# Commonwealth of Massachusetts
# Center for Health Information & Analysis (CHIA)
# Non-Government Application for MA APCD Limited Data Set
# [Exhibit A: Data Application]

*This form is required by all Applicants, except Government Agencies as defined in 957 CMR 5.02. All Applicants must also complete the Data Management Plan, attached to this Application. The Application and the Data Management Plan must be signed by an authorized signatory of the organization. This Application and the Data Management Plan will be used by CHIA to determine if your organization may receive CHIA data. Please be sure the documents are completed fully and accurately. You may wish to consult the Evaluation Guide that CHIA will use to review your documents. Prior to receiving CHIA Data, the organization must execute the Data Use Agreement. You may wish to review that document as you complete these forms. This application should be completed by the Primary Investigator, and must be signed by a party with authority to bind the organization seeking CHIA Data for the purposes described herein.*

***NOTE**: In order for your Application to be processed, you must submit the required application fee. Please consult the fee schedules for MA APCD data for the appropriate fee amount. A remittance form with instructions for submitting the application fee is available on the CHIA website.*

*All attachments must be uploaded to IRBNet with your Application. All applications documents can be found on the CHIA website* in Word and/or PDF format.

## I. GENERAL INFORMATION

| APPLICANT INFORMATION | |
|---|---|
| Applicant Name:<br>(Primary Investigator) | Dr Helen Suh |
| Title: | Professor |
| Organization Requesting Data:<br>(Recipient) | Northeastern University |
| Project Title: | Assessing Susceptibility to Air Pollution Using Hybrid Data Mining and Epidemiological Techniques |
| IRBNet ID: | 787847-1 |
| Address, City/Town, Zip Code | 316 Robinson Hall<br>Department of Health Sciences<br>Northeastern University<br>Boston, MA 02115 |
| Telephone Number: | 617-373-5925 |
| Email Address: | h.suh@northeastern.edu |
| Names of Co-Investigators: | |
| Email Addresses of Co-Investigators: | |
| Original Data Applicant Submission Date: | February 16, 2015 |
| Dates Data Application Revised: | December 7, 2016 |
| Project Objectives (240 character limit): | The purpose of this project is to evaluate the impact of air pollution on clinical morbidity indicators using novel techniques that combine data mining and epidemiological methods and to identify factors that affect susceptibility to these impacts. We will do so by linking CHIA data to a variety of air pollution |

| | |
|---|---|
| | measures, including fine particulate air pollutants (PM2.5), ozone, PM2.5-associated metals and black carbon, and PM2.5 sources. |
| Project Research Questions (if applicable) or Business Use Case(s): | 1. What is the association between air pollution and clinical morbidity indicators, including cause-specific hospital admissions, emergency department visits, and out-patient visits?<br>2. How do these impacts vary by air pollutant and health outcome?<br>3. What population groups are most susceptible to air pollution's harmful effects? How does susceptibility vary with age?<br>4. Does roadway density, green space, and socioeconomic conditions of neighborhoods modify air pollution-health impacts? |

## II. PUBLIC INTERST & PROJECT SUMMARY

1. Briefly explain why completing your project is in the public interest.

This project will help to predict air pollution-mediated health risks for Massachusetts residents and examine how these risks change as people age and how they may vary with neighborhood of residence. To do so, we will create a new database that links CHIA data with neighborhood data and with daily, monthly and yearly air pollution concentrations measured at multiple sites within Massachusetts and estimated using GIS-based spatio-temporal models. We will analyze these linked data using traditional epidemiological approaches and data mining approaches.

The ultimate goals of our analyses will be to examine the relationship of air pollution (fine particles (PM2.5), ozone, PM2.5-associated metals and black carbon, and PM2.5 sources) and cause-specific hospital admissions, emergency department visits, and outpatient visits. In so doing, we will identify populations at greatest risk, as well as critical time periods and/or conditions when health risks may change.

2. Has an Institutional Review Board (IRB) reviewed your project?

☒ Yes, a copy of the *approval letter* and *protocol* must be included with the application package on IRBNet.

☐ No, this project is not human subject research and does not require IRB review.

3. **Research Methodology**: Applicants must provide a written description of the project methodology (typically 1-2 pages), which should state the project objectives and/or identify relevant research questions. This document must be included with the application package on IRBNet, and must provide sufficient detail to allow CHIA to understand how the data will be used to meet objectives or address research questions. Applications that do not include this methodology statement cannot be reviewed or approved.

