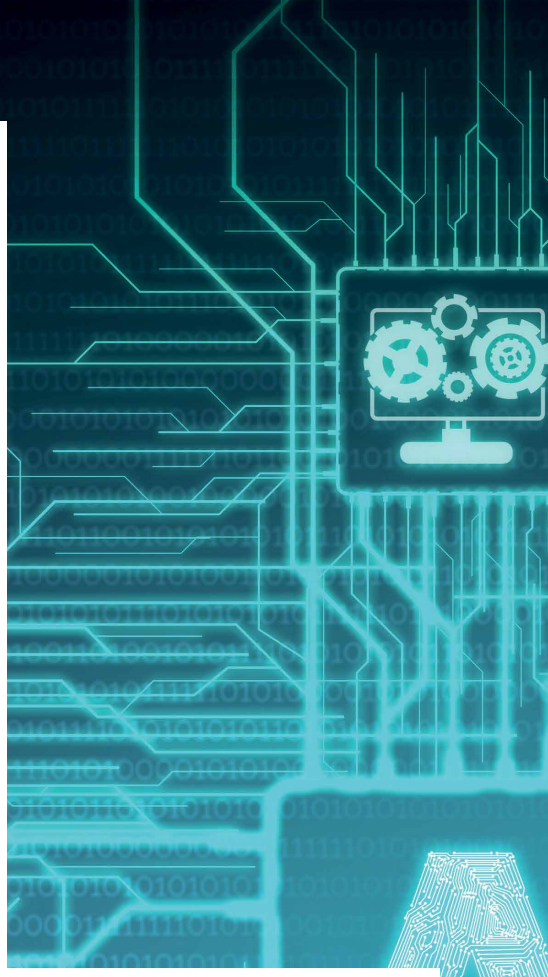
Faceted Query Systems for the Management of Clinical Data

Dr Guo-Qiang Zhang

Scientia

FACETED QUERY SYSTEMS FOR THE MANAGEMENT OF CLINICAL DATA



Scientists and clinicians rely on data to inform their practice and make decisions in a variety of medical settings. For data to be meaningful they need to be translated into actionable information and interpreted by the user. Access to a sheer amount of data can, in itself, pose a challenge. **Dr Guo-Qiang 'GQ' Zhang** from the University of Texas Health Science Center at Houston (UTHealth) has developed several innovative systems that provide a user-friendly interface for handling large-scale, multi-centre clinical data.

Simplifying Access to Healthcare Registries

Health research in the 21st century has become increasingly data-driven. Computer-aided exploration allows researchers to generate hypotheses and to share their findings more rapidly than ever before. However, the volume and complexity of the data generated and shared in our hyper-connected world grow at such a pace that traditional approaches are often inadequate at handling them.

Dr Guo-Qiang 'GQ' Zhang is the Vice President and Chief Data Scientist for the University of Texas Health Science Center at Houston (UTHealth). Dr Zhang and his collaborators aim to develop user-friendly query engines that simplify the process of clinical data management. The vision of Dr Zhang and his team of researchers is to enable users to interact with data in real-time, almost effortlessly, with browsing suggestions and contextual feedback immediately displayed at the query level.

Dr Zhang and his team are inspired in their effort by the way in which online shopping works for the general population. The type of interface used by online shopping websites allows users to narrow down their searches from numerous items to just several options of interest. For example, when buyers browse online for a pair of shoes, they can quickly find what they are interested in by filtering for simple attributes or facets, such as colour, brand, size and price range. This type of approach at handling big data is known as 'faceted search', and applying it to the interrogation of clinical data about complex human conditions requires improving the organisation of existing databases through their annotation using codified knowledge (also known as ontologies).

Dr Zhang has a track record spanning over a decade of contributions to biomedical data science. When working at Case Western Reserve University and the University of Kentucky, Dr Zhang and his research group developed the National Sleep Research Resource (NSRR), a comprehensive and easily accessible repository of sleep data.

The repository was built by integrating data from 10 large studies funded by the National Institutes of Health in the USA. The NSRR allows researchers to share their data and make these readily analysable by others by providing information on the source of the data, the time point at which they were collected, and the equipment and methodology used for the collection. A case-control interface also allows registered users to specify a case cohort and a control cohort for discovering viable scientific hypotheses.

The NSSR is currently the world's largest and richest data-sharing system in the field of sleep research. It provides a free single point of access to large sleep polysomnogram datasets, enabling researchers to investigate the impact of sleep disorders on important clinical outcomes.

