Report on the Third International Workshop on Semantic Web Meets Health Data Management (SWH 2020)

Haridimos Kondylakis
ICS-FORTH
Heraklion, Greece
kondylak@ics.forth.gr

Kostas Stefanidis
Tampere University
Tampere, Finland
kostas.stefanidis@uta.fi

Praveen Rao
Univ. of Missouri-Columbia
Columbia, USA
praveen.rao@missouri.edu

ABSTRACT

Creating a holistic view of patient data comes with many challenges but also brings many benefits for disease prediction, prevention, diagnosis, and treatment. Especially in the COVID-19 era, this is more important than ever before. The third International Workshop on Semantic Web Meets Health Data Management (SWH) was aimed at bringing together an interdisciplinary audience who was interested in the fields of Semantic Web, data management, and health informatics. The workshop goal was to discuss the challenges in healthcare data management and to propose new solutions for the next generation of data-driven healthcare systems. In this article, we summarize the outcomes of the workshop, and we present a number of key observations and research directions that emerged from presentations.

1. INTRODUCTION

In recent years, precision medicine has received much attention in the U.S. It deals with the treatment and prevention of diseases by taking into account the genetic makeup, environmental and lifestyle factors of an individual [2]. As a result, medical professionals can precisely prevent and treat diseases rather than using a "one-size fits all" approach. Key in achieving the vision of precision medicine as well as affordable, less intrusive and more personalized care, is to efficiently and effectively harness the value of healthcare data to gain meaningful insights. Ultimately this has the potential to improve patient outcomes, increase the quality of life of patients, lower healthcare costs, and lower mortality. To conduct meaningful and large-scale precision medicine studies, a researcher needs data from different hospitals to be shared so that a large representative patient population can be analyzed using artificial intelligence (AI) and machine learning techniques leading to sound conclusions. However, healthcare systems suffer from *data siloes* due to the use of different software vendors, data models, and bureaucratic reasons.

A typical hospital manages patient data securely using an electronic health record (EHR) system. An EHR system uses a relational database system to store data about a patient's medical history, demographics, diagnosis, medications, allergies, lab test results, billing information, and others. Essentially, EHR data of patients are rich and complex and contain hundreds of attributes [1]. Applications such as Picture Archiving and Communication System (PACS) are used for storing imaging data (e.g., radiology images, histopathology images, mammograms). With the momentum around precision medicine, genome sequences of patients are produced and analyzed by large hospitals and are usually stored outside of the EHR system. In addition, healthcare data exists in many different formats, from textual documents (e.g., physician notes) and web tables to imaging modalities (e.g., Digital Imaging and Communications (DICOM)) to well-defined relational data and APIs. Clinical data may be represented in different ways in two different EHR systems. Hence, achieving semantic interoperability is challenging. Healthcare data can also be found on social media through healthcare conversations, in wearables and monitoring devices that continuously stream information about a person's fitness and health.

To overcome data siloes that plague existing healthcare systems and hinder new innovations in medicine, much effort has been spent in developing interoperability standards over the last few decades. Health Level Seven's Fast Healthcare Interoperability Resources (FHIR) [9] is emerging as a popular standard for healthcare data exchange and developing new applications. Clinical terminologies such as the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), International Classification of Diseases (ICD)-9, Logical Observation

Identifiers Names and Codes (LOINC), etc. are widely used to provide precise semantics to clinical/medical concepts. They serve as a reference system for comparing and aggregating information recorded by different entities in a healthcare system. While interoperability continues to be a challenge for healthcare data exchange and conducting large-scale clinical studies to advance the vision of precision medicine, Semantic Web technologies can provide effective solutions for enabling semantic interoperability and common language across healthcare systems or even increase the quality of health recommendations [13, 14, 15]. They can lead to the disambiguation of health information through the adoption of various terminologies and available ontologies. As a result, healthcare data can be precisely represented and integrated easily across diverse data sources while enabling large-scale AI/machine learning techniques to be applied for clinical decision-making.

Motivated by the aforementioned reasons, the goal of SWH 2020 was to bring together researchers with interests cross-cutting the fields of Semantic Web, data science, data management, and health informatics to discuss the challenges in healthcare data management and to propose novel and practical solutions for the next generation of data-driven healthcare systems. Developing optimal frameworks for integrating, curating and sharing large volumes of clinical data has the potential for a tremendous impact on healthcare, enabling better outcomes at a lower and affordable cost. The ultimate goal is to enable new innovations in Semantic Web, knowledge management, and data management for healthcare systems to move the needle to achieve the vision of precision medicine.

2. TOPICS

In the sequel, we present the various topics that the workshop focused through the various presentations and invited talks.

2.1 Data Harmonization

Public health data collection is difficult due to the small sample size and high cost involved. However, clinical data in health-related systems have far more coverage but have a higher degree of variety and limited validation. There is also the need for standardization. A typical data science pipeline on clinical data starts with data engineering, such as de-duplication [5], normalization, and using appropriate mappings. Then, the data store can be harvested using machine learning/deep learning, biostatistics, and bioinformatics tasks. The insights gained can be used for intervention, publication, predictive modeling, visualization, and so on.