## III. DATA FILES REQUESTED

1. Please indicate the MA APCD databases from which you seek data, the year(s) of data requested, and your justification for requesting *each* file. Please refer to the MA APCD Release Data Specifications for details of the file contents.

| MA ALL-PAYER CLAIMS DATABASE FILES | Year(s) Of Data Requested<br>Current Yrs. Available<br>☒ 2011 ☒ 2012 ☒ 2013 ☒ 2014 ☒ 2015 |
|---|---|
| ☒ **Medical Claims** | **Please describe how your research objectives require Medical Claims data:**<br>To assess how health outcomes relate to a host of characteristics and exposures such as pollutant concentrations. To analyze impacts of exposures e.g. pollutants, weather patterns on health outcomes at various temporal resolutions. |
| ☒ **Pharmacy Claims** | **Please describe how your research objectives require Pharmacy Claims data:**<br>To enable control for confounding by medication use, which can be very important e.g. for some chronic metabolic outcomes associated with air pollution. |
| ☐ **Dental Claims** | **Please describe how your research objectives require Dental Claims data:** |
| ☒ **Member Eligibility** | **Please describe how your research objectives require Member Eligibility data:**<br>To allow control for member characteristics and relationships among subjects in the analysis of the associations between exposures and outcomes. |
| ☐ **Provider** | **Please describe how your research objectives require Provider data:** |
| ☐ **Product** | **Please describe how your research objectives require Product data:** |

## IV. GEOGRAPHIC DETAIL

Please choose one of the following geographic options for MA residents. *For releases with 5 digit zip code, CHIA will apply a substance abuse filter which will remove all claims that include a substance abuse diagnosis.*

| | |
|---|---|
| ☐ 3 Digit Zip Code (MA) (standard) | ☒ 5 Digit Zip Code (MA)*** |

*****Please provide justification for requesting 5 digit zip code. Refer to specifics in your methodology:**
To asses air pollution exposure (data will be linked to air pollution data and the member zipcode will help to locate specific monitors in the area and help us assess air pollution exposure status). This will be used to for assessing proximity to roadways, green spaces, and to generate aggregate values of neighborhood SES.

## V. DATE DETAIL

Please choose one option from the following options for dates:

| | | |
|---|---|---|
| ☐ Year (YYYY) (Standard) | ☐ Month (YYYYMM) *** | ☒ Day (YYYYMMDD) ***<br>[for selected data elements only] |

***** If requested, please provide justification for requesting Month or Day. Refer to specifics in your methodology:**
Many of the exposures we will be analyzing change on a daily basis, including air pollutant levels, weather patterns, traffic density levels etc. Therefore measures of health outcome occurrence at a daily level are very important for our analyses.

## VI. NATIONAL PROVIDER IDENTIFIER (NPI)

Please choose *one* of the following options for National Provider Identifier(s):

| | |
|---|---|
| ☒ Encrypted National Provider Identifier(s) (standard) | ☐ Unencrypted National Provider Identifier(s)*** |
| ***** If requested please, provide justification for requesting unencrypted National Provider Identifier(s). Refer to specifics in your methodology:** | |

## VII. MEDICAID DATA

Please indicate here whether you are seeking Medicaid Data:

☐ Yes
☒ No

Federal law (42 USC 1396a(a)7) restricts the use of individually identifiable data of Medicaid recipients to uses that are directly connected to the administration of the Medicaid program. If you are requesting Medicaid data from Level 2 or above, please describe, in the space below, why your use of the data meets this requirement. Requests for Medicaid data will be forwarded to MassHealth for a determination as to whether the proposed use of the data is directly connected to the administration of the Medicaid program.

## VIII. DATA LINKAGE AND FURTHER DATA ABSTRACTION

*Note: Data linkage involves combining CHIA data with other databases to create a more extensive database for analysis. Data linkage is typically used to link multiple events or characteristics within one database that refer to a single person within CHIA data.*

1. Do you intend to link or merge CHIA Data to other datasets?

☒ Yes
☐ No linkage or merger with any other database will occur

2. If yes, please indicate below the types of database to which CHIA Data be linked. [Check all that apply]

☐ Individual Patient Level Data (e.g. disease registries, death data)
☐ Individual Provider Level Data (e.g., American Medical Association Physician Masterfile)
☐ Individual Facility Level Data level (e.g., American Hospital Association data)
☒ Aggregate Data (e.g., Census data)
☐ Other (please describe):

3. If yes, describe the data base(s) to which the CHIA Data will be linked, which CHIA data elements will be linked; and the purpose for the linkage(s):

We will link the CHIA cause-specific hospital admissions, emergency department visit, outpatient visit, and medication use data from 2009-2015 to daily air pollution, behavioral risk factor, and census data. We will use this linked data set to examine the impacts of air pollution on health and to assess whether behaviors and neighborhood characteristics modify these impacts.