In a study published in 2014, Dr Zhang and his collaborators introduced MEDCIS, with an ontology-driven patient information capturing system aimed at facilitating data sharing for epilepsy clinical studies. MEDCIS is an intuitive and integrated system that makes extensive use of multi-level drop-down menus, reducing the possibility of data entry errors and variability in the use of epilepsy terms.

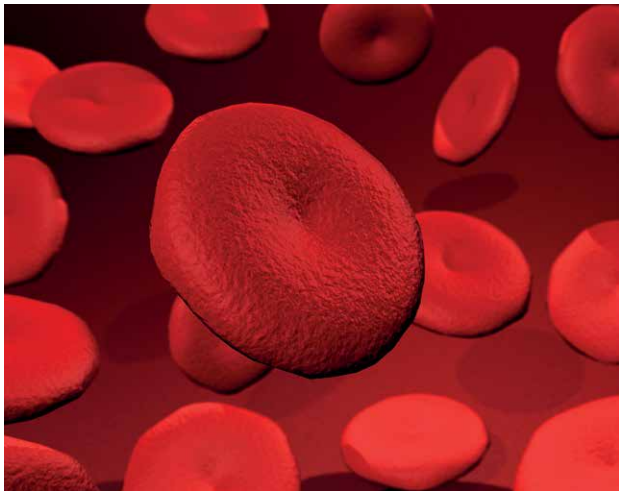
MEDCIS was developed to support the prevention and risk identification of sudden unexpected death in epilepsy (SUDEP). This line of research investigation resulted in a recent five-year grant from the National Institutes of Health to study SUDEP risk markers in the prospective data managed through MEDCIS. This study will ultimately lead to individualised, evidence-based, SUDEP risk assessment tools that help clinicians and patients manage potentially modifiable risks, leading to overall reduced SUDEP mortality and improved epilepsy patient care.

A User-Friendly Interface to Aid Cancer Data Exploration

Dr Zhang and his colleagues are currently working on the development of OncoSphere, a faceted query engine that will improve the interrogation of data available from cancer registries. Cancer data have been systematically collected to allow investigators and policymakers to access information on incidence and mortality. However, these software tools for accessing cancer registries do not support sophisticated data exploration. Dr Zhang's research team hopes to bridge the gap by developing an interface that allows professionals to easily identify patient cohorts for clinical trials and epidemiological studies.

The traditional workflow in cancer research involves making a hypothesis before the data are collected and analysed. OncoSphere will facilitate a more desirable workflow that starts with data exploration prior to the generation of a hypothesis. The system will streamline data access for several query modalities, allowing researchers to group patients in cohorts based on similar medical histories or to identify disparities in outcomes among different patient populations. Importantly, the system will assist with the identification of patients who might benefit from personalised cancer care and treatment programs.

OncoSphere is sponsored by the USA National Cancer Institute (NCI) and the system will anchor its query interface based on the NCI Thesaurus, a terminology reference system used to organise and annotate clinical oncology data. Dr Zhang and his group argue that the NCI Thesaurus is not, as it currently stands, a good fit for a faceted query system because it lacks the quality requirement on its structure to serve this new role. The team is working at improving the NCI Thesaurus, enabling it to fit this new interface role and facilitate hypothesis generation in the cancer domain in more robust and efficient ways.



Identifying Missing or Incorrect Relations

With the development of OncoSphere, Dr Zhang and colleagues hope to make a difference to the way in which clinical scientists around the world experience web exploration. The performance of the query engine will depend on its capability for producing search results that are complete and sound. The property of completeness relates to the ability of the interface to generate all the possible results that are relevant to a particular clinical scenario. An incomplete faceted search will yield results that omit important medical records. Soundness, on the other hand, means that all results from a faceted search are relevant to the query, with no room for incorrect entries.

Dr Zhang's team works very hard to identify potential missing and incorrect relations, such as those in the NCI Thesaurus. They reported, as an example, that when searching for 'neoplastic large T-lymphocytes' (white blood cells that are growing uncontrollably leading to a tumour of the blood), the NCI Thesaurus failed to include 'anaplastic T-lymphocytes' (malignant blood cells that have lost their usual shape and functions) as a subtype. As a consequence of this incomplete facet, patients with anaplastic T-lymphocytes would not be included in a cohort of patients with neoplastic large T-lymphocytes.

The team also pointed to instances of problematic relations that may cause the query engine to generate erroneous results. OncoSphere will be instrumental in the identification of missing or invalid relations in the NCI Thesaurus, improving its structural organisation and supporting its new role for the faceted query.

