



UNIVERSITY OF  
**LEICESTER**



University Hospitals of Leicester **NHS**  
NHS Trust

# The Challenges of Integrating Data for Environmental Health Research

**Dr. Jonathan Tedds**

Senior Research Fellow

Health and Research Data Informatics, Health Sciences

Co-chair RDA-WDS Publishing Data IG; UK Environmental Observation Framework DAG

**Dr. Joshua Vande Hey**

NERC Knowledge Exchange Fellow in Aerosols & Health

Earth Observation Science, Physics & Astronomy

<https://www.jiscmail.ac.uk/lists/ENVIROHEALTH-INFORMATICS>

RDA IG/WG Collaboration Meeting, Nottingham, 7 June 2016



UNIVERSITY OF  
**LEICESTER**

# Integrated Health and Environmental Data requires an Informatics-based Approach



# What is Health Informatics?

Let's ask the NHS:

“In its most simplest term, health informatics is about getting the right information to the right person at the right time. It is critical to the delivery of information to healthcare professionals so they can deliver the most appropriate care.”

“Health informatics is one of the fastest growing areas within healthcare.”



UNIVERSITY OF  
**LEICESTER**

What informatics are we talking about in health research?



# What informatics are we talking about in health research?

Health research informatics involves empowering health researchers to interact with and apply their data in efficient and comprehensive ways.



# What informatics are we talking about in health research?

Health research informatics involves empowering health researchers to interact with and apply their data in efficient and comprehensive ways.

However, this also relates to working with cohort patients.



# What informatics are we talking about in health research?

Health research informatics involves empowering health researchers to interact with and apply their data in efficient and comprehensive ways.

However, this also relates to working with cohort patients.

And there is potential for applying this technology and these techniques more broadly in healthcare.



UNIVERSITY OF  
**LEICESTER**

# What are Environmental Variables?





UNIVERSITY OF  
**LEICESTER**

# What are Environmental Variables?

In healthcare and health research, environment can mean many things:



# What are Environmental Variables?

In healthcare and health research, environment can mean many things:

- Parameters in the treatment environment

Hospital layout, cleanliness, lighting, noise, occupancy, amenities, indoor air quality (pathogens and chemicals)...



# What are Environmental Variables?

In healthcare and health research, environment can mean many things:

- Parameters in the treatment environment  
Hospital layout, cleanliness, lighting, noise, occupancy, amenities, indoor air quality (pathogens and chemicals)...
- Parameters in the home environment  
Cleanliness, ease of access, noise, indoor air quality, things influencing development such as toys and other interactive objects...



# What are Environmental Variables?

In healthcare and health research, environment can mean many things:

- Parameters in the treatment environment  
Hospital layout, cleanliness, lighting, noise, occupancy, amenities, indoor air quality (pathogens and chemicals)...
- Parameters in the home environment  
Cleanliness, ease of access, noise, indoor air quality, things influencing development such as toys and other interactive objects...
- Parameters in the social environment  
Family nutrition, community dietary habits, interaction with social networks, social or societal stress...



# What are Environmental Variables? II

Then we get to the things geoscientists usually think of first:

# What are Environmental Variables? II

Then we get to the things geoscientists usually think of first:

- Parameters in the physical environment

Outdoor air quality, radioactive isotopes in the soil, water quality, noise, vegetation, topography, solar radiation, light pollution, surface temperature...



# What are Environmental Variables? II

Then we get to the things geoscientists usually think of first:

- Parameters in the physical environment

Outdoor air quality, radioactive isotopes in the soil, water quality, noise, vegetation, topography, solar radiation, light pollution, surface temperature...

So we must be clear about what we mean when talking about environment and health



UNIVERSITY OF  
**LEICESTER**

# How can we use environmental data to understand health impacts?





# How can we use environmental data to understand health impacts?

- The environmental data being used must be carefully specified and appropriate for a given study



# How can we use environmental data to understand health impacts?

- The environmental data being used must be carefully specified and appropriate for a given study
- Geographically identifying cohort information must be carefully protected



# How can we use environmental data to understand health impacts?

- The environmental data being used must be carefully specified and appropriate for a given study
- Geographically identifying cohort information must be carefully protected
- Confounding variables must be carefully considered (correlates with both dependent and independent variables)



UNIVERSITY OF  
**LEICESTER**

# How can we use environmental data to understand health impacts?

So the answer is: very carefully!



