MEDMI Advisory Board: Summary Report November 2016 (www.data-mashup.org.uk)

What is MEDMI?

The **MED-MI** (**M**edical and **E**nvironmental **D**ata - a **M**ash-up **I**nfrastructure) Project was funded in 2013 by the UK Medical Research Council (MRC) and UK Natural Environment Research Council (NERC) as a Partnership Grant to explore the interconnections between the environment and human health using "big data." The MEDMI Project is overall a **Demonstration/Feasibility Project**.

The MEDMI Project attempts to combine and make available meteorologic, atmospheric, environmental, and human health and wellbeing data from the **MEDMI Partner Institutions** and other collaborators. These linked data are made available on the **MEDMI Platform** which allows for data storage, analysis, visualization, and exploration while maintaining data confidentiality. The MEDMI Platform has been established, explored and illustrated through **3 MEDMI Demonstration Projects** as well as other collaborative **Pilot Projects**. There have also been **MEDMI Engagement and Dissemination Events** to seek new collaborations, learn from the experience of other projects, and disseminate the experience and results of the MEDMI Project.

Project Oversight was provided by the **MEDMI Advisory Board** throughout the 33 months of funding; with a 6 month no cost extension, the MEDMI Project finished at the end of November 2016. As of that time, there are ongoing discussions between two of the MEDMI Partners, the UK Met Office and Public Health England (PHE), regarding the extension of MEDMI to continue to make the databases and other resources available to researchers.

Who are the MEDMI Partners?

- The **European Centre for Environment and Human Health** (www.ecehh.org) of the University of Exeter Medical School is the lead institution. The European Centre is focused on interdisciplinary research and training around the interconnections between the environment and human health. The University of Exeter (www.exeter.ac.uk) is a Russell Group University with an internationally recognized expertise in climate change and sustainability.
- The **UK Met Office** (www.metoffice.gov.uk) provides the weather and climate change forecasts for the UK and worldwide for the public, business, and government. The Met Office has climate, weather, and other environmental data and research expertise. The Met Office is interested interacting with MEDMI long term.
- **Public Health England (PHE)** (www.gov.uk/government/organisations/public-health-england) has the mission is to protect and improve the nation's health and to address inequalities. PHE has health and wellbeing data and research expertise. In addition, the PHE **Environmental Public Health Surveillance System (EPHSS)** (aka Environmental Tracking) is interested interacting with MEDMI long term.
- The **London School of Hygiene and Tropical Medicine (LSHTM)** (www.lshtm.ac.uk) an internationally renowned school of public and global health, working closely with partners in the UK

and worldwide to address contemporary and future critical health challenges. The LSHTM has expertise in large dataset analyses and international public health research and training expertise.

- The **University of Bristol** (www.bris.ac.uk) is internationally renowned, ranked in the top 30 universities globally, with expertise in environment and public health research and training.

What is MEDMI Project aiming to do?

The MED-MI Project Aim is to facilitate user-specific analyses of the relationships between subsets of geographically and temporally indexed environmental and health datasets, known as "**data mash-ups**" in computer science. The MEDMI servers host the raw and modelled data, and in some cases linked datasets, plus any added data as the result of future collaborations (with access subject to confidentiality and ethical safeguards); the datasets are downloadable to personal PCs or analysed on the Platform. The Platform also hosts an interactive web-based interface to allow both local and remote access to user-selected subsets of the data; and it provides a library of tools (e.g. visualization, mapping, commentary, notation) to facilitate access to the data, and in the future, to provide standard analysis and modelling (e.g. "mash-up") options.

Below are the more specific objectives along with an update of progress to date:

1. **Establish the database management system and MED-MI Platform housed on the University of Exeter IT System to manage and facilitate storage and access;**

Two servers were obtained in June 2014. They are based at the University of Exeter. In December 2014, environmental data from the Met Office were loaded onto the servers (including MIDAS, NCIC, AURN, and stratospheric ozone) and since then Dr Christophe Sarran (Met Office) has been working to organize and process these data using python coding to make the data callable as well as easily available for lagged statistical analyses.

It has been much harder to obtain health data. SGSS (formerly LABBASE) data on over 2000 pathogens (associated with infectious diseases and reported to PHE) were obtained from PHE in April 2015. These data were then geo-coded (adapted from postcode to latitude and longitude) so that it could be linked with environmental data; and added to the MEDMI server in May 2015. Recently restricted SGSS data containing confidential patient location as well as the diagnosing laboratory were added for PHE sensitivity analyses only. Other applications for health data including:

- ONS Mortality Data – mortality data has been received for 5 approved researchers and the data obtained to be installed on MEDMI with limited access.
- UK Biobank data on UV, Vitamin D and Osteoporosis was obtained for one of the MEDMI pilot projects.
- HeS (hospital episodes statistics) data – application approved – currently not available for MEDMI websites or server.
- TPP (private GP database) – Dr Shakoor Hajat has these data for research which have been linked with temperature data by TPP – currently not available for MEDMI websites or server.