Building on Responsiveness and Expressiveness

A good query interface is designed to empower human-data interaction. With this in mind, Dr Zhang's team aims to optimise the usability of OncoSphere in terms of responsiveness and expressiveness. Interface responsiveness is the ability of the query engine to not only execute queries in a speedy manner but also to interact with the user in near real time. OncoSphere achieves this by integrating the use of checkboxes and web widgets with a mouse hovering function that displays search suggestions instantaneously. The expressiveness of an interface relates to its ability to support a broad range of queries. To achieve this, the team is collaborating with the University of Kentucky's Markey Cancer Center (MCC). Investigators at MCC will have access to OncoSphere and will formulate queries on a broad range of categories. MCC members will then be asked to submit anonymous comments on the usability of the system and to suggest improvements.

Future Directions

The preliminary evaluation of OncoSphere will ascertain the degree to which it meets the design objectives, before the system can be tested by a larger number of users at a later stage. The plan for the future is to use crowdsourcing to assess the faceted search capabilities of OncoSphere at full scale, allowing it to become an essential resource for the cancer research community.

Dr Zhang and his colleagues will also continue to engage with the epilepsy community and they aim to expand the collection of clinical records from an increased number of patients in the coming years. They are working closely with their collaborators to broaden the sharing of data in order to advance understanding of the biological mechanisms behind SUDEP. The team also hopes to further develop its NSRR repository to aid sleep research.

More efforts are needed in developing a system that can break large, unstructured data files into minimal fragments that can be indexed on the fly. Dr Zhang and his colleagues are continuing their research efforts in facilitating cohort discovery by enhancing ontology exploration, query management and query sharing for large clinical data repositories.

Meet the researcher

Dr Guo-Qiang Zhang
University of Texas Health Science Center at Houston
Houston, TX
USA

Dr Guo-Qiang 'GQ' Zhang received his PhD in Computer Science from the University of Cambridge. His early research interests included theoretical computer science and the semantics of programming languages. Dr Zhang is now Vice President and Chief Data Scientist for the University of Texas Health Science Center at Houston (UTHealth), one of the six health science campuses of the University of Texas System. He also serves as a Co-Director of the newly established Texas Institute for Restorative Neurotechnologies. Before joining UTHealth, he was Professor of Internal Medicine and Computer Science at the University of Kentucky. Over the last decade, Dr Zhang's research has revolved around Human-Data Interaction, achieved through the development of innovative software and clinical informatics applications. Dr Zhang led the development of the data resources for the National Sleep Research Resource and the Center for Sudden Unexpected Death in Epilepsy Research. He also has a track record of research in biomedical metadata including ontologies and terminology systems. Dr Zhang uses cutting-edge computer science and informatics methodology to effectively address biomedical data/big data challenges through the translation of theory, algorithms, methods and best practices into functional and usable tools impacting the clinical research data ecosystem.

CONTACT

E: Guo-Qiang.Zhang@uth.tmc.edu
W: https://www.uth.edu/tirn/index.htm
https://www.uth.edu/data

KEY COLLABORATORS

Samden Lhatoo, UTHealth
Licong Cui, UTHealth
Shiqiang Tao, UTHealth

FUNDING

Center for SUDEP Research, NIH U01NS090408, U01NS090405
National Sleep Research Resource, NIH R24HL114473
Ontology-driven Faceted Query Engine, NIH R21CA231904
An informatics framework for SUDEP Risk Marker Identification and Risk Assessment, NINDS R01NS116287
The Kentucky Research Informatics Cloud, NSF ACI1626364

FURTHER READING

GQ Zhang, S Tao, N Zeng, L Cui, Ontologies as nested facet systems for human-data interaction, Semantic Web, 2020, 11(1), 79–86.

GQ Zhang, L Cui, R Mueller, et al, The National Sleep Research Resource: Towards a sleep data commons, Journal of the American Medical Informatics Association, 2018, 25(10), 1351–1358.

GQ Zhang, G Xing, L Cui, An efficient, large-scale, non-lattice-detection algorithm for exhaustive structural auditing of biomedical ontologies, Journal of Biomedical Informatics 2018, 80, 106–119.

L Cui, W Zhu, S Tao, et al, Mining Non-Lattice Subgraphs for Detecting Missing Hierarchical Relations and Concepts in SNOMED CT, Journal of the American Medical Informatics Association, 2017, 24(4), 788–798.

S Tao, L Cui, GQ Zhang, Facilitating cohort discovery by enhancing ontology exploration, query management and query sharing for large clinical data repositories, AMIA Annual Symposium Proceedings, 2017, 1685–1694.

GQ Zhang, L Cui, SD Lhatoo, et al, MEDCIS: Multi-Modality Epilepsy Data Capture and Integration System, AMIA Annual Symposium Proceedings, 2014, 1248–1257.

UTHealth
The University of Texas
Health Science Center at Houston