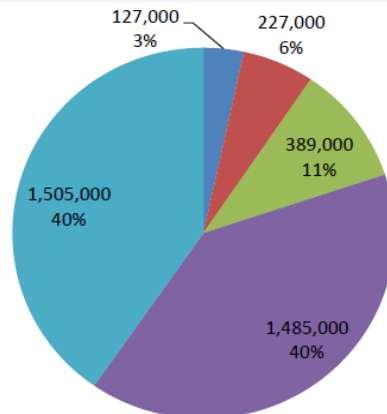
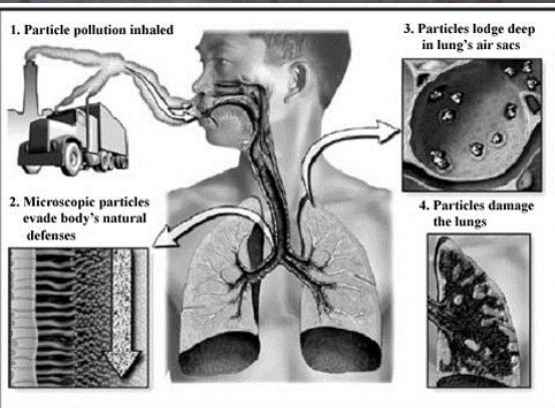
Enough introductory material—

An RDA type challenge?

How can we improve and/or expand the use of environmental data for health research while promoting standardization and best practice?

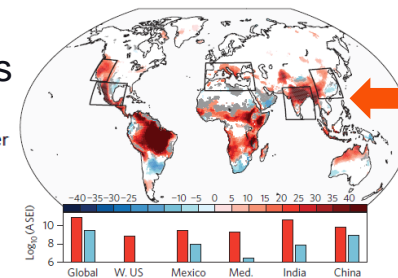
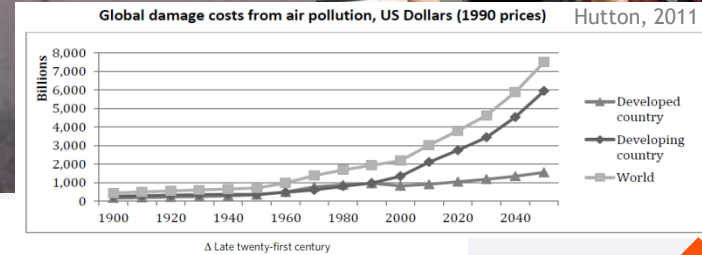
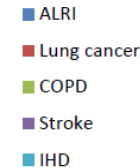
# Example: Aerosols and Health

**£9-19 billion estimated annual economic cost of air pollution in the UK (HoCEAC).**



AAP: Ambient air pollution; ALRI: Acute lower respiratory disease; COPD: Chronic obstructive pulmonary disease; IHD: Ischaemic heart disease.

**WHO:**  
AAP caused  
3.7 million deaths  
in 2012



Projected increase in annual stagnant days  
Horton et al, 2014



DEFRA: The particulate pollution burden in the UK was estimated to be equivalent to **29,000 deaths** and **340,000 life years lost** in **2008 alone**.

China is investing **>£90 billion** to reduce PM2.5 levels in Beijing by 25% **by 2017**.



UNIVERSITY OF  
**LEICESTER**

# Knowledge Exchange Activities: Aerosols & Health

Working with EU Met  
Services and Industry  
to Standardize Low  
Cost Sensor Networks

Linking Data Users  
with Data Providers

Connecting Health  
Researchers with  
Sensor and Data  
Analysis Providers

Scoping Market  
Potential of New  
Solutions

Developing New Infrastructure  
and Methodologies for  
Integration and Interrogation of  
Environmental + Health Data

Exploring  
Business Models for  
Novel Services

Creating Tools for  
Public Health  
Management of  
Exposure

Investigating New  
Applications for  
Sensor Technologies

Promoting Uptake of  
Exposure Mitigation  
Technologies with  
Industrial Partners