There are also data that we have permission to use which have been added to the MEDMI servers, including pollen data from the Met Office and Meniere's data from Dr Jess Tyrrell. Some of the other pilot projects have generated additional data that has also been added. For example, the pilot project "Statistical Downscaling of Gridded Air Quality Data" by Prof Sujit Sahu has produced estimates of air

quality for four of the most harmful pollutants in England and Wales for the five year period 2007-2011. These data are available on the MEDMI website.

As it was not possible to replace the Database Research Scientist after he left in August 2014 (due to challenges recruiting someone with the required diverse and highly technical dataset), the role was divided up to cover several different areas of work including finishing the statistical browser (Ceri Whitmore), creating the visualisation tool (Neil Kaye), and developing a database module for the MEDMI data (Dr Christophe Sarran), as well as the Pilot Projects (described below). Lastly, we have worked with PHE's Environmental Tracking Group and Software Development Unit to develop a browser interface to the data held on the MEDMI servers. The interface will be accessible through PHE's Environmental Public Health Surveillance System (EPHSS) website. Information about this will also be provided on the MEDMI website, with a link to the interface. The interface is due to go live in February 2017, following comprehensive testing.

2. Perform the data "mash-up" by linking datasets in a consistent temporal and spatial framework and with appropriate quality control and confidentiality, making climate, weather and environment data available for use on the MED-MI Platform;

Five different models for data mash-up have emerged under MEDMI and are demonstrated on the MEDMI Platform:

1. Public web application browser – this sits interactively on the MEDMI website, demonstrating some of the work in Demonstration Project 1, but without access to any confidential or proprietary data. This is open access. A second tool has been developed to visualize and analyse data from Demonstration Project 2.
2. Secure web application browser - a second secure URL has been established which links to the web-browser application. This is only accessible to MEDMI researchers (and other approved researchers primarily from the Partner Institutions). Access, after establishment of credentials with the PI, is controlled through a secure log-in via the main MEDMI website. This enables MEDMI to control who has access to the MEDMI data, in response to the issues of licensing and confidentiality of data.
3. Web interface to extract data from the MEDMI server – this will be located on the EPHSS website (with a link from the MEDMI website). Users will be able to select datasets through a series of dropdown menus and then submit a request for data. The browser will extract data by linking to a database module that has been developed by Christophe Sarran at the Met Office. This will make MEDMI data more available to approved researchers who do not have the programming skills to access the raw data directly from the MEDMI servers. The development of the interface is complete and is now undergoing comprehensive testing and will go live in February 2017.
4. Database module and raw data stored on the MEDMI server - it is possible to extract data and link the MEDMI data with other data by logging into the MEDMI server and using the database module. Training on this module has been provided by the Met Office in November 2016 to interested prospective users (currently about 30 people from MEDMI partners, as well as some public health analysts); additional classes are being made available due to interest in 2017. This method of access suits those with programming skills who want access to raw, unfiltered data or wish to use the temporal and special processing tools that form part of the database module. These processes are useful for linking datasets with each other and for producing time series. There are also some visualization tools attached to the database module.

5. Data that are held by various MEDMI researchers and have been used for MEDMI Demonstration Project, but for licensing or confidentiality reasons the data are not available to go on MEDMI website or server (i.e. TPP data – Shakoor Hajat at LSHTM; some SGSS data being used by PHE researchers). In the future, access to anonymised sets as "tasters" on the MEDMI Platform might be possible.
6. Facilitate linkages of other health and environmental datasets –For example, the work of MEDMI has led to the possibility of Met Office data being incorporated into UK Biobank. It is too complex to extract data from Biobank to undertake linkages on MEDMI; UK Biobank takes in other datasets and performs the linkages, then returns linked data to researchers, and vice versa (i.e. with osteoporosis pilot project). A joint Workshop between MEDMI and Biobank took place in November 2015 at the Wellcome Trust on increasing the possibilities of environment and health linkages for research.

3. Explore the feasibility and vision of the Platform by performing novel Partnership Demonstration Projects using the linked MED-MI databases;

As proof of concept and as part of the actual MEDMI Platform creation process, MEDMI has demonstrated the feasibility of linking diverse datasets and the utility of this approach for population health using interdisciplinary, hypothesis-driven Demonstration Projects as well as collaborative Pilot Projects.

Based on Advisory Board recommendations, we have also tried a "storyboarding" approach when interacting with the researchers around both the databases and the analyses; this storyboarding was used to start to set up the MEDMI Platform, particularly the browser analysis interface in Demonstration Project 1.

