Data Management Plan: CARE Platform

I. Types of data

The CARE Platform will utilize existing public health related data from a variety of sources, with a focus being on the health data available from *data.gov*, injury data available from FARS/NHSTA, and customized versions of data from these sources, as available from the Community Health Institute and SafeRoadMaps/CERS. The research project will create a software tool for conducting discovery research on these datasets. This tool will be made available to the research community at the conclusion of the project.

The CARE Platform will be hosted at the San Diego Supercomputer Center (SDSC). SDSC is a world-class facility for high performance data intensive computing with a state of the art, energy efficient datacenter. It provides leading edge computational, analysis, and storage capabilities uniquely suited for cyberinfrastructure research, hosting, and services, including colocation facilities and support for "condo" clusters. As described in the proposal, the CARE platform will host some data, e.g. the FARS database, and also serve as a cache for other data from remote sources, such as Healthy Communities (HCI). Derived data products may be created by the CARE platform utilizing specialized systems, methods and processes discussed within this proposal.

These data as well as all the documents relating to the project, including publications, reports, papers, and project summaries, will be captured on systems at SDSC. In addition, the center will also maintain and update supporting documentation relating to all software design, systems analysis, and system maintenance and support requirements

II. Data and Metadata Standards

The requirements analysis phase (I) of the research project will include the identification and confirmation of appropriate data and metadata standards for the CARE platform. One of our tasks will be to establish standards for the metadata tags that will be created by the CARE Platform for products derived by data analysis tools, as well as metadata and related documentation for data accessed and stored by the platform. This data and documentation will be made available to the research community, upon completion of the project.

For the two test cases, Community Health Institute and SafeRoadMaps, the data sources will be described and, where applicable, stored using well-defined domain models and xml standards, to allow for interoperability in future with other data systems and applications.

III. Policies for access and sharing and provisions for appropriate protection/privacy

As detailed in the project description, the CARE platform in intended to be a research cloud service that provides analytical middleware for use in analyzing health data. During the project, access will be limited to project team member and invited expert stakeholders through a password protected website. Commencing with Task 5 (month 26), means for access by the broader research community will be implemented. At that time, the project team will determine

whether there is a need for initiating access charges, which may be appropriate for securing the longer terms sustainability of the CARE platform and analysis tools.

All of the data that will be utilized are publicly available data sets that have been de-identified by public agencies and have passed their standards for privacy protection and assurance so that no individually identifiable data is provided. The datasets to be utilized within this project and other intellectual property have been released without restriction.

Over the course of the study, the project team will meet with both the Community Health Institute and the SafeRoadMaps/CERS team to arrive at a data-sharing agreement for post-project utilization of their data. Such an agreement will provide a model for not only this partnership, but for licensing the CARE Platform analytics for use by other health data sets.

IV. Policies and provisions for re-use, re-distribution

As noted in the project description, policies for provision and re-use will be developed as part of the research project. It is anticipated that there will be considerable interest in the platform and tools within the research and practice community, including academic researchers, health research agencies, and cloud service providers, among others. The need for such a tool was identified during a recent NSF sponsored symposium on Health Cyberinfrastructure, which was conducted by the PIs.

V. Plans for archiving and Preservation of access

The project website and service will contain all appropriate information and documentation for using the CARE platform and tool for health research discovery and analysis. The site will also contain all references, research papers, and related products developed throughout the course of the project.

The San Diego Supercomputer Facility at UC San Diego will host the data throughout the research project and provide a minimum of three years of online access beyond the completion of the project. Data storage will be performed at the nominal rates charged by SDSC to any project using the facility. These are relatively modest (~$1000/TB) and can be borne ahead of time for the 3-year period. Should the CARE platform not extend beyond the three years (post grant), the data could then be archived at SDSC at even lower cost. A decision would have to be made at that point in time regarding how exactly to archive the data, and on paying for the archival storage.