects (ESCAPE)*

Participants assigned address-level annual average estimates:

- Nitrogen dioxide
- Nitrogen oxides
- PM10
- PM2.5
- PM2.5 absorbance
- PM coarse

http://www.escapeproject.eu/

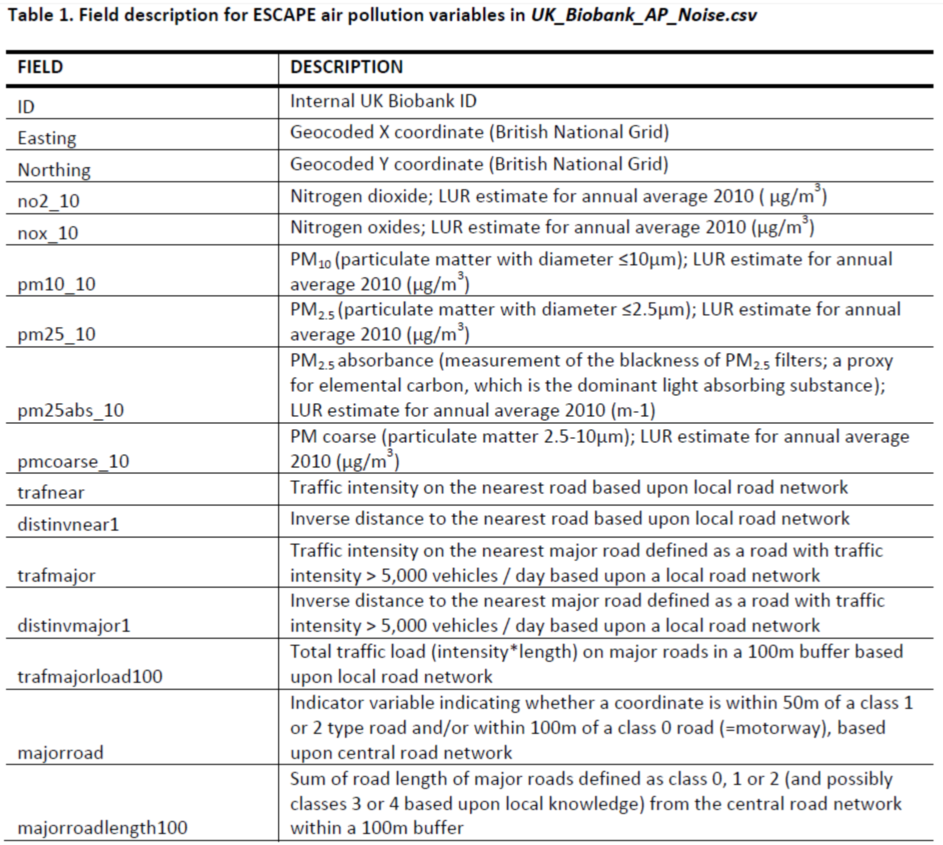


Table 1. Field description for ESCAPE air pollution variables in *UK_Biobank_AP_Noise.csv*

FIELD	DESCRIPTION
ID	Internal UK Biobank ID
Easting	Geocoded X coordinate (British National Grid)
Northing	Geocoded Y coordinate (British National Grid)
no2_10	Nitrogen dioxide; LUR estimate for annual average 2010 (μg/m^3)
nox_10	Nitrogen oxides; LUR estimate for annual average 2010 (μg/m^3)
pm10_10	PM$_{10}$ (particulate matter with diameter ≤10μm); LUR estimate for annual average 2010 (μg/m^3)
pm25_10	PM$_{2.5}$ (particulate matter with diameter ≤2.5μm); LUR estimate for annual average 2010 (μg/m^3)
pm25abs_10	PM$_{2.5}$ absorbance (measurement of the blackness of PM$_{2.5}$ filters; a proxy for elemental carbon, which is the dominant light absorbing substance); LUR estimate for annual average 2010 (m-1)
pmcoarse_10	PM coarse (particulate matter 2.5-10μm); LUR estimate for annual average 2010 (μg/m^3)
trafnear	Traffic intensity on the nearest road based upon local road network
distinvnear1	Inverse distance to the nearest road based upon local road network
trafmajor	Traffic intensity on the nearest major road defined as a road with traffic intensity > 5,000 vehicles / day based upon a local road network
distinvmajor1	Inverse distance to the nearest major road defined as a road with traffic intensity > 5,000 vehicles / day based upon a local road network
trafmajorload100	Total traffic load (intensity*length) on major roads in a 100m buffer based upon local road network
majorroad	Indicator variable indicating whether a coordinate is within 50m of a class 1 or 2 type road and/or within 100m of a class 0 road (=motorway), based upon central road network
majorroadlength100	Sum of road length of major roads defined as class 0, 1 or 2 (and possibly classes 3 or 4 based upon local knowledge) from the central road network within a 100m buffer

