In conjunction with AIME'09, a full-day tutorial will be held on 19th July.

Introduction to Clinical Data Mining
John H. Holmes, PhD
University of Pennsylvania School of Medicine

This full-day tutorial will illustrate, via demonstration and hands-on experience, the application of data mining methodologies to a clinical database. A knowledge discovery life cycle model will be employed as the conceptual framework for the tutorial. Attendees will obtain practical experience in mining a database for use in clinical research, and ultimately for assisting with statistical analysis. We will focus on several well-known datasets for exploration in the tutorial, and we will learn and use the Weka data mining suite. Weka is freely available in the public domain, and runs on even modestly equipped computers within a Java runtime environment (JRE). Weka and the JRE will be distributed to attendees on CD-ROM free of charge. Attendees will be encouraged to bring laptops to the tutorial. Those who do not bring laptops will benefit from the detailed demonstrations in the tutorial.

We will focus on several families of data mining methodologies, including trees, clustering, Bayesian classification, evolutionary computation, visualization, and statistical classifiers. After a discussion of the general characteristics of biomedical data, such as missing values and feature selection problems, and methods for preparing biomedical data for mining, we will introduce examine examples of the selected families of tools for mining biomedical data, including thorough algorithmic descriptions, functional examples, and live demonstrations of each on several real-world biomedical datasets. The applications will focus specifically on rule discovery, emergence of clinical prediction rules, classification, and clustering, as appropriate to each method. The advantages and disadvantages of each method will be discussed in detail. The tutorial will also include a rigorous discussion of methods for evaluating the results obtained from mining biomedical data, including classification and prediction accuracy and test characteristics such as sensitivity, specificity, area under the receiver operating characteristic curve, and predictive values, the choice and use of suitable validation datasets, methods for comparing models, and the use of human expert panels in providing content for qualitative model validation.

Intended audience: Clinical and basic science researchers will benefit most from this tutorial. The content level is 50% beginner, 50% intermediate

Prerequisites: None, although some prior exposure to the basic methodologies of data mining may be helpful.