Demonstration Project 1 - Extreme temperatures, air quality, and mortality (Hajat, Sarran, Whitmore)
Demonstration Project 1 is based on an existing published time series analysis of Dr Hajat on mortality, temperature and some other environmental factors in a short time period in London. The project has been a "proof of concept" to show that it is possible to replicate analysis undertaken off-line using a statistical package (in this instance Stata) using a web-browser application created by the researchers. Hajat and Whitmore have developed the web-browser using a story-boarding approach. The browser has been developed using web technologies including a JavaScript framework, for the interactive interface, which calls python serverside code for running the time series analysis code. The browser has been developed and successfully used to reproduce the results from the initial analysis. Hajat and Whitmore have added a second data-set (i.e. pollen and hospital admission data (HES)) to demonstrate that the browser can be adapted for different datasets; due to confidentiality/ownership issues, the latter is only available on the secure part of the MEDMI Platform. (Publication in Science of the Total Environment "Development of a browser application to foster research on linking climate and health datasets: challenges and opportunities", S. Hajat, et al in press).

Demonstration Project 2 - Climate, weather, and infectious diseases (Cherrie, Cichowska, Djennad, Kaye, Kessel, Lo Iaconno, Nichols, Sarran, Whitmore). As part of a Pilot Project, Professor Nichols and Dr Cherrie performed regression trawling exercise of pathogens to look at their seasonality using the SGSS data; this has identified >40 pathogens that have seasonal variation. The SGSS data were linked

with a wide range of environmental factors and these form the basis of the new visualization browser developed by Neil Kaye (Met Office) available on the secure part of the MEDMI website

The project is also undertaking a sensitivity analysis to see if the laboratory location is a sufficient substitute for residency location with regards to the environmental data. Obtaining residency data for many pathogens is difficult due to privacy/confidentiality, as there is a risk of identifying individuals if residency is used for some rare diseases, especially in less densely populated postcodes. If the laboratory can be used as a substitute to residency, it will enable more detailed analysis as laboratory data are available for a much longer time period and the issue of confidentiality is overcome. A manuscript has been submitted for publication.

In collaboration with the National Institute of Health Research (NIHR) funded Health Protection Research Unit (HPRU) in Environmental Change and Human Health, Drs Djennad, Nichols and Lo Iaconno have expanded an earlier paper by Professor Nichols and Dr Sarran on campylobacter and temperature/season. They have undertaken time-series analysis, using 3 different models of time-series analysis for a 5 year period as well as wavelet analyses. A manuscript has been submitted for publication. Dr Lo Iaconno has lead a systematic review of the methods used to analyse waterborne infectious diseases in relation to climate and other environmental change. A manuscript has been submitted for publication.

Demonstration Project 3 - Climate, coastal ocean dynamics, harmful algal blooms (HABS), and human health "blue sky" initiative (Barciela, Fleming, Sarran, Nichols, Taylor, Davidson) A NERC-funded CASE PhD was obtained to support this work. Professor Fleming (Academic lead), Dr Barciela (Met Office Lead), Professor Nichols (PHE Lead), Professor Davidson (SAMS) and Dr Richard Sharpe (Cornwall Public Health/UEMS) as the Supervisory Team. Ms Lucy Lintott MSc started March 2016.

Professor Davidson from Scottish Association for Marine Science (SAMS), which has Scottish HAB data, is involved in supervising the PhD student, as well as providing access to the HABs data. The Met Office has signed a MOU with CEFAS to obtain the HABs data for England and Wales. With the student, Dr Barciela is expanding her oceanographic analyses and will validate them with the HABs data; Ms Lintott will use HeS and other data to examine connections between the HABs and human health events with Professors Fleming and Nichols. Ms Lintott will also produce a scoping literature review of climate change, HABs and human health with Dr Sharpe and the other Supervisors. This project is now also listed as part of the NIHR funded HPRU in Environmental Change and Human Health (Project 3.7).

4. **Establish an initial website, and later, a portal for use by outside entities and individuals (including medical and public health researchers, planners and policy makers) inside and outside of the UK host to allow both local and remote access to user-selected subsets of the data as well as a library of tools (e.g. visualization, mapping, commentary, notation) to facilitate access;**

There are 2 servers in place (one for the MEDMI Platform and one for the Website, the latter leveraging funds from the European Regional Development Fund [ERDF]) with the backup housed at the Streatham Campus with University of Exeter IT support. A MEDMI Website was commissioned from a local web development company, fffunction, in Cornwall. The website went live in March 2016 www.data-mashup.org.uk.

The website is the gateway to several tools that have been developed by the project. A web browser application has been developed by Mr Ceri Whitmore and Dr Shakoor Hajat for Demonstration Project 1.

One version is available on the public website. A duplicate of the web-browser with additional data is also accessible through a secure log-in on the MEDMI website. A second visualization tool has been developed for Demonstration Project 2, which is accessible through the MEDMI website.

Lastly, a web interface has been developed through collaboration with PHE's Environmental Tracking group. This will enable users to select and extract data held on the MEDMI server. This tool will be available through PHE's EPHSS website, but there will be a link from the MEDMI website when it goes live later this year.