** Eeftens et al. 2012; Beelen et al. 2013*

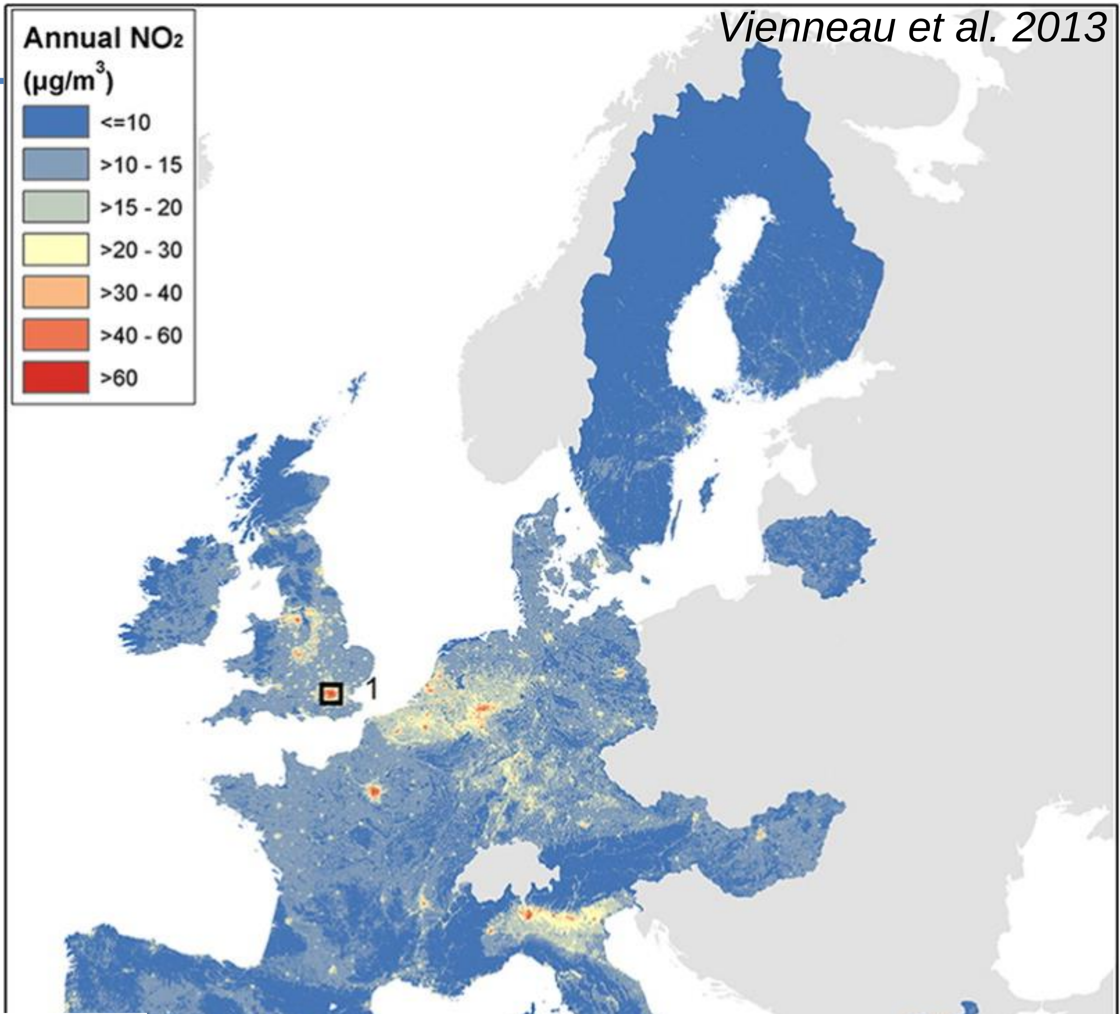
MRC-HPA Centre for Environment & Health

Air pollution

EU-wide air pollution maps based on a LUR model for Europe enhanced with satellite derived air pollution estimates

Used to assign:

- Nitrogen dioxide
- PM10



Vienneau et al. 2013

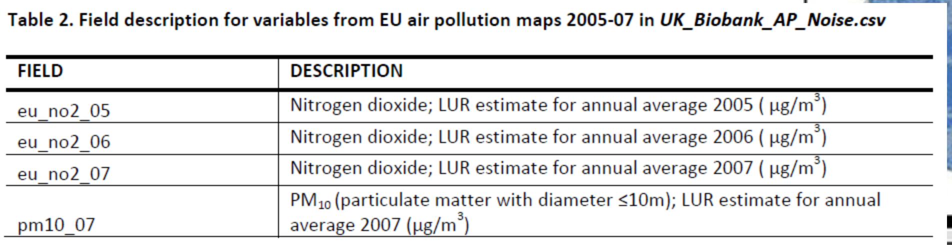


Table 2. Field description for variables from EU air pollution maps 2005-07 in *UK_Biobank_AP_Noise.csv*

FIELD	DESCRIPTION
eu_no2_05	Nitrogen dioxide; LUR estimate for annual average 2005 (µg/m^3)
eu_no2_06	Nitrogen dioxide; LUR estimate for annual average 2006 (µg/m^3)
eu_no2_07	Nitrogen dioxide; LUR estimate for annual average 2007 (µg/m^3)
pm10_07	PM_{10} (particulate matter with diameter ≤10m); LUR estimate for annual average 2007 (µg/m^3)

Road traffic noise

Developed a harmonised pan-European noise exposure model (Morley et al 2015)

Based on a modified Common NOise aSSessment methOdS (CNOSSOS) model*

Participants from each cohort assigned annual average estimates of noise exposure:

