Abstract

Early, accurate diagnosis of disease can dramatically improve prognosis. Clinical diagnostic model research attempts to optimize early diagnosis by designing diagnostic models based on variables obtained by the least invasive means.
Diagnostic model research currently involves a complex, multidisciplinary workflow involving data collection by clinicians on the one hand, and data preprocessing and machine-learning by machine-learning experts on the other.

Due to the traditional lack of integration between software packages used in this workflow, preparing data for analysis can require considerable manual effort. Following data extraction, data have to be inspected for conversion issues. The absence of information about a Case Report Form (CRF)'s structure in extracted data further requires manual guidance during preprocessing. As a result, data analysis is typically only performed once, after the data set reaches a certain predetermined size, based on rules of thumb or Monte Carlo simulations.

This thesis presents the Clinical Data Miner (CDM) software framework, which integrates data collection, data preprocessing and machine-learning in a single platform. This integration eliminates the error-prone, time-consuming steps of preparing data for analysis, and enables the automation of preprocessing steps that rely on information about a CRF's structure. The increased automation streamlines the diagnostic model research workflow. With its built-in functionality for generating learning curves, it furthermore provides study coordinators insight into how predictive performance evolves as patient set sizes grow. This allows them to make an informed decision about whether to continue or terminate data collection, thereby respectively avoiding both the creation of weakly performing models, as well as unnecessary data collection. Thus, as Electronic Data Capture (EDC) has done for patient data collection, the CDM software framework's functionality should improve the efficiency of diagnostic model studies.