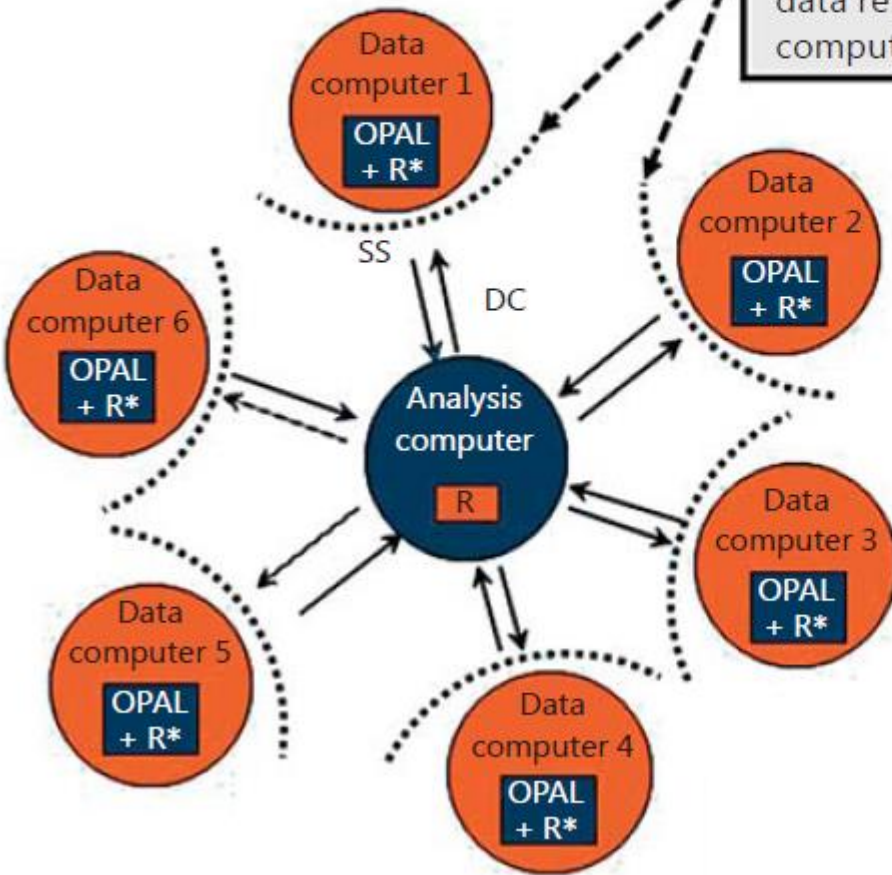




# Cohort challenge: DataSHIELD



Only nonidentifying summary statistics (SS) and DataSHIELD commands (DC) allowed passing between computers. Individual-level data retained on the local data computer



- Paul Burton (Bristol), ALSPAC “Children of the 90s” cohort co-lead, and D2K collaborators have developed a tool called DataSHIELD which allows secure query of multiple cohorts without the need for sharing of data