- Daytime
- Evening
- Overnight
- 16 hour mean
- A-weighted 24 hours

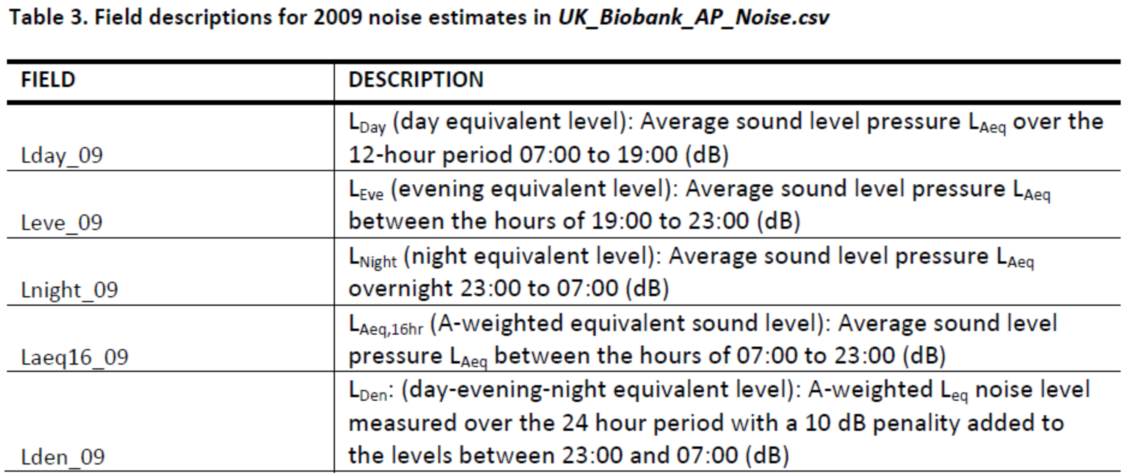
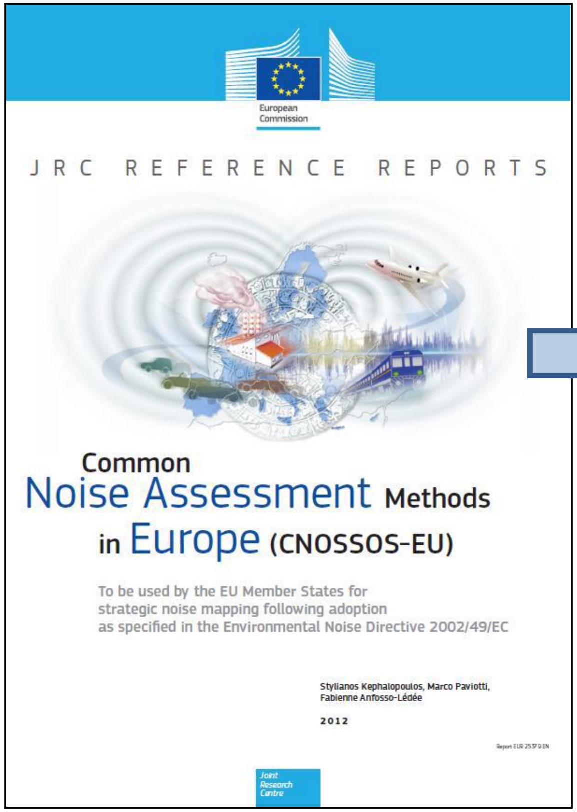


Table 3. Field descriptions for 2009 noise estimates in *UK_Biobank_AP_Noise.csv*

FIELD	DESCRIPTION
Lday_09	L_{Day} (day equivalent level): Average sound level pressure L_{Aeq} over the 12-hour period 07:00 to 19:00 (dB)
Leve_09	L_{Eve} (evening equivalent level): Average sound level pressure L_{Aeq} between the hours of 19:00 to 23:00 (dB)
Lnight_09	L_{Night} (night equivalent level): Average sound level pressure L_{Aeq} overnight 23:00 to 07:00 (dB)
Laeq16_09	$L_{Aeq,16hr}$ (A-weighted equivalent sound level): Average sound level pressure L_{Aeq} between the hours of 07:00 to 23:00 (dB)
Lden_09	L_{Den}: (day-evening-night equivalent level): A-weighted L_{eq} noise level measured over the 24 hour period with a 10 dB penality added to the levels between 23:00 and 07:00 (dB)

*Kephalopoulos et al. 2012; Kephalopoulos et al. 2014

The ‘EnviroSHAPER’



European Commission

JRC REFERENCE REPORTS

Common Noise Assessment Methods in Europe (CNOSSOS-EU)