4. If yes, for each proposed linkage above, please describe your method or selected algorithm (e.g., deterministic or probabilistic) for linking each dataset. If you intend to develop a unique algorithm, please describe how it will link each dataset.

We will link CHIA, air pollution, behavioral and census data by CHIA date of admission and zip code. Each CHIA record will be linked deterministically.

5. If yes, please identify the specific steps you will take to prevent the identification of individual patients in the linked dataset.

CHIA data will be provide by zipcode, with each admission, visit, and medication usage accompanied by information regarding, gender, race/ethnicity, date of admission or visit, and zipcode of residence. As a result, linked CHIA data will contain no patient identifiers.

Regardless, all CHIA data files will be stored and maintained is a secure environment, with access only to approved NEU employees and approved third parties. The complete inventory of CHIA data files will be stored on the large data server, with only designated users allowed to read or execute data. Groups will be created to limit access to specific data files as necessary. For this project, designated users will not be allowed to download or write data outside the project's large data server and compute node.

Note that for this project, designated users will be allowed to link CHIA data files to other publicly available data files, such as those containing data on air pollution, behavioral and census variables. These linked data files will also be stored on the project's large data server, with access again restricted to designated users and no downloading or writing of these data sets outside the projects large data server and compute node. Ability to link data sets will be restricted to a minimal number of approved parties.

Northeastern's privacy and security program is compliant with federal government regulations. Our IT systems have completed the full federal C&A process for a number of government agencies. Northeastern has established protocols that are multilayered to secure computer systems and the data they contain. At MGHPCC we have a hardware firewall for all access to all NEU networks (management, storage and compute) in the data center. In addition physical access is restricted to only approved NEU employees via a dual key card and pod key entry with security personnel that check credentials before entry and pod keys are provided. Northeastern cluster accounts are validated using the Northeastern Windows Active Directory for only those users that have been approved to access the cluster. These accounts terminate automatically if a user is no longer affiliated to Northeastern University. They cannot use cluster resources when this happens. For third party access, a Northeastern faculty member has to sponsor the account, which in this case is Dr. Helen Suh, the PI of the study. Once sponsored by Dr. Suh, the account will then be enabled in the Northeastern Windows Active Directory for cluster access. Again once the sponsorship period ends, the account will be disabled in the Northeastern Windows Active Directory and cluster access is no longer possible. Northeastern currently has other government projects that require similar compliance with privacy and security of data, including census data from the Bureau of Census. The cluster can enforce "Secure" capability in accordance with IAW FIPS 200 (http://csrc.nist.gov/publications/fips/fips200/FIPS-200-final-march.pdf) and NIST 800-53(http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf) for example. To remotely connect to the compute environment a two factor authentication is used (user has to be enabled in Northeastern University Active Directory and user has to have clearance on login nodes for access).

Information Technology Services – Research Computing (ITS – RC) maintains applications, data and security standards on the cluster that support management of research projects (define permissions, users, projects at any point in time), software, user and group accounts, software and compute schedulers and related storage. Data between these systems may be exchanged on a real time basis.

The following are in place for security of non-anonymized or non-neutralized data that contain sensitive user or other information. These data are typically made available to specific researchers in fully disclosed format. The researcher can then decide what level of access users of the data can get.

1) Data in storage are in lock down with only designated users allowed to read, write or execute data. For this project, we will allow designated users to read and execute data, but not to download or write data outside the project's large data server and compute node.
2) Groups are created such that users in particular groups have structured and limited access to data as decided by the researcher who is the owner of the data.
3) When data are moved into the cluster encryption is used and when data are deleted secure deletion is employed.
4) When data are staged for run on the compute nodes the staging location is also in lock down.
5) Individual jobs can be restricted in their resource utilization, based on user/group/project membership.
6) Researchers can provide their own applications that are installed via restricted modules so that only designated persons can use them on the cluster.
7) All software is evaluated for stability and viruses before installing on the cluster.
8) The Federal Government and USCB have very strict access and security requirements. These security requirements are outlined in Census Bureau IT Security Program Policy, the DOC IT Security Program Policy and NIST SP 800-53r4I. The cluster adheres to these.
9) Specifically:
   a) Users can only access data that they have authorized access to via group and user specific memberships.
   b) Intermingling of data sets can be locked down. So any user cannot have access to or merge or query two separate data sets. These can be enforced by the researcher that owns the data sets easily on requesting what is required to ITS – RC.
   c) Separate backups if needed can be enforced without comingling with other data.
   d) Security tools are in place to monitor intrusions to the cluster like "ssh black lists", "tripwire", and "iptables".
   e) The cluster sits behind a firewall. All access to and out of the cluster is monitored and information is stored historically.