Real-World Challenges. One of the harsh realities of aggregating data from multiple healthcare systems is that they are built differently by different vendors. In addition, the coding standards are adopted differently by different hospitals. For example, ICD-9 is still being used by some U.S. hospitals, while the majority of them have switched to ICD-10. Note that ICD-9 lacks specificity in defining diagnoses and procedures compared to ICD-10.[1] For example, ICD-9 uses the same code for "burn on the left arm" and "burn on the right arm". However, ICD-10 has two separate codes. Therefore, there is a need for better data harmonization albeit the many standards to effectively harness multi-site clinical data. Data scientists of such data will benefit by working closely with clinicians and domain experts rather than independently interpreting clinical data. To this direction, the Cerner Health Facts^{TM} (HF) is a de-identified dataset collected from 100 non-affiliated U.S. organizations (2010-2018) that contains 65M+ patients with billions of lab data and clinical events [8]. HF uses different clinical terminologies (e.g., ICD-9, ICD-10, LOINC) for standardization, used for different research studies, including opioid prescriptions for migraines.

Identifying Inconsistencies in SNOMED. Motivated by the presence of inaccurate representations in SNOMED, a new method was developed to analyze stop words in SNOMED concepts using lexical analysis to identify inconsistencies [3]. The method is able to identify missing hierarchical relationships and missing attribute relationships. A missing hierarchical relationship arises when a SNOMED concept, for example, "pneumonia and influenza" does not have as parents "pneumonia" and "influenza" in the hierarchy. A missing attribute relationship arises due to missing role groups for concepts such as "ornithosis with pneumonia" in SNOMED. In additional to detecting these inconsistencies, the method also suggests corrections. The method was able to identify 70% of missing attribute relationships and 28% of missing hierarchical relationships.

Annotating FHIR RDF Graphs. Motivated by the need to integrate and annotate medical data with additional knowledge for decision support systems, an approach was proposed to annotate RDF graphs constructed over Fast Healthcare Interoperability Resources (FHIR) [11]. Specifically, medication information is annotated to existing patient data. A distributed services architecture was de-

[1]https://www.ama-assn.org/media/7546/download

veloped to enable patients to store data in multiple FHIR sources. Medical ontologies were used to construct RDF graphs, which can then be processed by a semantic reasoner.

Drug Ontology Construction. The Mexican Drug Ontology [12] was constructed by the Secretary of Public Health in Mexico. The ontology is used to exchanging medical information between different information systems, to enable dynamic search of pharmaceutical information, compliant with the Mexican specifications that must be met by health service providers in Mexico. The Hermit reasoner is used to ensure that no contradictions are found in the ontology and also to enable query answering using Description Language (DL) querying. The ontology was developed using the Methontology methodology, whereas it was evaluated using a set of competency questions.

2.2 Closing the cycle from doctor to patient

The objective of an Information and Communication Technology (ICT) Health Ecosystem is to create a holistic view of patients data. However, closing the cycle from doctor to patients opens a number of technical, ethical and scientific challenges that are presented in the sequel.

Gathering user requirements. Gathering user requirements, in this process, is not an easy task, since medical doctors are extremely hard to find and gathering their needs takes a significant amount of time [7]. The aim is to understand what they need and propose them how their needs can be addressed by the systems that will be developed.

Regulating access to patient information. Another important thing is the issue of consent that dictates the extent to which patient data (clinical and genomic data) can be accessed and reused and by whom, inside and outside the context of a project, and for which purpose. Consent should set out potential risks and benefits to participants, as well the data anonymization procedures to be undertaken. The consent forms should be specified by the Data Protection Officers and their Legal Team. To do this, they should know the data flow and the processes implemented in the ICT Ecosystem. A result of this effort is the specification of the access control framework, i.e., who has access, and for what purpose and to what type of data. This task is not easy, since the medical doctors strongly believe that their patients' data is their data and should have total control over them.

Integrating with national registries. In addition, a patient-centric design is required that will support the integration with national registries and e-health systems [7]. Following such a design, there are benefits to the patient who does not need to carry his/her health record around to the healthcare system. Especially, in COVID-19 era, semantics and AI could highly contribute, along with personal health apps to the effective, secondary usage of available data. To this direction, the authors in [10] provide tools for public authorities, healthcare professionals, citizens and their families acting as a central point for COVID-19 management, whereas the available data are homogenized using an ontology. However applications like that face additional challenges, not only from the technological perspective but also from public health and governmental perspective as well, which might hamper the actual usage of the tool in practise.

2.3 Mining new knowledge

In nowadays, a big amount of healthcare data is used for providing analytics and for constructing applications useful for patients, thus opening a number of technical, ethical and scientific challenges.

