Poc3 – Big Data Processing Platform – Identifying open source health care data sources for Phase 3 - Smart Room

1 INTRODUCTION

1.1 OVERVIEW

This document has been written to specify the tasks and deliverables associated with **identifying health care data sources for** Phase 3 **Smart Room**.

This project (Poc3 – Phase 3 – Smart Room) sits part of a continuing asset demonstration UA16-016 - "Systems and Methods of Analyzing Healthcare Data." The previous phases were about technology selection, assessment and demonstration, summarized as follows:

- Phase 1 (March 2016 – August 2016) was concerned with technology selection and demonstrating that these technologies are able to work together and process a small healthcare data set.

- Phase 2 (November 2016 – January 2017) was concerned with providing speech-to-text and natural language processing capabilities to the solution. The target for this research was to provide the basis of analyzing doctor-patient interaction and even assistance with patient note taking by the doctor.

This project (Phase3) is about developing a Big Data Analytics Processing platform to assist in improving patient outcomes through processing and analysis of a wide range of health care data sets.

1.2 WHAT IS SMART ROOM?

Smart Room consists of a four-step data storage and retrieval system. Through combination of data from these various sources and examine the effectiveness of the data fusion by implementing a real-life Big-Data processing solution.

To develop Smart Room, we will be leveraging all the technology from the previous phases. This will mean staying aligned with cutting edge technologies and utilizing a wider range of data sources.

1.3 IMPORTANCE OF IDENTIFYING A WIDE VARIETY OF DATA SOURCES

The wide variety of open source medical data sets that are available today provide untapped opportunities for analysis and insight generation through Big Data Analytics technologies, that could improve patient outcomes. To this end a key operation task in this phase is to identify a selection of these datasets.

1.4 TASKS AND DELIVERABLES

Task #1: Identification and listing of open source health care datasets

Prerequisites:

Read the following documents:

- PoC1 - Deliverable 1.docx
- PoC1 - Deliverable 2.docx
- PoC2 - Project Deliverable - Final Report.docx

Deliverable #1: The specific deliverables are as follows:

1. *Identify a list of 20 varied open source data sets.*
2. *Download representative samples from each data set. If possible, the overall dataset should be at least 200MB. Although the majority of the data is expected to be CSV files, see if any of the downloaded samples could be text, audio and video as well.*
3. *Construct a spreadsheet using CSV file format that lists internet links to these sources, along with a description of each data source, it's origin and any other metadata available*
4. *Construct a spreadsheet using CSV file format to list the filename and path to each downloaded sample and metadata.*
5. *Write a report describing the investigation and results*