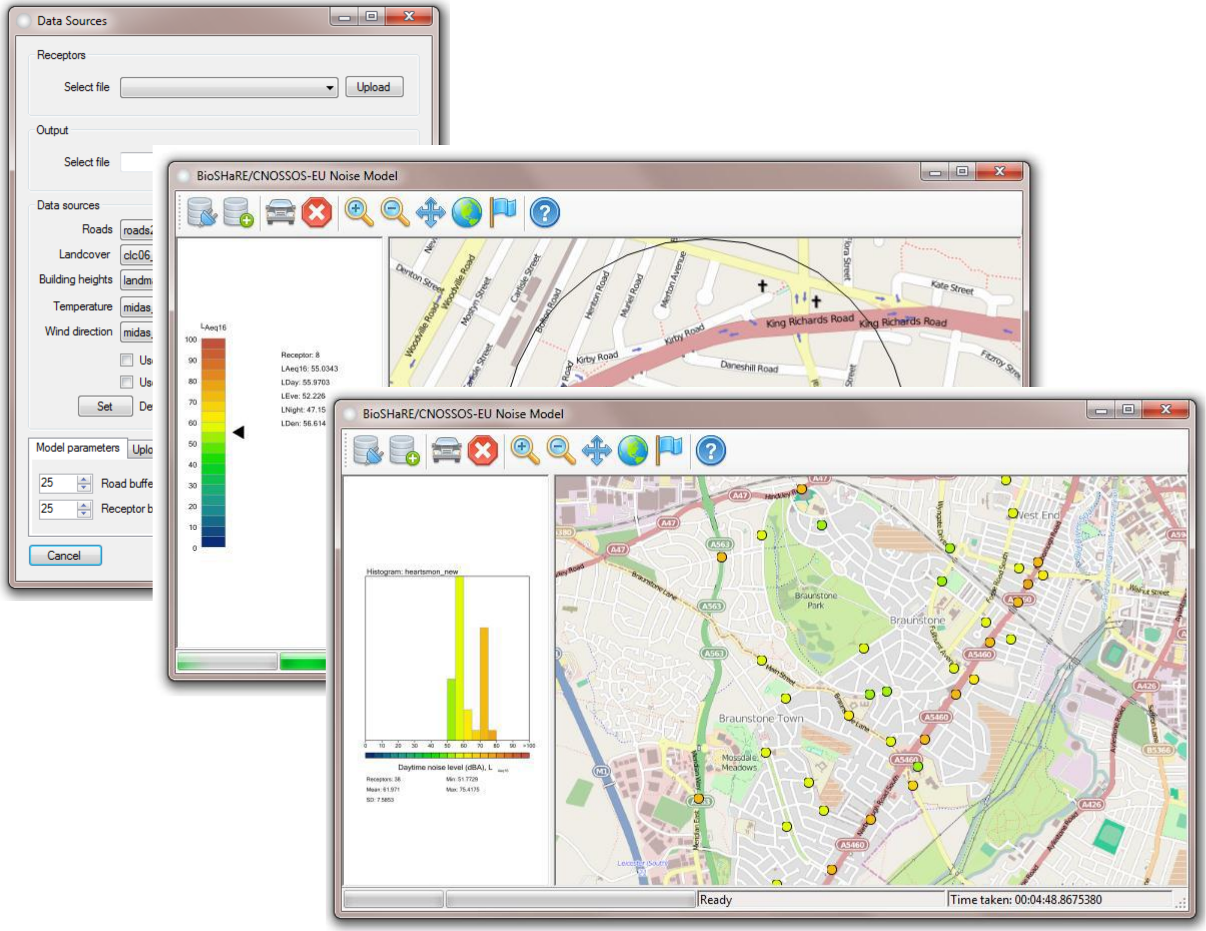
To be used by the EU Member States for strategic noise mapping following adoption as specified in the Environmental Noise Directive 2002/49/EC

Stylianos Kephalopoulos, Marco Paviotti, Fabienne Anfosso-Lédée

2012

Report EUR 25379 EN

Joint Research Centre



Data Sources

Receptors

Select file [▾] Upload

Output

Select file

Data sources

Roads: roads2

Landcover: clc06_

Building heights: landm

Temperature: midas_

Wind direction: midas_

□ Us

□ Us

Set De

Model parameters | Uplo

25 Road buffe

25 Receptor b

Cancel

BioSHaRE/CNOSSOS-EU Noise Model

Receptor: 8
LAeq16: 55.0343
LDay: 55.9703
LEve: 52.226
LNight: 47.15
LDen: 56.614

BioSHaRE/CNOSSOS-EU Noise Model

Receptors: 38
Mean: 61.971
SD: 7.5853
Min: 51.7729
Max: 75.4175

Ready

Time taken: 00:04:48.8675380

BioSHaRE.eu

MRC-HPA Centre for Environment & Health

Imperial College London

MRC Medical Research Council

Health Protection Agency

King's College London

Harmonisation - Exposures

Cohort	ESCAPE Air polln	EU-wide Air polln	CNOSSOS Noise
Lifelines	√	√	√
HUNT	x	√	√
EPIC-Oxford	√	√	√
UK Biobank	√	√	√

Retrospective harmonisation

Founded on the DataSHaPER (DataSchema and Harmonization Platform for Epidemiological Research) harmonization approach, which involves a:

- Project specific ‘DataSchema’
  - Describes a set of harmonised variables of value in a particular scientific context.
- Corresponding Harmonisation Platform
  - Contains pairing rules that determine whether the information collected by each participating cohort can be used to construct these predefined harmonised variables.
- Algorithms
  - Applied to each cohort to create the new harmonised variables.

(Doiron et al. 2013).

For the BioSHaRE Environmental Project, 46/60 target variables in the DataSchema able to be harmonised across UK Biobank, HUNT, LifeLines and EPIC-Oxford…

Harmonisation - Outcomes

- Mortality/morbidity outcomes
  - Coded via the International Classification of Diseases (ICD)
  - Standard diagnostic tool
  - HARMONISED!
- Blood pressure (systolic/diastolic)
  - Available in all cohorts
  - BUT measurement protocols varied by cohort
    - In EPIC-Oxford, BP was measured by a trained health professional; the mean of 2 measures was taken
    - In HUNT, three measurements were taken for each participant, and the mean of the 2nd and 3rd measurement used
    - In Lifelines, 10 measurements were made, with the final two averaged and used
  - HARMONISED?

Harmonisation - Outcomes

- Blood biochemistry data
  - Available in HUNT and Lifelines, but not EPIC-Oxford, or (yet) UK Biobank
    - total serum cholesterol
    - High-sensitivity C-reactive protein
    - Triglycerides
    - high-density lipoprotein (HDL) cholesterol
    - glucose
  - Non-fasting blood samples (HUNT) vs fasting blood samples (Lifelines)
  - HARMONISED?

