

## Tian Kang

Ph.D. candidate

Department of Biomedical Informatics, Columbia University

E-mail: [tk2624@cumc.columbia.edu](mailto:tk2624@cumc.columbia.edu) Phone: 917-288-9570

Address: 622 West 168 Street, New York, NY 10032

---

### DISSERTATION SUMMARY

---

My Ph.D. dissertation title is “**Machine Reading Comprehension for Medical Evidence**”. The research aims to bridge the gaps of information need for Evidence-based Medicine in a timely manner by developing a novel attention model for Machine Reading Comprehension that is interpretable and informed by extracted domain knowledge, and eventually enabling efficient question answering for clinical decision making.

---

### EDUCATIONS

---

<b>Ph.D. in Biomedical Informatics</b> , Columbia University, 2021 (anticipated), advised by Dr. Chunhua Weng	GPA 3.90/4.00
<b>M.A. in Biomedical informatics</b> , Columbia University, 2016	GPA 3.76/4.00
<b>B.S. in Bioinformatics</b> , Huazhong University of Science and Technology, China, 2014	GPA 3.50/4.00

---

### RESEARCH EXPERIENCE

---

#### **A PICO-informed Self-Attention model for Machine Reading Comprehension for medical evidence.**

Department of Biomedical Informatics, Columbia University, March 2020 – now

- Developing a novel attention model for Machine Reading Comprehension that is interpretable and informed by extracted domain knowledge.

#### **Improving PICO Element Recognition by UMLS-based Data Augmentation**

Department of Biomedical Informatics, Columbia University, Dec. 2019 – March 2020

- Developing a data augmentation method, called UMLS-EDA, by simple text transformation and incorporation of domain knowledge such as UMLS, is effective in helping improve both PICO element recognition tasks in both entity and sentence level.
- Proposed UMLS-EDA improved the biLSTM-CRF model significantly. By adding BERT into the model, the model achieved state-of-the-art performance.

#### **Knowledge Representation of Medical Evidence based on PICO framework for PubMed Literatures**

Department of Biomedical Informatics, Columbia University, Nov. 2018 – now

- Developing an information extraction system for fully standardizing medical abstracts (Named Entity Recognition, study arm resolution, relation identification, etc.) to facilitate Evidence-Based Medicine and information retrieval for PubMed.

#### **Cataloging and mapping treatments for patients from an Online Autism Community**

Department of Biomedical Informatics, Columbia University, Sep. 2016 – Jan. 2017

- Applied NLP to understand the treatments used in real world for Autism Spectrum Disorder using patient-generated text from online health communities and compared it to treatment guidelines.
- Wrote the manuscript and presented it in 2017 WWW conference.

#### **EliIE (Free-text Eligibility Criteria Information Extraction System): A machine learning-based IE system to formalize clinical research eligibility criteria into OMOP Common Data Model**

Department of Biomedical Informatics, Columbia University, Aug. 2015 – Apr. 2016

- Lead an annotation team, designed annotation guidelines and created training data of clinical trial eligibility criteria.
- Developed an IE system called EliIE applying NLP techniques and machine learning approach to parse and formalize clinical trial eligibility criteria in order to facilitate trial recruitment.(GitHub repository: <https://github.com/Tian312/EliIE>)
- Wrote the manuscript and presented the paper at JAMIA Journal Club Invited Talk.

#### **Detecting Speculations and Negations in Chinese Clinical Notes using Character and Word Embedding**

Department of Biomedical Informatics, Columbia University, Aug. 2015 – Dec. 2015

- Explored how Word Embedding can improve Chinese clinical NLP tasks. Both speculation and negation detection results achieved the state of the art when the project completed.
- Wrote both manuscripts.

### Readability Assessment of Clinical Trial Eligibility Criteria

Department of Biomedical Informatics, Columbia University, Oct. 2014 – Feb. 2015

- Assessed readability of the world's largest online clinical trial registry – [clinicaltrials.gov](http://clinicaltrials.gov) using NLP techniques to evaluate its accessibility to health consumers.
- Wrote the manuscripts and presented the paper in AMIA 2015.