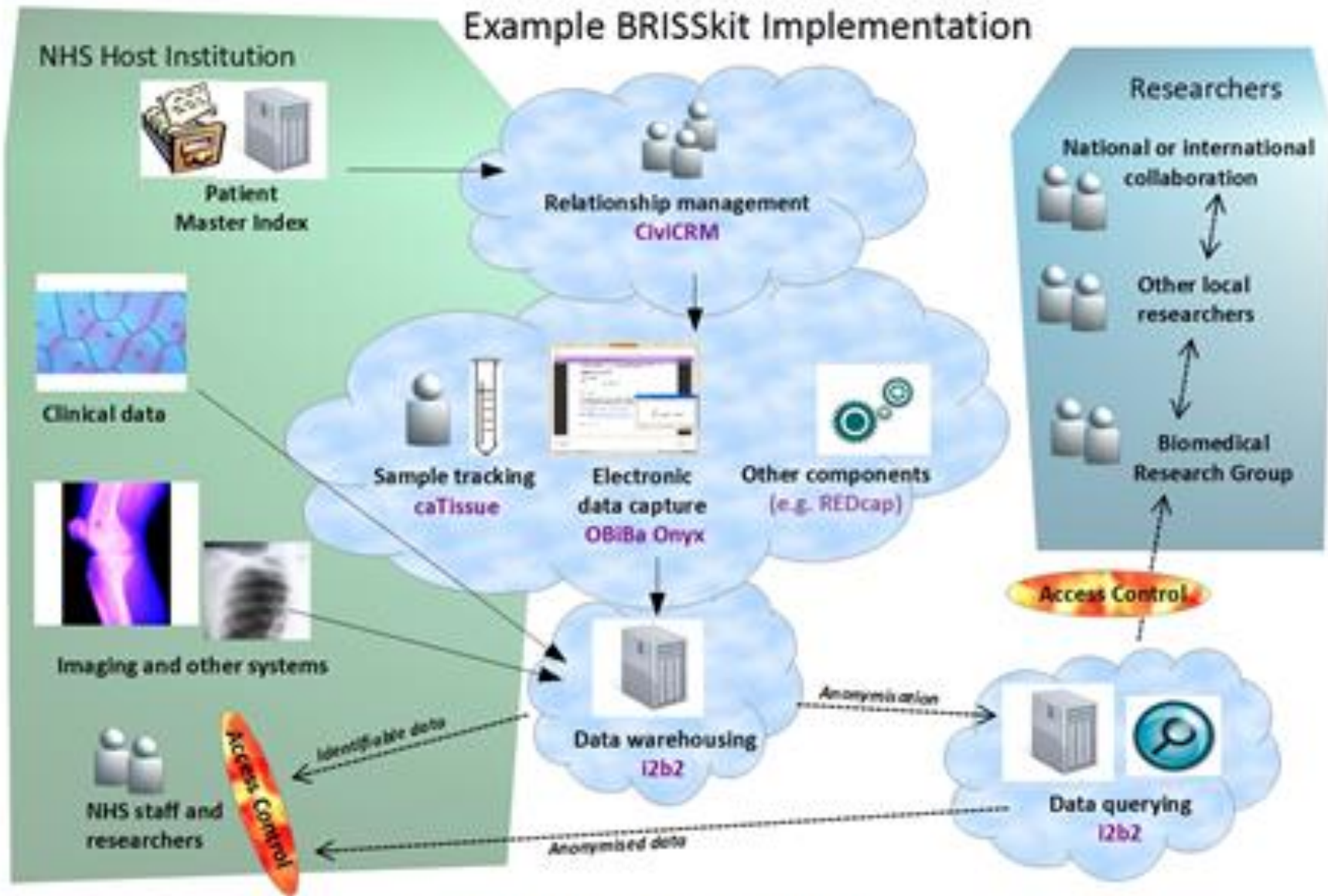


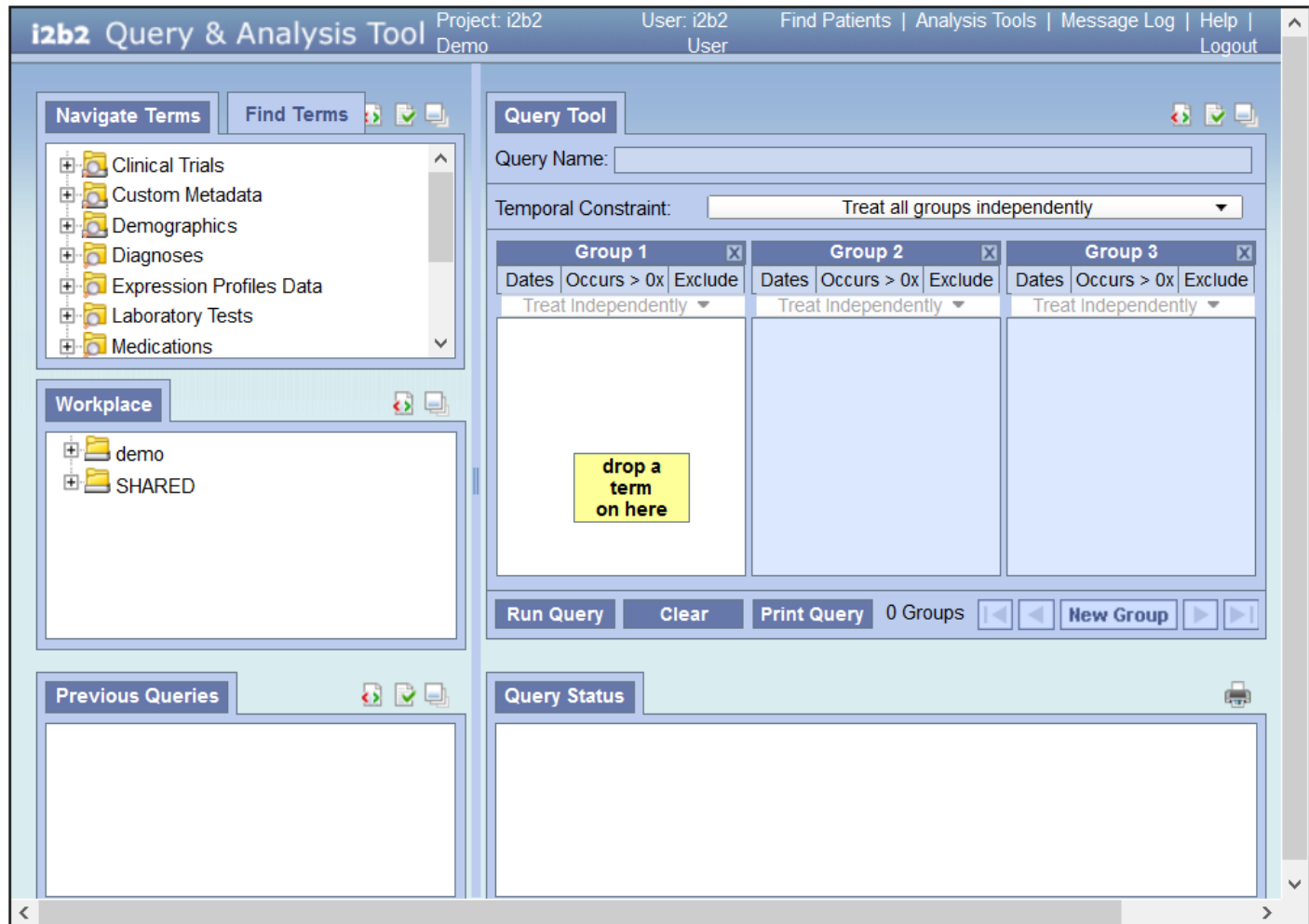


## Environmental DataSHIELD Approach

We are working to extend the DataSHIELD statistical approach concept

# Back to health informatics!



[Upload XLS - New Project](#)[Upload XLS - Existing Project](#)[Ontology Mapper](#)[Delete Project](#)[View Data](#)[Your i2b2 instance](#)