Harmonisation - Covariates

[Earlier…body size / socioeconomic resources…]

In BioSHaRE…

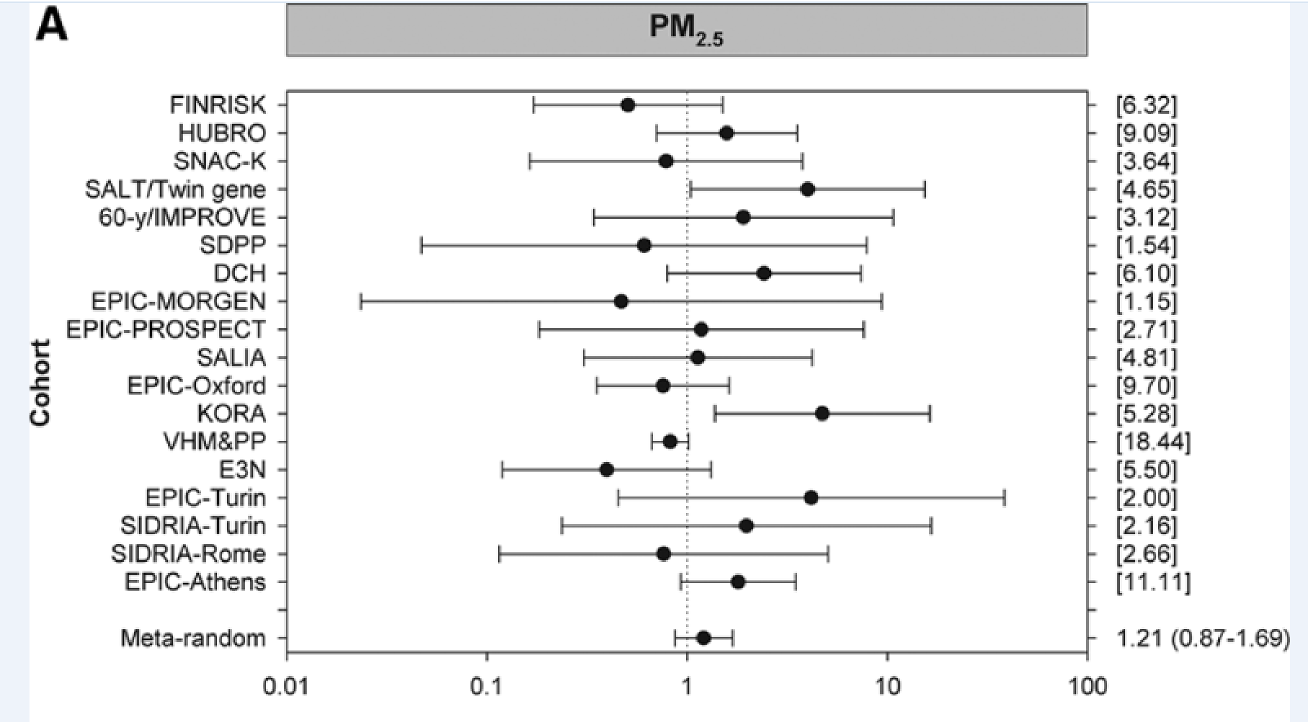
- Age at interview – easy!
  - Directly recorded (HUNT, EPIC-Oxford)
  - Calculable using DoB & date of interview (UK Biobank)
- Alcohol consumption (grams) per week – difficult!
  - EPIC-Oxford – participants asked '*grams alcohol/day*'; x 7 = consumption per week
  - Lifelines - participants asked '***How often** did you drink alcohol in the past month?*' and '*On days that you drank alcohol, **how many** glasses did you drink on average*?'. 'How often' x 'how many' x 9.9 grams of alcohol in one standard serving = consumption per week
- Diet – impossible!

Analysis

Meta-analysis vs individual analysis

ESCAPE study conducted *study-level meta-analysis* of air pollution effects on health:

- Only exploits within-cohort exposure contrasts
- Adjustment for confounding differs by cohort, leaving differing degrees of residual confounding
- Inflexible; exploratory analysis of sub-groups/interactions hindered



Beelen et al 2014

Analysis

In BioSHaRE, harmonisation allows cross-cohort analyses to be undertaken *at the individual-level*

+ greater exposure differentials

+ greater statistical power/efficiency

+ greater flexibility for exploratory analyses

BUT:

- harmonisation reflects the 'lowest common denominator' achievable across cohorts

Cohort specific analyses should also be undertaken using the best available exposure/covariate/outcome measures for that cohort to aid interpretation...