5. Strengthen and explore collaborations (including new pilot projects) with outside entities;
6 Pilot Projects of approximately £5K/project were awarded funding in April 2015, with an additional Pilot Project added in February 2016 to Professor Sabina Leonelli. These projects are all completed. A more detailed **Summary of Projects** is attached.

a) **Weather and Symptom Fluctuations in in Meniere's Disease** - Dr Tyrrell, Dr Schmidt (ECEHH), Dr Sarran (Met Office) – completed and paper accepted for publication.

b) **Linkage Tools for Pre-Processing of Pollen data** - Dr McInnes and Dr Sarran (Met Office) – pilot project work complete, additional work is required to finalise all the pre-processing tools.

c) **Osteoporosis & Solar Irradiance using Biobank** - Dr Cherrie, Dr Osborne (ECEHH) and Dr Sarran (Met Office) – Data obtained from UK Biobank. Paper being finalized.

d) **Childhood Obesity and Neighbourhood Environments: Integrating SAIL and MEDMI** - Dr Sarah Rodgers, Mr Fry (SAIL, Swansea University), Dr Ben Wheeler (ECEHH) – finished, no paper but useful feedback has led to new web interface project to increase MEDMI data accessibility for researchers. The collaboration between ECEHH and SAIL has led to the development of a new project proposal to NIHR on "green-blue space exposure and individual-level wellbeing". It has been successful at the first round of the process. A full application will be submitted in December 2016.

e) **Statistical Downscaling of Gridded Air quality data** – Professor Sahu (Southampton University). The data generated by this project is available on the MEDMI website. This collaboration has led to further funding from the ESCRC for the Met Office and Southampton University's work on air quality data.

f) **Pathogen reasonability – regression trawling** – Professor Nichols (PHE), Dr Mark Cherrie (ECEHH), and Neil Kaye (Met Office) –Paper is being finalized. Visualisation tool is available on the MEDMI website.

g) **Tracing Data Journeys Across Climate, Environment and Human Health: A Qualitative Study of the Medical & Environmental Data Mash-Up Infrastructure Project –**Professor Sabina Leonelli, Dr Niccolo Tempini (Egenis, University of Exeter) to document MEDMI project development and outcomes using MEDMI as a case study for Professor Leonelli's ERC funded DATA_SCIENCE project – ongoing interviews with report and publication expected to be submitted in Nov 2016.

A **Workshop with UK Biobank** took place in November 2015 at the Wellcome Institute to look at the challenges and opportunities of linking environmental data with large health datasets, particularly UK Biobank. This Workshop included researchers who have experience of linking environmental data with UK Biobank, some of who are from MEDMI partner institutions, as well as others (SAIL, Swansea; SAHSU, Imperial, and Dept of Psychiatry at Oxford University). The Workshop was a collaboration with the UK Biobank and with the NIHR funded Health Protection Research Unit (HPRU) in Environmental Change and Human Health. The Workshop resulted in a Letter to the UK Cabinet concerning the

importance of environment and health data and their linkages as a national resource, challenge and opportunity. The Biobank also appears to be more receptive to environment and health linkages (e.g. Dr Ben Wheeler green space data linkage; Dr Mark Cherrie solar irradiance data linkage). The recent MRC NERC call mentioned both MEDMI and this Workshop specifically http://www.nerc.ac.uk/innovation/activities/environmentaldata/health-call/.

6. **Disseminate the MED-MI Platform with appropriate MRC-approved confidentiality and ethical safeguards supported by the Strategic Partnership;**
   a) **Dissemination**

   i) Big Data Workshop took place in June 2014: Connecting Environment and Human Health at the Health and Wellbeing Innovation Centre in Truro (Cornwall). The event was attended by over 60 people from academia, government, businesses (including SMEs), 3rd sector and interested persons (www.ecehh.org/events/big-data-workshop/). Funding was leveraged from the NIHR funded Health Protection Research Unit (HPRU) in Environmental Change and Human Health as well as with leveraged funding from a NERC Impact Accelerator Pilot Grant and ERDF funding from the European Centre.

   ii) A meeting took place in London in January 2015 for MEDMI researchers and potential collaborators, including Dr Rogers (SAIL), Dr Bates (TPP), Professor Semenza (ECDC) and Professor Rees and Dr Watkins (CEH) as well as representatives from ONS, Biobank, and SGSS databases. This was an opportunity to share information about MEDMI with other institutions and researchers, and explore new collaborations and databases. Pilot project funding has been used to fund work between researchers at ECEHH and SAIL (Secure Anonymised Information Linkage Databank at Swansea University). Discussions about a potential workshop with CEH around presenting data for policy influence are ongoing and continued at the Workshop with UK Biobank in November 2015.

   iii) MEDMI has also had bi-annual **Investigator Meetings** and supported smaller informal meetings of the researchers and additional collaborators at the various Partner Institutions. There have been annual meetings of the **Project Advisory Board**.