Both ITS – Research Computing, and the Office of Information Security periodically review cluster security and make recommendations to ensure continued compliance.

6. Once the linkage is made, what non-MA APCD data elements will appear in the new linked file?

Air pollutant concentrations, ZIP-code level averages of socioeconomic and demographic variables.

## XI. PUBLICATION / DISSEMINATION / RE-RELEASE

1. Describe your plans to publish or otherwise disclose CHIA Data, or any data derived or extracted from such CHIA Data, in any paper, report, website, statistical tabulation, seminar, conference, or other setting. All publication of CHIA Data must comply with CHIA's cell size suppression policy, as set forth in the Data Use Agreement. Please explain how you will ensure that any publications will not display a cell less than 11, and no percentages or other mathematical formulas will be used if they result in the display of a cell less than 11.

We will publish results from our analyses in peer-reviewed publications in relevant scientific journals and will present these results at scientific conferences. All results will be de-identified aggregates with no cells less than 11. If these occur, they will be represented as missing.

2. Do you anticipate that the results of your analysis will be published and/or publically available to any interested party? Please describe how an interested party will obtain your analysis and, if applicable, the amount of the fee, that the third party must pay.

Results from our analyses will be made available through journals and scientific conferences. We will not charge any fees for analyses involving CHIA data.

3. Will you use CHIA Data for consulting purposes?

☐ Yes
☒ No

4. Will you be selling standard report products using CHIA Data?

☐ Yes
☒ No

5. Will you be selling a software product using CHIA Data?

☐ Yes
☒ No

6. Will you be reselling CHIA Data in any format?

☐ Yes
☒ No

If yes, in what format will you be reselling CHIA Data (e.g., as a standalone product, incorporated with a software product, with a subscription, etc.)?

7. If you have answered “yes” to questions 4, 5 or 6, please describe the types of products, services or studies.

8. If you have answered “yes” to questions 4, 5, or 6, what is the fee you will charge for such products, services or studies?

## X. APPLICANT QUALIFICATIONS

1. Describe your qualifications (and the qualifications of your co-investigators) to perform the research described.

Professor Helen Suh is a Professor in the Department of Health Sciences in the Bouve College of Health Sciences at Northeastern University and is the Director of the Population Health doctoral program at Bouve, adjunct faculty at the Harvard School of Public Health and a Senior Fellow at NORC at the University of Chicago. Dr. Suh earned a Sc.D. and an MS in Environmental Health from the Harvard School of Public Health, and a SB in Biology from the Massachusetts Institute of Technology.

An internationally recognized expert in air pollution health effects in the areas of environmental epidemiology, exposure assessment and air pollution, Dr. Suh has led multidisciplinary teams in environmental exposure assessment and epidemiology for over 20 years. Her research focuses on three general areas within air pollution health effects, including: 1) assessment of the impact of lifestyle and neighborhoods on air pollutant exposures and human health; 2) examination of multi-pollutant impacts on human health; 3) development of GIS-based spatio-temporal modeling tools for epidemiological research. Dr. Suh is the lead investigator in numerous research projects, including an NIH-funded study investigating the impacts of air pollution and lifestyle of cognitive and cardiac health and an EPRI-sponsored study examining the association of chronic air pollution exposures and mortality. As part of these studies, Dr. Suh develops and uses innovative analytic tools, such as GIS-based spatio-temporal models that are able to predict air pollution exposures and consider their diverse sources and properties. Results from Dr. Suh's studies have already and will continue to advance our understanding of air pollution health impacts and the development of appropriate policies and regulations to reduce their health risks.

Dr. Suh's research has been published in leading journals, including *Environmental Health Perspectives* (EHP), *Epidemiology*, and *Circulation*. Her papers have had significant impact, as evidenced by her h-index of 50 and her almost 8000 citations. Dr. Suh's expertise is well recognized. Dr. Suh is an Associate Editor of the *Journal of Exposure Science and Environmental Epidemiology* (JESEE), a leading journal that publishes exposure assessment and exposure-motivated environmental epidemiology research. In addition, Dr. Suh has been appointed by the US EPA Administrator to the Clean Air Scientific Advisory Committee (CASAC), a seven-member committee that provides independent advice to the EPA Administrator on the technical bases for EPA's national ambient air quality standards. She is also a member of the CASAC subcommittees for ozone, nitrogen oxides, and sulfur dioxide and previously served as a member of the CASAC Committee for Particulate Matter. She was recently a member of the National Academy of Sciences (NAS) Committee on Scientific Tools and Approaches for Sustainability, charged with an evaluation of scientific tools and approaches for incorporating sustainability concepts into assessments used to support EPA decision-making. Previously, Dr. Suh served on several Institute of Medicine and National Academy of Science committees and has been invited to speak at numerous workshops and conferences.

