New toolkit provides more efficient analysis of health data to drive improvements in patient care
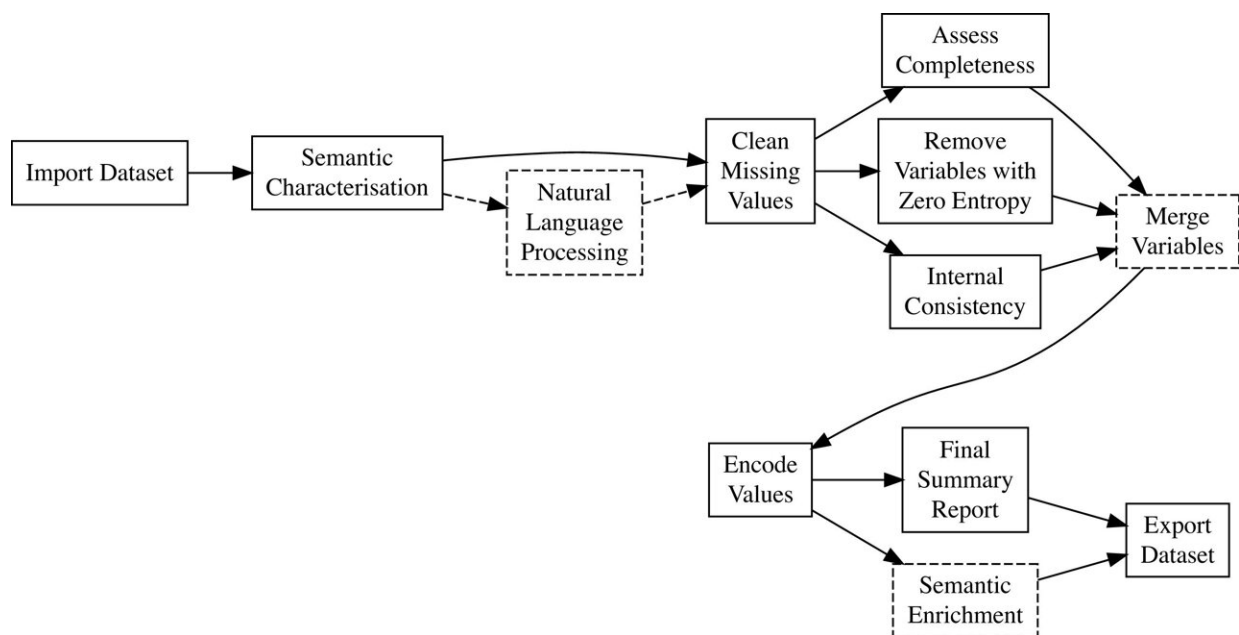
August 29 2023

Overview of eHDPrep workflow. The ordering of steps reflects logical dependencies. Dashed arrows and boxes signify optional steps. Following import, the semantic characteristics of the data are established, missing values are dealt with, and a series of operations may be performed. "Natural Language Processing" is only required if free-text variables are present. "Merge Variables" is an optional step for user-defined merging operations with functionality to measure information loss. Variables are encoded in a machine-interpretable format, and a summary report is generated for review by the user. Additionally, functionality to review each step is provided. "Semantic Enrichment" optionally involves aggregation of variables according to semantic commonalities identified

by an ontology such as SNOMED CT. Credit: *GigaScience* (2023). DOI: 10.1093/gigascience/giad030

Researchers from Queen's University have developed a new toolkit that harnesses the power of "Big Data" for digital health with the aim of driving improvements in patient care and outcomes through data-driven innovation.

The toolkit named eHDPrep has been made freely available to allow both the researchers themselves and other researchers to more effectively and reliably analyze large health datasets.

The understanding that arises from these analyses is hoped to produce better and more effective clinical tools that provide information to assist health professionals in making clinical decisions, such as determining which treatment may be more effective to treat a certain type of cancer.

The research has been published in the journal *Gigascience* and is a collaboration between the Data Intensive Biomedicine Group from the Patrick G Johnston Centre for Cancer Research (PGJCCR) at Queen's University, the Centre for Secure Information Technologies from the Institute of Electronics, Communications and Information Technology (ECIT) at Queen's, the Cancer Epidemiology Group from the Centre for Public Health (CPH) at Queen's, and the LifeArc Data Sciences Group.

Big Data refers to large data sets consisting of both structured and unstructured data that are analyzed to find insights, trends, and patterns. Health care Big Data involves collecting, analyzing, and leveraging consumer, patient, physical, and clinical data that is too vast or complex to be understood by standard data processing approaches.

Big Data is often processed and analyzed by data scientists, who deploy advanced computational approaches. These analyses can guide decision-making, improve patient outcomes and decrease health care costs. There is significant potential for the application of Big Data in health care, but there are still issues to overcome for us to realize and benefit from its full potential.

The eHDPrep tool enhances data quality which is a current major issue with effective use of health data. For example, providing methods for elimination of inconsistencies, removal of redundancy, increasing completeness and appropriately coding the data so that it is machine-interpretable, which is crucial for computational analyses.

The tool also enables a better understanding of health data by joining information together into higher level concepts that can reveal non-obvious links between different patients—in a process called "semantic enrichment."

This semantic enrichment process provides greater statistical power to make discoveries, for example highlighting key factors that drive disease progression in cancer and cardiovascular disease.

The research team have applied the eHDPrep tool to two datasets from colorectal cancer, one from Northern Ireland and another from The Cancer Genome Atlas (U.S.). The data cleaning and enrichment processes from eHDPrep is an important enabling step for them to develop new ways of grouping patients in order to advance colorectal cancer precision medicine.

The researchers hope this new understanding will ultimately lead to new treatments and diagnostics that will benefit colorectal cancer patients.

Commenting on the importance of the research, Tom Toner, Ph.D.

student from the Overton Research Group in Patrick G Johnston Centre for Cancer Research at Queen's University and first author on the research, said, "The exponential growth of Big Data in health care presents a significant challenge in extracting meaningful insights and driving improvements in patient care. With eHDPrep, we address the crucial issue of data quality in Big Data and enhance the analysis process by incorporating semantic enrichment."

"We are excited about the potential impact of eHDPrep on advancing precision medicine, particularly in the field of colorectal cancer. By making this toolkit freely available, we're ensuring that other researchers can also benefit from its capabilities and contribute to the collective efforts in improving patient outcomes."

Dr. Ian Overton, Data Intensive Biomedicine Research Group Leader and Reader (Associate Professor) from the Patrick G Johnston Centre for Cancer Research at Queen's University, said, "Data quality is fundamental for success in cancer research. Our new toolkit eHDPrep cleans Big Data for health by throwing the garbage out and is already helping with research work on colorectal cancer in my own group."

"Also, by enriching our datasets to discover otherwise non-obvious connections we can find new links between patients and are using these in our work towards new medicines in the fight against cancer."

More information: Tom M Toner et al, Strategies and techniques for quality control and semantic enrichment with multimodal data: a case study in colorectal cancer with eHDPrep, *GigaScience* (2023). DOI: 10.1093/gigascience/giad030

Citation: New toolkit provides more efficient analysis of health data to drive improvements in patient care (2023, August 29) retrieved 27 April 2024 from https://medicalxpress.com/news/2023-08-toolkit-efficient-analysis-health-patient.html