---

## PUBLICATIONS

---

- **Kang, T.**, Zou, S. and Weng, C., 2019. Pretraining to Recognize PICO Elements from Randomized Controlled Trial Literature. *Studies in health technology and informatics*, 264, pp.188-192.
- Wei, D.H., **Kang, T.**, Pincus, H.A. and Weng, C., 2019. Construction of Disease Similarity Networks Using Concept Embedding and Ontology. *Studies in health technology and informatics*, 264, pp.442-446.
- Rogers, JR, Callahan, TJ, **Kang, T.**, Bauck, A, Khare, R, Brown, JS, Kahn, MG and Weng, C 2019 A Data Element-Function Conceptual Model for Data Quality Checks. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 7(1): 17, pp. 1–14.
- Yuan, C., Ryan, P.B., Ta, C., Guo, Y., Li, Z., Hardin, J., Makadia, R., Jin, P., Shang, N., **Kang, T.** and Weng, C., 2019. Criteria2Query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association*.
- Butler, A., Wei, W., Yuan, C., **Kang, T.**, Si, Y. and Weng, C., 2018. The Data Gap in the EHR for Clinical Research Eligibility Screening. *AMIA Summits on Translational Science Proceedings*, 2017, p.320
- Sen, A., Goldstein, A., Chakrabarti, S., Shang, N., **Kang, T.**, Yaman, A., Ryan, P.B. and Weng, C., 2017. The representativeness of eligible patients in type 2 diabetes trials: a case study using GIST 2.0. *Journal of the American Medical Informatics Association*.
- Zhang, S., **Kang, T.**, Qiu, L., Zhang, W., Yu, Y. and Elhadad, N. and Weng, C., 2017, April. Cataloguing Treatments Discussed and Used in Online Autism Communities. *In Proceedings of the 26th International Conference on World Wide Web (pp. 123-131)*. International World Wide Web Conferences Steering Committee.
- **Kang, T.**, Zhang, S., Tang, Y., Hruby, G.W., Rusanov, A., Elhadad, N. and Weng, C., 2017. EliIE: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association*, 24(6), pp.1062-1071.
- **Kang, T.**, Zhang, S., Xu, N., Wen, D., Zhang, X. and Lei, J., 2017. Detecting negation and scope in Chinese clinical notes using character and word embedding. *Computer methods and programs in biomedicine*, 140, pp.53-59.
- Zhou, F., Cao, H., Zuo, X., Zhang, T., Zhang, X., et al., 2016. Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease. *Nature genetics*, 48(7), p.740.
- Zhang, S., **Kang, T.**, Zhang, X., Wen, D., Elhadad, N. and Lei, J., 2016. Speculation detection for Chinese clinical notes: impacts of word segmentation and embedding models. *Journal of biomedical informatics*, 60, pp.334-341.
- **Kang, T.**, Elhadad, N. and Weng, C., 2015. Initial readability assessment of clinical trial eligibility criteria. *In AMIA Annual Symposium Proceedings (Vol. 2015, p. 687)*. American Medical Informatics Association.

---

## CONFERENCE PRESENTATIONS

---

**Paper** *Pretraining to Recognize PICO Elements from Randomized Controlled Trial Literature*. MEDINFO 2019, Lyon, France.

**Paper** *Cataloguing treatments discussed and used in online autism communities*, 2017 International World Wide Web Conference (WWW), Perth, Australia.

**Paper** *We make choices we think are going to save us": Debate and stance identification for online breast cancer CAM discussions*, 2017 International World Wide Web Conference (WWW), Perth, Australia.

**Paper** *Initial Readability Assessment of Clinical Trial Eligibility Criteria*, AMIA Annual Symposium 2015, San Francisco, CA.