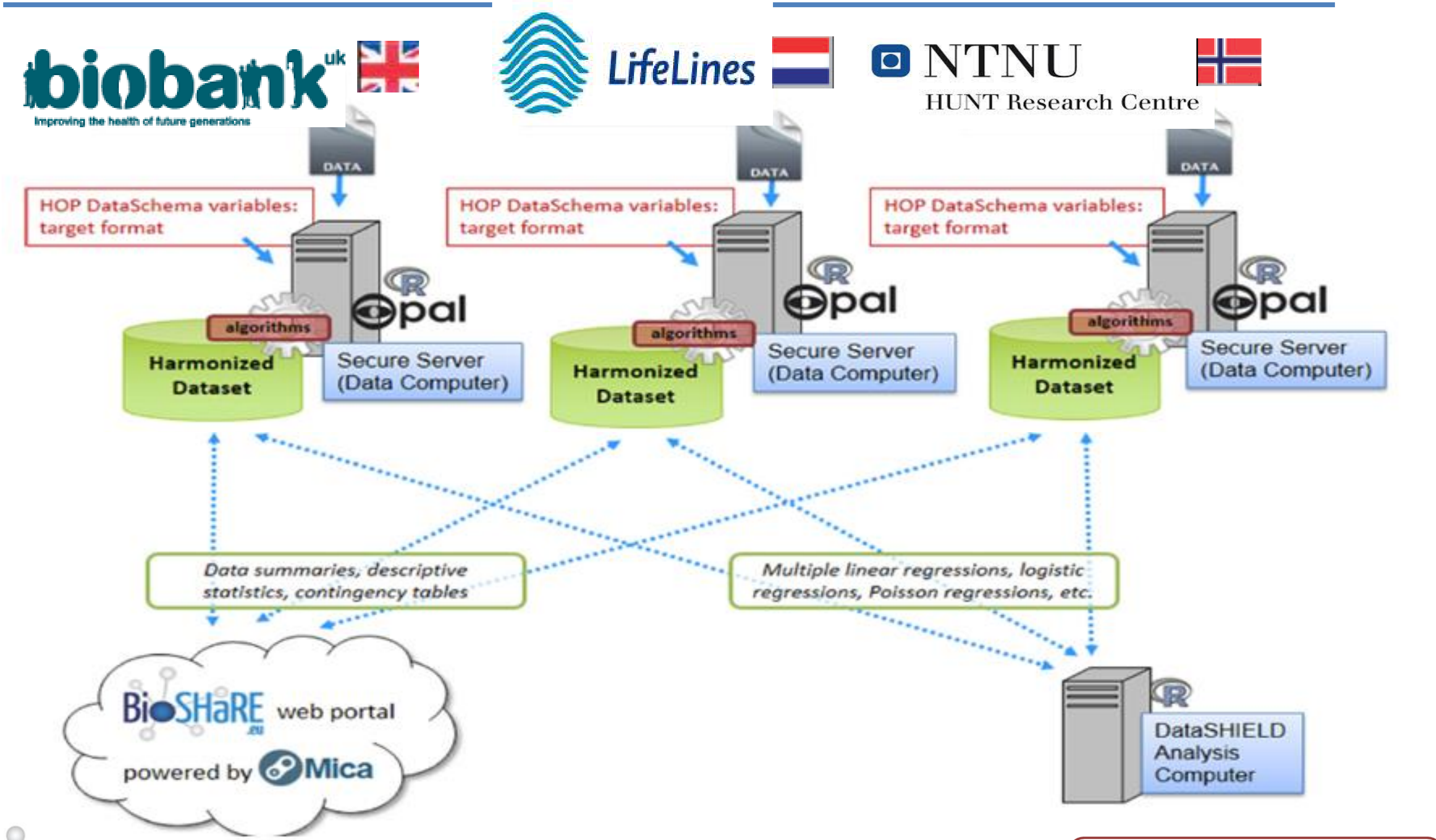
DataSHIELD

In BioSHaRE, data from the different cohorts held locally and *virtually pooled* using Data Aggregation Through Anonymous Summary-statistics from Harmonized Individual levEL Databases (DataSHIELD) *[Gaye et al. 2014].*

DataSHIELD:

- provides a solution when ethico-legal considerations prevent data-sharing
- promotes and facilitates collaborations by empowering data owners and affording them better control over their data.
- improves the governance and management of data by allowing them to be maintained locally.

DataSHIELD


biobank uk
Improving the health of future generations
LifeLines
NTNU
HUNT Research Centre
DATA
DATA
DATA
HOP DataSchema variables: target format
HOP DataSchema variables: target format
HOP DataSchema variables: target format
R
Opal
R
Opal
R
Opal
algorithms
algorithms
algorithms
Harmonized Dataset
Harmonized Dataset
Harmonized Dataset
Secure Server (Data Computer)
Secure Server (Data Computer)
Secure Server (Data Computer)
Data summaries, descriptive statistics, contingency tables
Multiple linear regressions, logistic regressions, Poisson regressions, etc.
BioSHaRE.eu web portal
powered by Mica
R
DataSHIELD Analysis Computer

http://datashield.org/

MRC-HPA Centre for Environment & Health

Discussion

Opportunities from cross-cohort working

Via BioSHaRE, we have:

- Enriched ‘harmonized’ cohort data
- Undertaken proof of principal studies
- Developed knowledge, skills, infrastructure & tools, and people

In terms of epidemiology, our approach has ensured:

- A large sample size
- Greater exposure differential
  - Both required to assess the risks associated with environmental exposures

