Preface

This volume contains the papers presented at HSDM20: ACM WSDM Workshop on Health Search and Data Mining held on February 3, 2020 in Houston.

There were 7 submissions. Each submission was reviewed by at least 2, and on the average 2.9, program committee members. The committee decided to accept 6 papers. The program also includes 2 invited talks.

There are many interesting challenges in delivering intelligent decision support in the health domain. Collections of documents such as health records, scholarly publications, clinical trials, or drug orders grow at high rates and are distributed around the globe in a fragmented manner. Health data is highly multi-modal (clinical notes, time series, medical images, genomics etc.) and its interpretation is domain specific. Users of health information systems have different levels of expertise, and information needs, e.g., a patient *vs.* a primary care physician *vs.* cancer researcher. At the same time, the data is highly sensitive and subject to legal requirements regarding privacy, security, and confidentiality. This breadth of challenges requires interdisciplinary approaches. The Information Retrieval (IR) and Data Mining (DM) communities are particularly well-positioned to tackle these problems.

Search, recommendation, and information extraction systems help lay and expert users explore ever-growing collections. Decision support systems assist in complex decision making processes. Intelligent user interfaces present the right information at the right time and allow for unobtrusive interaction all the way from the lab to the bedside. Mobile device applications and other sensors help provide a more holistic view on the patient's case than what can be gleaned in an 10-minute physician interview.

Health-related topics of interest include, among others:

- Search over images/genomics/structured data
- Federated multi-modal search combining different data sources
- User interfaces for biomedical/clinical search supporting complex information needs
- Analysis of search logs and social media
- User search behavior studies
- Building and use of medical knowledge bases or ontologies
- Privacy-preserving techniques for clinical data
- Adverse event detection and prediction
- Mobile (mHealth) applications
- Wearables
- Spoken interaction with health data
- Whole exposome modeling and estimation
- Applications of data mining and machine learning
- Ethics, bias, and fairness

This conference was sponsored by UPMC Enterprises.

February 3, 2020
Houston, Texas, USA

Yubin Kim
Carsten Eickhoff
Ryen White

Table of Contents

Machine Learning for Healthcare: Beyond i.i.d. Prediction 1
Zachary Lipton

Applying Information Retrieval to the Electronic Health Record for Cohort Discovery and Rare Disease Detection 2
William Hersh

Comparing Rule-based, Feature-based and Deep Neural Methods for De-identification of Dutch Medical Records 3
Jan Trienes, Dolf Trieschnigg, Christin Seifert and Djoerd Hiemstra

Healthcare NER Models Using Language Model Pretraining 12
Amogh Kamat Tarcar, Aashis Tiwari, Dattaraj Rao, Vineet Naique Dhaimodker, Penjo Rebelo and Rahul Desai

Lung nodule classification using Convolutional Autoencoder and Clustering Augmented Learning Method(CALM) 19
Soumya Suvra Ghosal, Indranil Sarkar and Issmail El Hallaoui

A Query Taxonomy Describes Performance of Patient-Level Retrieval from Electronic Health Record Data 27
Steve Chamberlin, Steven Bedrick, Aaron Cohen, Yanshan Wang, Andrew Wen, Sijia Liu, Hongfang Liu and William Hersh

Streaming Gait Assessment for Parkinson's Disease 34
Cristopher Flagg, Ophir Frieder, Sean MacAvaney and Gholam Motamedi

Clustering Large-scale Diverse Electronic Medical Records to Aid Annotation for Generic Named Entity Recognition 43
Nithin Haridas and Yubin Kim

Program Committee

Steven Bedrick	OHSU
Leonid Boytsov	3M M*Modal
Dina Demner	U.S. National Library of Medicine, NIH
Carsten Eickhoff	Brown University
Alba García Seco De Herrera	University of Essex
Yubin Kim	UPMC Enterprises
Bevan Koopman	CSIRO
Henning Müller	HES-SO
Joao Palotti	Qatar Computing Research Institute
Zhen Qin	Google
Kirk Roberts	The University of Texas Health Science Center at Houston
Karin Verspoor	The University of Melbourne
Wei Wei	University of Pittsburgh Medical Center
Ryen White	Microsoft
Elad Yom-Tov	Microsoft

Additional Reviewers

S

Sergeeva, Elena