**i2b2 Query & Analysis Tool** Project: i2b2 Demo User: i2b2 User Find Patients | Analysis Tools | Message Log | Help | Logout

**Navigate Terms** Find Terms

- Clinical Trials
- Custom Metadata
- Demographics
- Diagnoses
- Expression Profiles Data
- Laboratory Tests
- Medications

**Workplace**

- demo
- SHARED

**Previous Queries**

**Query Tool**

Query Name:

Temporal Constraint:

Group 1			Group 2			Group 3		
Dates	Occurs > 0x	Exclude	Dates	Occurs > 0x	Exclude	Dates	Occurs > 0x	Exclude
Treat Independently			Treat Independently			Treat Independently		

drop a term on here

Run Query Clear Print Query 0 Groups New Group

**Query Status**

[open i2b2 in new window](#)<http://www.brisskit.le.ac.uk>



UNIVERSITY OF  
**LEICESTER**

# Introducing Environmental Parameters Into Health Cohort Query Tools



# BRISKit / i2b2 Data Query

Navigate Terms

Find Terms



- NO2\_2008
- NO2\_2012
- PM\_25\_2008
- PM\_25\_2012
- Race
- SMOKED\_AGE\_STARTED
- SMOKED\_EVER

Workplace



demo

ExportXLS



Specify Data

View Results

Plugin Help

Settings

Click the one of the buttons on the right to download the following table in the appropriate format.

CSV Export

HTML/XLS Export

Patient Information

for Patient Set 'BEE-PM\_-NO2-PM\_@11:53:10 [2-24-2015] [demo] [PATIENTSET\_3]'

	Patient ID	DIASTOLIC	NO2_2008	NO2_2012	PM_25_2008	PM_25_2012
1	5	diastolic = 79.00000	NO2exposure08 = 2.00000	NO2exposure12 = 6.00000	PM25exposure08 = 18.00000	PM25exposure12 = 16.00000



# BRISKit / i2b2 Data Query

Navigate TermsFind Terms

- NO2\_2008
- NO2\_2012
- PM\_25\_2008
- PM\_25\_2012
- Race
- SMOKED\_AGE\_STARTED
- SMOKED\_EVER

Workplace

- demo

ExportXLS

Specify DataView ResultsPlugin HelpSettings

Click the one of the buttons on the right to download the following table in the appropriate format.

CSV ExportHTML/XLS Export

Patient Information  
for Patient Set 'BEE-PM\_-NO2-PM\_@11:53:10 [2-24-2015] [demo] [PATIENTSET\_3]'

	Patient ID	DIASTOLIC	NO2_2008	NO2_2012	PM_25_2008	PM_25_2012
1	5	diastolic = 79.00000	NO2exposure08 = 2.00000	NO2exposure12 = 6.00000	PM25exposure08 = 18.00000	PM25exposure12 = 16.00000

## Example Query Parameters

Query Name: No Query Name is currently provided.  
Temporal Constraint: Treat all groups independently

### Group 1

Date From: none Date To: none Excluded? false Occurs X times: > 0 Relevance %: 100 Temporal Constraint: Treat Independently

Path	Concept/Term	Other Information
\aq2\BEER_PER_WEEK\	BEER_PER_WEEK	GT : 5

### Group 2

Date From: none Date To: none Excluded? false Occurs X times: > 0 Relevance %: 100 Temporal Constraint: Treat Independently

Path	Concept/Term	Other Information
\aq2\PM_25_2008\	PM_25_2008	GT : 12 ugm3



UNIVERSITY OF  
**LEICESTER**

# Developing Architecture for Secure and Sensible Interrogation of Linked Data







...so we can leverage and integrate existing informatics approaches (as far as older funding paradigms allow?)....

BUT actually the big problem is how can the environmental data be meaningfully linked?



# How can the environmental data be linked?

Let's look at a simple example where postcode information is available in the health cohort.