   iv) **MEDMI Website/Platform -** The Website went live in March 2016. This is the primary resource for disseminating information about MEDMI. Details have been shared with researchers across the country. There are ongoing discussions as to how the Website will be used in the future as well as the target population(s).

   v) MEDMI Researchers have presented MEDMI related papers and posters at several conferences in 2016 to disseminate the outcomes and lessons learnt from the project. These include:

   - *The Applied Epidemiology Scientific Conference March 2016, Warwick (PHE)*
   - *International Symposium on Environment and Health (ISEH) 2016 & Geoinformatics 2016 on Environment, Health, GIS & Agriculture, Galway, Ireland, August 2016*
   - *Royal Meteorological Society Conference, Manchester, July 2016*
   - *International Population Data Linkage Conference, Swansea, August 2016*
   - *International Society for Environmental Epidemiologists, ISEE, Rome, Sept 2016*
   - *Liverpool Challenger Event, Sept 2016*
   - *International Conference on Harmful Algae (ICHA), Brazil, Oct 2016*
   - *AGU, San Francisco, Dec 2016*

   vi) **Training on the MEDMI database module at Met Office (November 2016)**

Dr Christophe Sarran organized 3 training sessions at the Met Office in November to explain the data available on MEDMI and demonstrate how the database module can be used for extracting and linking datasets. The training was attended by 27 potential users (from the Met Office, PHE, LSHTM, University of Exeter and local public health analysts from the South West). More training sessions are being scheduled for 2017 due to user demand.

vii) **Big Data, Environment and Health** (TBA 2017) UoE, PHE and the Met Office are discussing holding a meeting with all interested NIHR HPRUs about future uses of the MEDMI Database resources.

b) Confidentiality and Ethical Safeguards for disseminating data

Now that there are data available on the MEDMI servers, we have put together procedures and systems for accessing and disseminating data. This process has been supported by Professor Anthony Kessel (MEDMI Project Investigator at PHE, who is a Caldicott Guardian) with ongoing discussions on the procedures that need to be put in place for confidentiality and ethical safeguards as well as governance. This includes processes and controls for who can access data; methods for accessing data and security settings. These issues are documented on the MEDMI Website as well as caveats about data interpretation. Discussions with groups (e.g. SAIL and the Farr institutes) and with Professor Leonelli and Dr Tempini have been invaluable since they are involved in similar issues.

7. Develop and seek new collaborations and funding (including the potential commercialisation of MED-MI products) to expand and sustain the Strategic Partnership and MED-MI Platform.

In the past few months, the MEDMI group has been in discussion with the PHE Environmental Tracking Group led by Professor Giovanni Leonardi (PHE). The **PHE Environmental Tracking** Group has spent the last few years in intense discussions with potential stakeholders about how to perform surveillance, modelling, forecasting, and research on environment and human health interactions; they have been developing their own web-based tools and primarily health datasets but they lack access to environmental data. The web interface between MEDMI and EPHSS is a feasibility study for PHE, testing whether this is a workable method for accessing environmental data. Discussion is ongoing about further future collaboration between MEDMI and PHE.

Funded

- NERC IAA Pilot Funds supported the Big Data Workshop in June 2015
- NIHR HPRU in Environmental Change was funded for 5 years, led by LSHTM (Kovats PI), PHE (Vardoulakis), UoE (Fleming), Met Office (Falloon), and UCL (Davies) which leveraged the MEDMI Project for Theme 1 on Climate Change (Hajat), Theme 2 on Built Environment (Vardoulakis), and Theme 3 on Natural Environment (Fleming). The 3 Demonstration Projects "track" to the scope of the HPRU as well as other activities.
- NERC funded CASE PhD in Climate Change, HABs and Human Health with Professor Fleming (Academic lead), Rosa Dr Barciela (Met Office Lead), Gordon Professor Nichols (PHE Lead), and Nick Dr Taylor (CEFAS Lead) and Professor Davidson (SAMS) as the Supervisory Team.
- Horizon 2020 BlueHealth Project funded for 4.5 years in Jan 2016 led by UoE (Fleming PI)

Not funded

- A NEWTON-MRC Application with all MEDMI Partners with Brazil FioCruz and INPE not funded.

- Application to NERC-MRC-CSO Improving Health with Environmental Data call was not funded.
- Application to Newton Fund Researcher Links Workshop Grant 2016 for a 3 day workshop on "Big Data Mashups in Environment and Human Health" in Pretoria, South Africa not funded but asked to resubmit in 2017.