2. **<u>Resumes/CVs</u>**: Please include with your application package on IRBNet résumés or curricula vitae of the Applicant/principal investigator, and co-investigators. (These attachments will not be posted on the internet.)

## XI. USE OF AGENTS AND/OR CONTRACTORS

**Please note: by signing this Application, the Organization assumes all responsibility for the use, security and maintenance of the CHIA Data by its agents, including but not limited to contractors.**

Provide the following information for all agents and contracotrs who will have access to the CHIA data. *Add agents or contractors as needed.*

| | |
|---|---|
| Company Name: | |
| Contact Person: | |
| Title: | |

| | |
|---|---|
| Address, City/Town, Zip Code | |
| Telephone Number: | |
| E-mail Address: | |
| Organization Website: | |

1. Will the agent or contractor have access to or store the CHIA Data at a location other than the Applicant's location, off-site server and/or database?

☐ Yes, a separate Data Management Plan **must** be completed by each agent or contractor

☐ No

2. Describe the tasks and products assigned to this agent for this project; their qualifications for completing the tasks; and the Organization's oversight of the agent, including how the Organization will ensure the security of the CHIA Data to which the agent or contractor has access.

| |
|---|
| |

| | |
|---|---|
| Company Name: | |
| Contact Person: | |
| Title: | |
| Address, City/Town, Zip Code | |
| Telephone Number: | |
| E-mail Address: | |
| Organization Website: | |

1. Will the agent or contractor have access to or store the CHIA Data at a location other than the Applicant's location, off-site server and/or database?

☐ Yes, a separate Data Management Plan **must** be completed by each agent or contractor

☐ No

2. Describe the tasks and products assigned to this agent for this project; their qualifications for completing the tasks; and the Organization's oversight of the agent, including how the Organization will ensure the security of the CHIA Data to which the agent or contractor has access.

| |
|---|
| |

## XII. FEE INFORMATION

Please consult the fee schedules for MA APCD Data and select from the following options:

☐ Researcher

☐ Others (Single Use)

☐ Others (Multiple Use)

Are you requesting a fee waiver?
☒ Yes
☐ No

If yes, please refer to the Application Fee Remittance Form and submit a letter stating the basis for your request (if required). Please refer to the fee schedule for qualifications for receiving a fee waiver. If you are requesting a waiver based on the financial hardship provision, please provide documentation of your financial situation. Please note that non-profit status alone isn't sufficient to qualify for a fee waiver.

## XIII. ATTESTATION

By submitting this Application, the Data Applicant attests that it is aware of its data use, privacy and security obligations imposed by state and federal law *and* is compliant with such use, privacy and security standards. The Data Applicant further agrees and understands that it is solely responsible for any breaches or unauthorized access, disclosure or use of any CHIA Data provided in connection with an approved Application, including, but not limited to, any breach or unauthorized access, disclosure or use by its agents.

Applicants requesting data from CHIA will be provided with data following the execution of a Data Use Agreement that requires the Data Applicant to adhere to processes and procedures aimed at preventing unauthorized access, disclosure or use of data.

**By my signature below, I attest to: (1) the accuracy of the information provided herein; (2) that the requested data is the minimum necessary to accomplish the purposes described herein; (3) the Data Applicant will meet the data privacy and security requirements describe in this Application and supporting documents, and will ensure that any third party with access to the data meets the data use, privacy and security requirements; and (4) my authority to bind the organization seeking CHIA Data for the purposes described herein.**

| | |
|---|---|
| Signature:<br>(Authorized Agent) | |
| Printed Name : | |
| Title: | |
| Signature:<br>(Applicant/Primary Investigator) | |
| Name: | Dr Helen Suh |
| Title: | Professor |
| Original Data Request Submission Date: | February 16, 2015 |
| Dates Data Request Revised: | December 07, 2016 |

Attachments. Please indicate below which documents have been attached to the Application and uploaded to IRBNet:
☒ 1. IRB approval letter and protocol (if applicable)
☒ 2. 1-2 page Research Methodology
☒ 3. Resumes of Applicant and co-investigators
☒ 4. Data Management Plan (including one for each agent of contractor that will have access to or store the CHIA Data at a location other than the Applicant's location, off-site server and/or database)

☒ 5. Fee Remittance Form (including any required documentation if a fee waiver is being requested)