## HEALTH COHORT 1 DATA

ID	Health Variable	Postcode
1.1		
1.2		
1.3		



## HEALTH COHORT 1 DATA

ID	Health Variable	Postcode
1.1		
1.2		
1.3		

## HEALTH COHORT 2 DATA

ID	Health Variable	Postcode
2.1		
2.2		
2.3		
2.4		

## Environmental Data Kept Somewhere Else

## HEALTH COHORT 1 DATA

ID	Health Variable	Postcode
1.1		
1.2		
1.3		

## HEALTH COHORT 2 DATA

ID	Health Variable	Postcode
2.1		
2.2		
2.3		
2.4		

## ENVIRONMENTAL DATASET

### FULL EXTENT OF STUDY REGION

[illegible]

## Full environmental dataset imported into cohort databases

## HEALTH COHORT 1 DATA

ID	Health Variable	Postcode
1.1		
1.2		
1.3		

## HEALTH COHORT 2 DATA

ID	Health Variable	Postcode
2.1		
2.2		
2.3		
2.4		

## ENVIRONMENTAL DATASET

[illegible]

Full environmental  
dataset imported  
into cohort databases

### HEALTH COHORT 1 DATA

ID	Health Variable	Postcode
1.1		
1.2		
1.3		

### HEALTH COHORT 2 DATA

ID	Health Variable	Postcode
2.1		
2.2		
2.3		
2.4		

### ENVIRONMENTAL DATASET

Postcode	Environmental Variable



Data is sorted by postcode and merged

### HEALTH COHORT 1 DATA

ID	Health Variable	Postcode	Environmental Variable
1.1			
1.2			
1.3			

### HEALTH COHORT 2 DATA

ID	Health Variable	Postcode	Environmental Variable
2.1			
2.2			
2.3			
2.4			





UNIVERSITY OF  
**LEICESTER**

Okay, now unleash the informatics?



UNIVERSITY OF  
**LEICESTER**

# Okay, now unleash the informatics?

- Almost...



# Pre-analysis harmonization / subsetting

- 1) Use metadata to find spatial and temporal resolution of environmental data

## Pre-analysis harmonization / subsetting

- 1) Use metadata to find spatial and temporal resolution of environmental data
- 2) Query health study locational information in health cohort metadata (resolution, accuracy, etc.)



## Pre-analysis harmonization / subsetting

- 1) Use metadata to find spatial and temporal resolution of environmental data
- 2) Query health study locational information in health cohort metadata (resolution, accuracy, etc.)
- 3) Check that an uncertainty can be assigned to the environmental exposure metric

