Scientific Reproducibility Using the Provenance for Clinical and Healthcare Research Framework

Satya Sahoo, Case Western Reserve University*; Joshua Valdez, Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University; Michael Rueschman, Departments of Medicine, Brigham and Women's Hospital and Beth Israel Deaconess Medical Center, Harvard University; Matthew Kim, Departments of Medicine, Brigham and Women's Hospital and Beth Israel Deaconess Medical Center, Harvard University; Susan Redline, Departments of Medicine, Brigham and Women's Hospital and Beth Israel Deaconess Medical Center, Harvard University*

Scientific reproducibility is key to scientific progress as it allows the research community to build on validated results, protect patients from potentially harmful trial drugs derived from incorrect results, and reduce wastage of valuable resources. The National Institutes of Health (NIH) recently published a systematic guideline titled "Rigor and Reproducibility" for supporting reproducible research studies, which has also been accepted by several scientific journals. These journals will require published articles to conform to these new guidelines. Provenance metadata describes the history or origin of data and it has been long used in computer science to capture metadata information for ensuring data quality and supporting scientific reproducibility. In this paper, we describe the development of Provenance for Clinical and Healthcare Research (ProvCaRe) framework together with a provenance ontology to support scientific reproducibility by formally modeling a core set of data elements representing details of research study. We extend the PROV Ontology (PROV-O), which has been recommended as the provenance representation model by World Wide Web Consortium (W3C), to represent both: (a) data provenance, and (b) process provenance. The ProvCaRe ontology is being used to extract provenance metadata to generate Resource Description Framework (RDF) provenance graphs from biomedical text to characterize provenance metadata described in published scientific studies that can be used for replicating scientific studies. We describe the results of analyzing 50 research publications using the ProvCaRe framework.