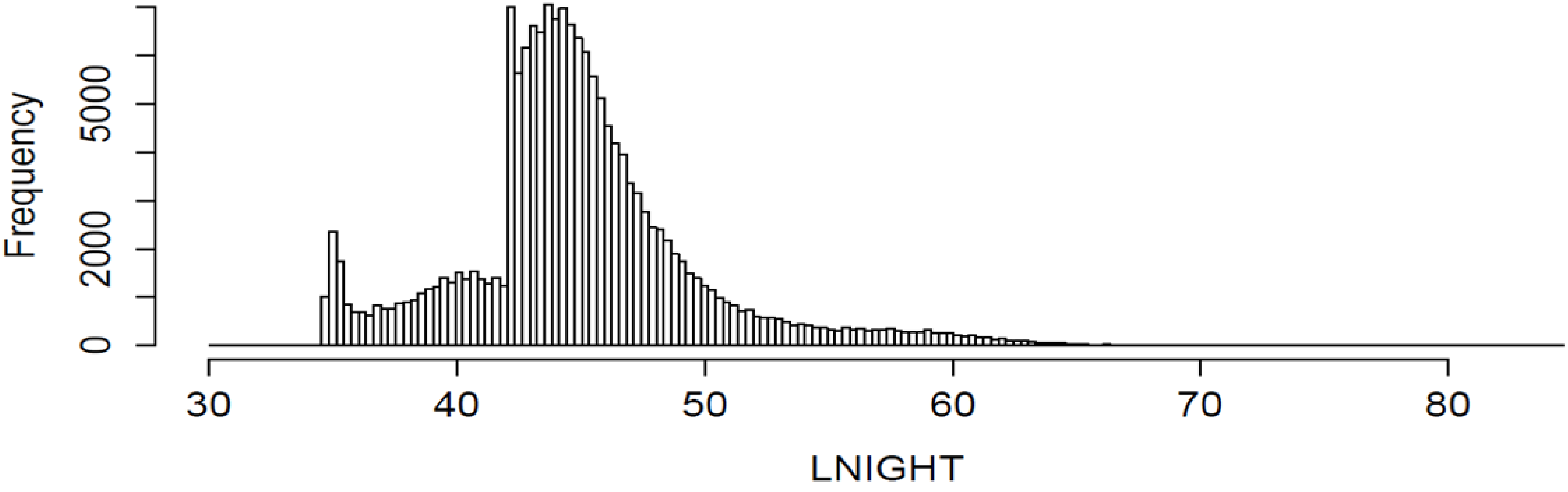
Challenges of cross-cohort working

- Obtaining cohort data
  - Data access time consuming, complex, non standardised
- Harmonisation
  - Assigning ‘harmonised’ environmental exposures (e.g. no ESCAPE data for HUNT)
  - Harmonisation not possible for 14/60 target variables
  - Harmonisation = lowest common denominator
    - Meta-analysis, with full adjustment within cohort, might offer better insights?

Challenges of cross-cohort working

- Exposure differentials
  - Limited within cohorts but non comparable across cohorts?

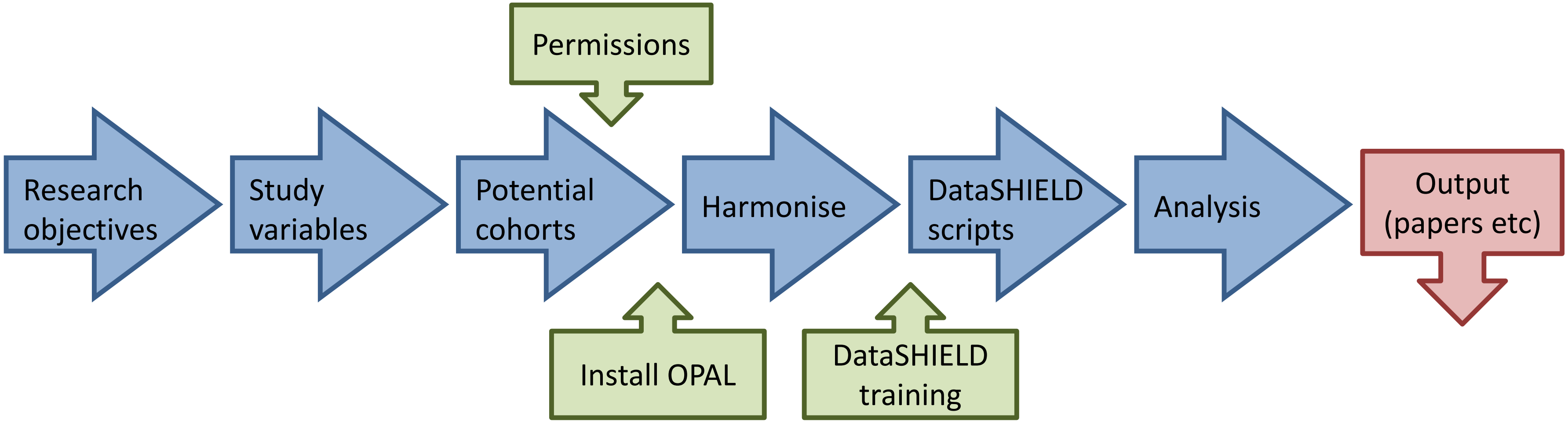
Night-time (23:00-07:00) noise (pooled)



