Design of a Colorectal Cancer Data Warehouse

TREO Talk Paper

Bunyamin Ozaydin, PhD
University of Alabama at Birmingham
bozaydin@uab.edu

Upender Manne, MS, PhD
University of Alabama at Birmingham
upendermanne@uabmc.edu

Sue Feldman, RN, MEd, PhD
University of Alabama at Birmingham
sfeldman@uab.edu

Abstract

Colorectal cancer researchers spend a substantial amount of effort performing integration, cleansing, interpretation, and aggregation of raw data from multiple sources, including health records and clinical research data. These efforts are often replicated for each project, with investigators running up against the same challenges and experiencing the same pitfalls discovered by those before them. The overall objective of this project is to design and implement a colorectal cancer data warehouse infrastructure to improve acquisition, management, and analysis of relevant health records, clinical research, and tumor registry data from our institution and state. The current data preparation processes, at best, are inefficient, costly, time-consuming, and cumbersome (Puppala et al. 2015). Moreover, without the needed information technology (IT) infrastructure, the potential of the ever-growing heterogeneous data accumulated in disparate data sets would be still untapped.

Our previous colorectal cancer work included discovery and validation of biomarkers, the roles of tumor location and race/ethnicity, treatment efficacy, and prognostic/predictive models that considered the effect of molecular, demographical, epidemiological, and clinico-pathologic features on outcomes, such as mortality, relapse, and survival (Bovell et al. 2013). The data sources for these projects included data exports from clinical records and spreadsheet files created for each research project. Data management for each project is usually performed in an ad-hoc manner, involving manual processes of data entry, matching, and merging. This process is error-prone and inefficient for data reuse, and not suitable to incorporate additional data sources.

This work proposes to initially design and implement a colorectal cancer data warehouse infrastructure that incorporates existing molecular-level and patient-level research data with continuous data feed from institutional enterprise data warehouse (EDW) in a multidimensional database format. We then plan to expand the scope of this data infrastructure to include social determinants of health (SDH), geospatial census, and state level tumor registry data. Such a data management platform will enable us to efficiently analyze disparities among various populations, create state-wide map projections and dashboards, and analyze certain outcomes (e.g. risk and aggressiveness of the disease) to identify differences between rural and urban populations. Creation of a colorectal cancer data warehouse infrastructure would allow us to store, clean, and manage the existing data sources efficiently, increase the quality and reliability of underlying data for our research, and incorporate new data sources to facilitate future research.

References