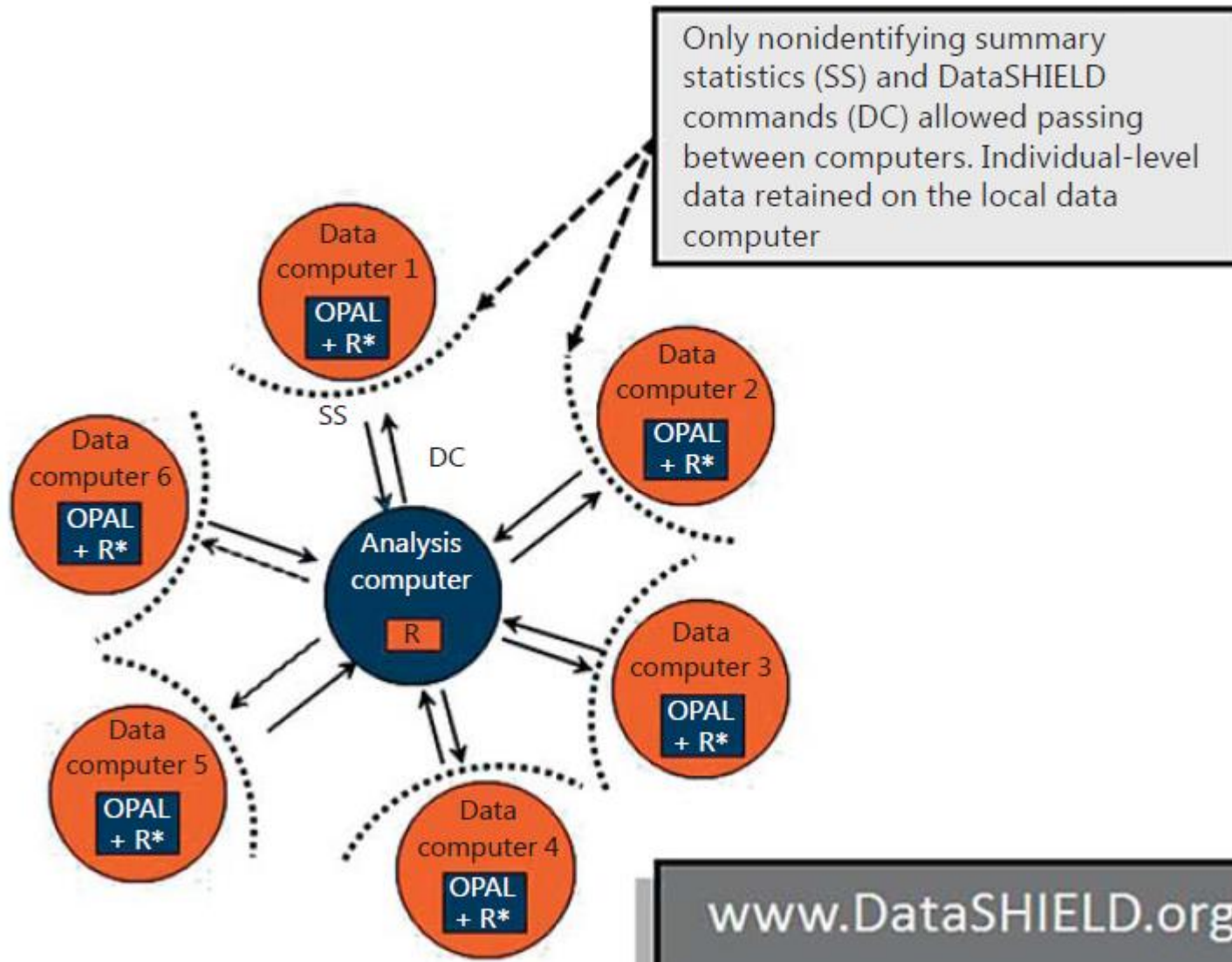


## Pre-analysis harmonization / subsetting

- 1) Use metadata to find spatial and temporal resolution of environmental data
- 2) Query health study locational information in health cohort metadata (resolution, accuracy, etc.)
- 3) Check that an uncertainty can be assigned to the environmental exposure metric
- 4) Filter data into sub-cohorts if necessary—for now data centre would have to do this.



# Once the data has been filtered/harmonized



Then it is possible to run a GLM on all cohorts to explore relationships across all cohorts in combination



UNIVERSITY OF  
**LEICESTER**

# We need an Environmental Data Dictionary





# What is a Data Dictionary?

In a health informatics framework, the data dictionary is where health-related variables are defined succinctly for

- data collectors
- data managers
- data users



FOM1 Files		Clinic : Focus on Mothers 1	Ran: Dec 2008 to July 2011	
Filename	Var name	Var Label		
FOM1	mult_mum	FOM1: Entry is a duplicate - remove if only looking at Mothers; i.e. not matched to children		
FOM1	fm1a010a	Month of attendance: FOM1		
FOM1	fm1a010b	Year of attendance: FOM1		
FOM1	fm1a011	Age at attendance (years): FOM1		
FOM1	fm1a015	DV: Source of data collection: FOM1		
FOM1	fm1ms001	Anthropometry fieldworker: FOM1		
FOM1	fm1ms002	Anthropometry room: FOM1		
FOM1	fm1ms100	Height (cm): FOM1		
FOM1	fm1ms101	Sitting Height (cm): FOM1		
FOM1	fm1ms103	DV: Leg length (cm): FOM1		
FOM1	fm1ms105	Pacemaker fitted: FOM1		
FOM1	fm1ms110	Weight (kg): FOM1		
FOM1	fm1ms111	DV: BMI: FOM1		
FOM1	fm1ms115a	Waist Circumference, 1st measure (cm): FOM1		
FOM1	fm1ms115b	Waist Circumference, 2nd measure (cm): FOM1		
FOM1	fm1ms115	DV: Waist Circumference, mean (cm): FOM1		
FOM1	fm1ms120a	Hip Circumference, 1st measure (cm): FOM1		
FOM1	fm1ms120b	Hip Circumference, 2nd measure (cm): FOM1		
FOM1	fm1ms120	DV: Hip Circumference, mean (cm): FOM1		
FOM1	fm1ms125	Arm Circumference (cm): FOM1		
FOM1	fm1ms126a	Head Circumference (cm): FOM1		
FOM1	fm1ms126b	Lasso Head Circumference (cm): FOM1		
FOM1	fm1ms126	DV: Combined Head Circumference (cm): FOM1		
FOM1	fm1dx001	Consent given for DXA scan: FOM1		
FOM1	fm1dx002	Consent given to be informed if low BMD on DXA : FOM1		



## Father biological samples File

Filename	Var name	Var Label
F_Sam	Hb_FOF1	Haemoglobin, fasting FOF1
F_Sam	chol_FOF1	Cholesterol mmol/l, fasting FOF1
F_Sam	trig_FOF1	Triglycerides mmol/l, fasting FOF1
F_Sam	hdl_FOF1	HDL cholesterol mmol/l, fasting FOF1
F_Sam	crp_FOF1	C-Reactive Protein mg/l, fasting FOF1
F_Sam	glucose_FOF1	Glucose mmol/l, fasting FOF1
F_Sam	vldl_FOF1	vLDL cholesterol mmol/l, fasting FOF1
F_Sam	ldl_FOF1	LDL cholesterol mmol/l, fasting FOF1