Publications

	Paper	Lead Authors	Status
1	**"Development of a browser application to foster research on linking climate and health datasets: challenges & opportunities"**	S. Hajat, C. Whitmore, L. Fleming, C. Sarran; A. Haines; B. Golding; H.Gordon-Brown; A. Kessel	ACCEPTED "Science of the Total Environment" on the 30/8/16.
2	**"The Weather and Meniere's Disease: a longitudinal analysis in the UK"**	J. Tyrrell, W. Schmidt, L. Fleming, G. Barrett, D. Whinney, N.Osborne. C. Sarran	ACCEPTED Otology & Neurology, with minor amendments.
3	**"The seasonality and effects of temperature and rainfall on Campylobacter infections"**	G. Nichols, A Djennad,	Resubmited paper to Epidemiology and Infection Journal.
4	**"A comparison of weather variables linked to infectious disease patterns using laboratory addresses and patients' home addresses"**	G. Loiacano, A Djennad, G Nichols, C Sarran, L Fleming, A Kessel, A Haines,	Paper to be submitted to International Journal of Health Geographics.
5	**Infectious Diseases - regression trawling**	G. Nichols, M. Cherrie	M. Cherrie produced draft paper. GN to share comments and then paper to be finalised.
6	**Osteoporosis**	M. Cherrie, N. Osborne	M. Cherrie and N. Osborne working on paper now data received from UK Biobank.
7	**Asthma exacerbations and pollen in the UK**	Alcock, Osborne, Wheeler, Hajat, Sarran, Clewlow, MacInnes, Hemming, White, Vardoulakis, Fleming	Paper being submitted in Nov 2016
8	**Medication use does not increase the risk of heat related illness among type 2 diabetes patients: a crossover analysis of over 4 million GP consultations across England**	S. Hajat. L. Fleming, A. Haines	Submitted to OEM in November 2016
9	**Tracing Data Journeys**	S. Leonelli, N. Tempini	Aiming to submit paper by December 2016

10	**Chapter on Big Data in Environment and Health Research for the Oxford Research Encyclopedia on Environment and Health**	L. Fleming	Accepted for submission. Draft shared with co-authors, feedback by 15/11/16

Additional Note

The MEDMI Project is a unique Partnership grant to demonstrate the feasibility of data mashups of environment and health data. One of the challenges revealed during the MEDMI progress to date has been the severe shortage of highly skilled personnel in the combined areas of software development and database management, as well as with Geographic Information Services (GIS) and web tool expertise. With the departure of Mr Whitmore as the only fulltime employee on MEDMI and the person who combines these skills (he remains as a consultant for parts of the MEDMI), we have had to piece together several different individuals with support from the Met Office to cover these functions. We have flagged to the University, the Partner Institutions, and the MRC, the lack of individuals with these important skill sets for big data mashups.

Of note, thanks to Professor Depledge, we have interacted with the **Black Swan**, a private company which uses an application to grab and analyze all types of data (primarily social media). The conversations confirmed the difficulty of the personnel shortage. Unfortunately, due to lack of time and resources, MEDMI did not subcontract Black Swan, but collaboration on the MRC Programme grant is possible in the future.

A Summary of MEDMI Pilot Projects

1. **Weather and Symptom Fluctuations in Ménière's's Disease**
   Jessica Tyrrell, Wiebke Schmidt (ECEHH), Christophe Sarran (Met Office)

Ménière's disease is an inner ear disorder that is long term, progressive and affects both the balance and hearing functions of the inner ear. Ménière's is a debilitating unpredictable disease, with high levels of psychosocial comorbidity and reduced quality of life amongst diagnosed individuals. Currently, there is no known cure for Ménière's, and only symptomatic drug treatment. Therefore significant emphasis is placed on self-management, with patients expected to identify and avoid triggers where possible. Atmospheric pressure is considered to be a possible trigger for spikes in Ménière's symptoms and attacks. A recent pilot study by researchers at the European Centre for Environment and Human (ECEHH) in collaboration with the Met Office and a local business (Buzz Interactive) has added to the evidence base that low atmospheric pressure may exacerbate symptoms. A mobile phone app has been developed which allows participants to monitor the 4 key symptoms of Ménière's (acute attaches, dizziness, hearing status and tinnitus), whilst collecting their GPS location data enabling amalgamation with Met Office weather data for that specific day.

Initial findings from 6 months of data suggested inverse associations between atmospheric pressure and symptom severity. This pilot project extended the analysis with a further 6 months of data, as well as including a broader range of weather variables (sea level pressure, station air pressure, temperature, wind speed, humidity, visibility). The environmental data was extracted from the MEDMI server using SSH Command line and then linked with Ménière's symptom data.

Key findings within this data set confirmed the inverse association between atmospheric pressure and the symptoms of Ménière's and identified potential associations with high humidity and increase odds of attacks and an inverse association between temperature and tinnitus. These findings remained when all other weather variables were considered and seasonality was adjusted for.