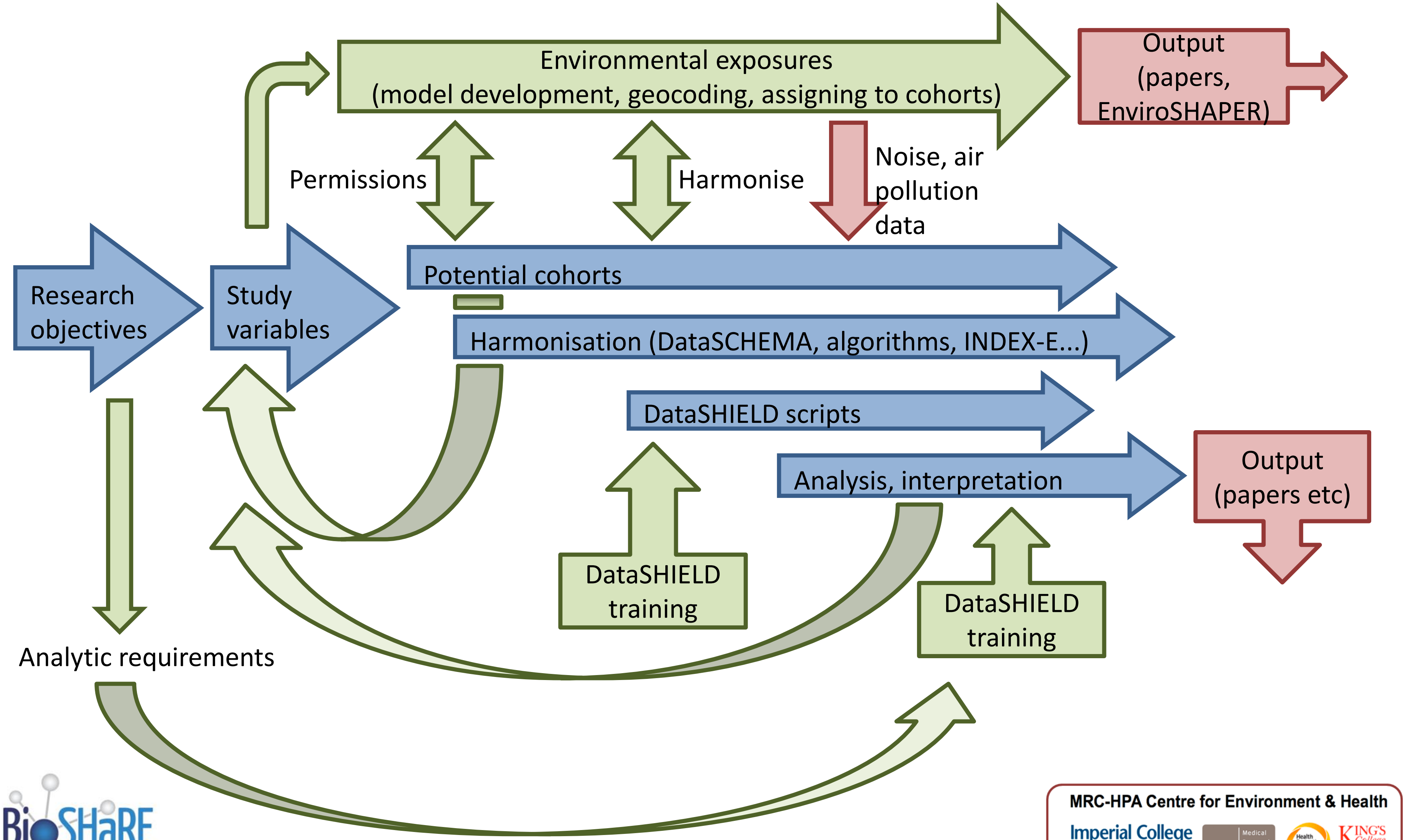
Challenges of cross-cohort working

- Analysis (e.g. via DataSHIELD)
  - Pioneering, but problematic!
- Ideal process…vs reality…

Timeline (ideal)



Timeline (reality)



MRC-HPA Centre for Environment & Health

Cross-cohort research - moving forwards

Cross cohort working exciting and promising:

- Provides the sample sizes needed to assess risk factors in complex diseases

Recommend:

- Further validation of developing tools and techniques
- Buy-in from existing & yet to be established cohorts
- Standardisation (i.e. prospective harmonisation!) to avoid ‘lowest common denominator’ issue
- Harmonised data access/application process to facilitate cross-cohort research?

References

Acknowledgements

MRC-PHE centre:
Samuel Cai
Dr Marta Blangiardo
Dr David Morley
Dr John Gulliver
Federico Fabbri
Dr Anna Hansell
Prof Paul Elliott.

Swiss TPH:
Dr Kees de Hoogh

Key BioSHaRE Collaborators:
Prof Paul Burton
Dr Amadou Gaye
Dr Isabel Fortier
Dr Dany Doiron
Dr Yannick Marcon
Dr Stephane Mbatchou
Wilma Zijlema
Prof Judith Rosmalen

Data providers:
EPIC-Oxford
Prof Tim Key
Dr Paul Appleby

Lifelines
Prof Ronald Stolk

HUNT
Prof Kristian Hveem

UK Biobank

EPIC-Turin
Prof Paolo Vineis

MRC-HPA Centre for Environment & Health

Beelen RG, Hoek D, Vienneau M et al. 2013. Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe – The ESCAPE project. Atmospheric Environment 72:10-23

Eeftens M, Beelen R, de Hoogh K et al. 2012. Development of Land Use Regression Models for PM2.5, PM2.5 Absorbance, PM10 and PMcoarse in 20 European Study Areas; Results of the ESCAPE Project. Environmental Science & Technology 46(20):11195-11205

Vienneau D, de Hoogh K, Bechle MJ et al. 2013. Western European Land Use Regression Incorporating Satellite- and Ground-Based Measurements of NO2 and PM10. Environmental Science & Technology 47(23):13555-13564

Morley DW, de Hoogh K, Fecht D et al. 2015. International scale implementation of the CNOSSOS-EU road traffic noise prediction model for epidemiological studies. Environmental Pollution 206:332-341

Doiron D, Burton P, Marcon Y, et al. 2013. Data harmonization and federated analysis of population-based studies: the BioSHaRE project. Emerg Themes Epidemiol. 10(1):12.

Beelen R, Stafoggia M, Raaschou-Nielsen O. 2014. Long-term Exposure to Air Pollution and Cardiovascular Mortality: An Analysis of 22 European Cohorts. Epidemiology 25(3):368–378

Gaye A, Marcon Y, Isaeva J, et al. 2014. DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol. 43(6):1929-44.