A Files		Questionnaire: Your Environment			
Completed by: Mother		At: 8 weeks gest			
Filename	Var name	Var Label			
A	a001	Questionnaire version			
A	a002	Time lived in Avon			
A	a003	Years since last move			
A	a004	Weeks since last move			
A	a005	NO of moves in 5 YRS			
A	a006	Home ownership status			
A	a007	Whose home mum lives in			
A	a008	Dwelling type			
A	a009	Lowest level of ACCOM			
A	a010	Winter temp of living rooms			
A	a011	Winter temp of bedrooms			
A	a012	Central or storage heating			
A	a013	Wood stoves or fires			
A	a014	Coal fires			
A	a015	Paraffin heating			
A	a016	Mains gas fires			
A	a017	Calor gas fires			
A	a018	Other heating			
A	a019	Central heating fuel			
A	a020	Central heating type			
A	a021	Situation of boiler			
A	a024	Hot water bottle use during PREG			
A	a025a	ELEC U blanket use during PREG			
A	a025b	ELEC O blanket use during PREG			
A	a025c	Owns ELEC blanket			
A	a025d	Age of ELEC blanket			
A	a025e	In bed with ELEC blanket on in winter			
A	a025f	In bed with ELEC blanket on in summer			
A	a025g	In bed with ELEC blanket on during PREG			



## Environmental Data Dictionary – RDA role?

**Objective:** To develop formats and demonstrators for environmental data dictionary entries that could give non-experts in these variables the ability to determine whether they could be used sensibly as secondary variables in a health study, and if so use them.

**Note:** Goal is not to remove environmental scientists from the loop, but to encourage a unified approach to identifying how different variables can be used.



## Example Environmental Meta Data Entries for the Data Dictionary

- Variable name and brief description in top layer,  
other meta data in second layer
- PM2.5 particulates annual map
- Personal sensor data



UNIVERSITY OF  
**LEICESTER**

# Entry in progress:

## PM2.5 Defra Background Map

Variable\_Name: PM025\_DBM\_2011

Variable\_Description: mass concentration of surface-level atmospheric particulate matter  $\leq 2.5$  micron diameter (PM2.5)

Units: micrograms per cubic metre at selected outdoor location

Variable\_Definition: modelled outdoor PM2.5 annual average mass concentration data derived from emission inventories, meteorological data, and an observational PM2.5 network, emission source type speciation is available

Data\_Source: DEFRA

Spatial\_Extent: UK

Spatial\_Resolution: 1 km by 1 km

Coordinates: OSGB1936 coordinates represent centrepoinets of grid squares

Temporal\_Extent: year of product specified in last 4 digits of variable name

Temporal\_Resolution: annual average for year specified in last 4 digits of variable name

Example\_Uses: can be used as an indicator of average outdoor fine particulate air pollution at residential or work address, might be useful for long term exposure studies where a strong health response signal is anticipated



UNIVERSITY OF  
**LEICESTER**

# Entry in progress:

## Mobile Personal Sensor NO2

Variable\_Name: NO2\_MPS1

Variable\_Description: volume concentration of nitrogen dioxide measured from personal sensor

Units: parts per billion

Static or mobile: mobile

Variable\_Definition: nitrogen dioxide concentrations are measured by drawing air across metal oxide sensors, corrections for interfering gases are applied

Data\_Source: university of leicester zephyr sensor

Spatial\_Extent: data exists where sampling has taken place

Spatial\_Resolution: in situ point measurement

Temporal\_Extent: data exists where sampling has taken place

Temporal\_Resolution: 1 minute

Example\_Uses: output can be used as a measure of personal exposure, but uncertainties due to sampling issues driven by position and movement should be considered. output can be converted to mass concentration for reference to air quality standards





# Summary

- Environmental health is a key global concern
- Researching it requires a multi discipline approach using informatics
- Even where environmental or health datasets might be standardised - are they interoperable?
- Perhaps RDA can play a role by helping enable registry level coordination on the where, what, how collected questions and associated training?
- Where can we add most value at RDA level?
- <https://www.jiscmail.ac.uk/lists/ENVIROHEALTH-INFORMATICS>



# Acknowledgments

- Team BRISKit - <http://www.brisskit.le.ac.uk>
- Roland Leigh and Paul Monks and their atmospheric science teams, and colleagues in Earth Observation
- Paul Burton, Rebecca Wilson, and the D2K team at University of Bristol
- Health professionals who have taken the time to offer input