2. **Linkage Tools for Pre-Processing Pollen Data**
   Rachel McInnes and Christophe Sarran (Met Office)

Atmospheric pollen monitoring in the United Kingdom is carried out by the National Pollen Monitoring Network coordinated by the Met Office. The network consists of some, 10 to 20 sites that operate during the pollen season. Counts of pollen grains, differentiated by species, are used to forecast levels of pollen on a regional basis. As pollen concentrations present significant variations on the scale of kilometres, this means that at best, pollen measurements are indicative of risk at a regional level.

This Pilot Project offers a first step towards statistical estimates of atmospheric pollen concentration by linking pollen and meteorological datasets and using a Gaussian plume model to link with a high resolution land cover map. The Pilot Project provides the tools to compute atmospheric pollen concentration appropriate for use in health studies where estimates of exposure are important. The project has made use of pollen and meteorological data available on MEDMI. The tools are designed to allow compatibility with other input maps to allow researchers to analyse pollen concentrations from different sources. This can also be used to study the effect of future land use change or climate change on pollen concentration in the UK.

Four steps were identified for the development of the required datasets and tools. A preliminary step included generic MEDMI Database tools and an implementation of the Gaussian plume model. Step 1 aimed at writing code to combine pollen monitoring data with wind data using the large-time solution for the concentration for a continuously emitting source derived by Thomson and Manning. Step 2 is to calculate deviations from the expected mean pollen count at each monitoring site and derive estimates of pollen emissivity close to each site. Finally, step 3 was to use spatially interpolated emissivity and the large-time solution for pollen concentration to produce estimates of atmospheric pollen concentration for any given day and location.

The project developed and provided datasets and tools for MEDMI users towards providing statistical estimates of atmospheric pollen concentrations at any location in the United Kingdom. For the Gaussian plume model, it will be necessary to complete a numerical implementation by passing the correct Python objects with pollen emission values, wind speed and direction, to the function with the plume model. Once this is implemented, it will be possible to test different options for the plume turbulence scale for each species as well as the scaling distance for the deposition.

For the remaining deliverables, the Met Office plans to complete the work once a workable spatial coordinate conversion solution has been found and the Gaussian plume model is ready for use.

3. Osteoporosis and Solar Irradiance

Mark Cherrie, Nick Osborne (ECEHH) and Christophe Sarran (Met Office)

Osteoporosis is a progressive bone disease that is characterised by a decrease in bone mass and density, which can lead to an increased risk of fracture. Osteoporosis is common in the UK with three million sufferers and 300,000 people receiving treatment from the NHS each year for fragility fractures i.e. those occurring from a fall from standing height or less. A lack of vitamin D is associated with osteoporosis, individuals with levels of 25-hydroxy vitamin D below 30 nmol/l are at greatest risk, although some individuals between 30-50 nmol/l may also be insuffucient.This is significant given that approximately 47% of Britons have levels below 40 nmol/l during winter and spring and 15% during summer and autumn. Vitamin D is obtained through sun exposure and to a much lesser extent from dietary sources, but it can also be supplied pharmaceutically.

The UK Biobank provides an excellent opportunity to understand the determinants of osteoporosis due to the large number of incident fracture cases (e.g. 3,000 hip fracture cases by 2017), quantitative ultrasound measures of the heel, dual energy X-ray absorptiometry at multiple sites and extensive phenotypic information. Existing ultraviolet radiation (UVR) datasets from JAXA Satellite Monitoring for Environmental Studies (JAXA/EORC/JASMES) were utilised, with linkage facilitated by the MEDMI project, and merged with information on residential location, sun behaviours and dietary habits.The researchers investigated the complex interaction between modifiable factors responsible for osteoporosis.

The researchers hypothesised that lifetime UVB was inversely associated with osteoporosis risk and positively associated with skin cancer. Key effect modifiers, such as pharmaceutical use of vitamin D, were investigated to inform on which individuals may benefit most from specific guidelines.

There is potential to apply similar methods to extra-skeletal diseases, with which vitamin D has been associated previously.

4. Childhood Obesity and Neighbourhood Environments: integrating MEDMI and SAIL

Sarah Rodgers, Richard Fry (SAIL, Swansea Univeristy), Ben Wheeler (ECEHH)

Evidence is accumulating that neighbourhood green/blue space may promote good health and well-being through various mechanisms, including supporting and promoting physical activity.

This pilot project investigated the opportunity to combine the strengths of MEDMI and SAIL (The Secure Anonymised information Linkage Dataset at Swansea University) to link environmental, health and socio economic data. Specifically the project looked at the association between neighbourhood characteristics (especially greenspace and socio-economic deprivation), environmental phenomena (rainfall) and childhood obesity.

The analysis involved extracting National Child Measurement Programme data from SAIL. These data are highly protected, meaning that in order to merge them geographically with environmental data from MEDMI, a reasonable proportion needed to be suppressed. The loss of data is non-random (since suppression is related to prevalence and area population), meaning considerable potential bias has been introduced into the analysis. In future, it is proposed that the environmental data is extracted and linked to the health data securely within SAIL, where suppression will not be required.