Patient trajectories. Recently, challenges related to the management of personal and sensitive healthcare data have been addressed through decentralized solutions for patient data, often implemented and modelled using distributed agents and semantic technologies [4]. To this direction, ontologies are used to represent patient trajectories, agent-based architectures are employed to model decentralized patient data exchanges, whereas the agent cooperation and negotiation strategies designed for healthcare data interactions are implemented through multi-agent systems for real-time processing. Patients can delegate the management of personal trajectory data to dedicated agents, which in turn can automatically negotiate and cooperate with other agents, for instance, to share and aggregate anonymized data, to grant access to agents of medical staff, or to allow ML processing and prediction. Although the vision and the technological components are in place, the full workflow has not yet materialized into practice.

New knowledge from scientific documents. Finally, with the constant generation of new scientific papers and guidelines, the automatic extraction of newborn development content from scientific documents is essential. Such extracted knowledge can be used to recommend relevant advices and insights. An approach presented in [6] extracts concepts and relations between conceptual entities in the data sources by using existing techniques from natural language processing and cognitive computing, and

generates a knowledge base for the extracted information. The solution presented cannot replace the advice of a specialist, but it can provide quick information from reliable sources.

3. CONCLUSIONS

A number of key observations and research directions emerged in the discussions. For example, in order to achieve the vision of precision medicine, only focusing on homogenizing data access through semantics is only one side of the coin. There are many additional areas that the community should provide tangible solutions such as regulating access to patient information, consent management and overcoming the barriers imposed by national and international regulations.

To this direction, an important topic is also privacy protection on the collected information. Privacy protection has not only to do with the patient's right to manage their own data, but also with providing with necessary mechanisms for tracking their own data or even providing feedback, reusing it for their own benefit in a different context, or even exploring it through a set of third party services. Consent should not be a static document that is signed once but it should dynamically regulate access to health data provided by the individuals.

Furthermore, incomplete, inconsistent and erroneous entries in clinical terminology systems can be a bottleneck in acquiring new knowledge from health data. On the other hand, the detection of these inconsistencies is impractical and difficult due to the size of these terminologies. To this direction, further semi-automated solutions focusing on augmenting data quality are needed.

Finally, the amount of health information available online is enormous. As such, intelligently mining personal recommendations, tips and information tailored to the unique situation of an individual is highly important. Although usually insights and articles are selected manually by curators, advances in cognitive computing, natural language processing, ontology engineering and big data can lead to effective solutions.

Overall, the third instance of the Semantic Web Meets Health Data Management Workshop made clear that a lot of research work still needs to be done in the area of semantic health data management.

4. REFERENCES

[1] What information does an electronic health record (EHR) contain?, 2019. https://www.healthit.gov/faq/what-information-does-electronic-health-record-ehr-contain.

[2] What is precision medicine?, 2019. `https://ghr.nlm.nih.gov/primer/precisionmedicine/definition`.

[3] R. Burse, G. McArdle, and M. Bertolotto. Stop-word based contextual auditing to identify inconsistencies in SNOMED. In *SWH*, 2020.

[4] J. Calbimonte, D. Calvaresi, and M. Schumacher. Decentralized management of patient profiles and trajectories through semantic web agents. In *SWH*, 2020.

[5] V. Christophides, V. Efthymiou, T. Palpanas, G. Papadakis, and K. Stefanidis. An overview of end-to-end entity resolution for big data. *ACM Comput. Surv.*, 53(6), 2021.

[6] S. Consoli, K. Wouters, R. A. Otte, and A. Heinrich. A newborn development insights mining and recommendation system from scientific literature and clinical guidelines. In *SWH*, 2020.

[7] I. Fundulaki, T. Saveta, V. Papakonstantinou, and Y. Roussakis. An ecosystem for precision medicine: Closing the cycle from doctor to patient. In *SWH*, 2020.

[8] E. F. Glynn and M. A. Hoffman. Heterogeneity introduced by EHR system implementation in a de-identified data resource from 100 non-affiliated organizations. *JAMIA Open*, 2(4):554–561, 08 2019.

[9] HL7. Fast healthcare interoperability resources, 2019. `http://hl7.org/fhir`.

[10] D. G. Katehakis, A. Kouroubali, G. Kavlentakis, N. Stathiakis, F. Logothetidis, Y. Petrakis, V. Tzikoulis, S. Kostomanolakis, and H. Kondylakis. Data management, semantics and personal health apps for staying safe in COVID-19. In *SWH*, 2020.

[11] G. Kober. Annotating FHIR-RDF-graphs with medication knowledge. In *SWH*, 2020.

[12] C. R. Peña, M. T. Vidal, M. Bravo, and R. Motz. Drug ontology for the public mexican health system. In *SWH*, 2020.

[13] M. Stratigi, H. Kondylakis, and K. Stefanidis. Fairness in group recommendations in the health domain. In *ICDE*, 2017.

[14] M. Stratigi, H. Kondylakis, and K. Stefanidis. Fairgrecs: Fair group recommendations by exploiting personal health information. In *DEXA*, 2018.

[15] M. Stratigi, H. Kondylakis, and K. Stefanidis. Multidimensional group recommendations in the health domain. *Algorithms*, 13(3):54, 2020.