The project also provided some useful lessons about the ease of accessing MEDMI data and has contributed to plans for the development of a new web-interface for accessing MEDMI data."

5. Statistical Downscaling of Gridded Air-Quality Data

Sujit Sahu, Mark Bass (Southampton University)

This project set out to produce model based downscaling of gridded air quality data for England and Wales. Daily estimates of air pollution for four most harmful pollutants NO2, O3, PM10 and PM2:5 were available from the Air Quality Unified Model (AQUM), developed by Savage et al. (2013). AQUM is a 3-dimensional weather and chemistry transport model used to deliver the UK national air quality forecast for DEFRA (Department for Environment, Forestry and Rural Affairs) and for scientific studies of atmospheric composition and air quality. The AQUM was run in hindcast mode to re-create hourly varying, air pollution concentrations for a 1-kilometer square grid covering the UK for the five year period 2007 to 2011. These hourly values were averaged to obtain the daily estimates which constituted the gridded air-quality data in this project. However, these air pollution estimates are biased as noted by Savage et al. (2013) and to overcome these biases, this project aimed to use Bayesian model based methods.

Complex Bayesian models for bias correction of AQUM outputs have been developed by Mukhopadhyay and Sahu (2016). The models make use of all publicly available observed data from 144 active AURN (Automatic Urban and Rural Network) air-pollution monitoring sites in England and Wales. These data, downloaded from the website of DEFRA1, represent ground truth and are of the best possible quality in contrast to the biased AQUM estimates. However, these data are only available for 144 sites within our study region and there is a large percentage of missingness due to various reasons including instrument malfunction. The main task of the Bayesian modelling effort was to develop a space-time bias correction method for the AQUM estimates using the observed data as the ground truth for each of the

four pollutants separately. Further details about the modelling effort and the level of accuracy achieved are provided in the technical report by Mukhopadhyay and Sahu (2016), which has been submitted for a peer reviewed publication.

In this pilot project we use the model developed by Mukhopadhyay and Sahu (2016) to obtain the best biased corrected estimates of air quality at each corner point of each of the 151,284 1-kilometer grid points covering the study region. The Bayesian model, implemented using iterative Markov chain Monte Carlo methods, requires substantial computer code development and computing resources to obtain the daily predictions at each of the 151,284 points. The MEDMI funding enabled us to perform these prediction tasks for three most recent years 2011, 2010 and 2009 for each of the four pollutants. These estimates have been made available to the MEDMI project for downloading purposes.

In summary, this project has achieved its main goal of statistical modelling to produce accurate high-resolution spatio-temporal estimates of air quality for four most harmful pollutants in England and Wales for the three years 2009, 2010 and 2011. The project has exceeded expectations by providing local-authority wise estimates of pollution levels along with their uncertainties. Further funding is required to produce similar estimates for the remaining two years 2008 and 2007.

6. Climate and Pathogen Seasonality in England and Wales

Gordon Nichols (Public Health England), Mark Cherrie (Exeter University)

Seasonality may be a proxy for some organisms whose incidence is affected by the environment in general and by climate in particular. This project aimed to systematically document the seasonality of pathogen-related infectious diseases reported in England and Wales to explore the possible influence of climate.

Human infectious disease data from Public Health England's LabBase2 national surveillance database and modelled climate date from the Met Office were utilised, with linkage facilitated by the MEDMI project. The researchers conducted a time series analysis of 277 different pathogens (i.e. the top 75% in terms of total case count). Each organism's time series was decomposed at weekly, monthly and quarterly periodicities, and forecasted using the Tbats package in R. Seasonality was detected using model fit statistics.

The results showed that whilst the majority of infectious disease organism serotypes displayed a seasonal component, only 36 serotypes displayed a moderate to strong correlation with climatic factors. This group included pathogens (e.g. Campylobacter) that have been known for some time to follow seasonal patterns, as well as other less-studied organisms or seasonal components with less frequently used climatic variables.

A key outcome of the project was a table of 36 pathogen serotypes with the details of their potential links with climate. In addition, an Rshiny app was developed to help with dissemination of results http://markcherrie.shinyapps.io/medmi_app. The user is able to filter the pathogens by seasonality, prevalence and serotype. Once an individual serotype is selected, a range of information is provided visually.

7. **<u>Tracing Data Journeys Across Climate, Environment and Human Health: A Qualitative Study of MEDMI</u>**
   Sabina Leonelli and Niccolo Tempini, University of Exeter

This is a study of MEDMI's project development and outcomes, with particular attention to the challenges in data integration, storage, dissemination and re-use identified by project participants and how those relate to international developments in data infrastructure, Open Data practices and relevant guidelines and policies.

In addition, MEDMI will be used as a case-study for the DATA-SCIENCE project led by Sabina Leonelli with funding from the ERC, thus positioning MEDMI alongside other major data integration projects in the biological and biomedical science that are currently being identified by the project.