# YIRU GONG

100 Haven Avenue, New York, NY 10032 | (917) 742-0036 | yg2832@cumc.columbia.edu | LinkedIn

## EDUCATION

**Columbia University Mailman School of Public Health** *New York, NY*
Master of Science (MS), Biostatistics; GPA: 3.85 09/2021 – 05/2023 (Expected)

- Public Health Data Science Track; Relevant Courses: *Deep Learning, Introduction to Databases, Natural Language Processing, Cloud Computing, Data Science II, Graphical Models For Complex Health Data*

**University of Edinburgh - Zhejiang University Joint Institute** *Edinburgh, UK & Zhejiang, China*
Bachelor of Science with Honors in Biomedical Science, Dual degree program; GPA: 3.82 09/2017 – 06/2021

- Grant funding: 2020 Overseas-Exchange Scholarship Award of ZJE Institute (15000 RMB)

## SKILLS

- **Programming:** R, Python, SQL and MATLAB; Bash, Linux Shell, SAS, PASS, C language, and VBA
- **Software Development:** frontend: JavaScripts, TypeScripts, Angular, Jinjia, HTML/CSS; backend: Python Flask; database: MySQL, SQLite, PostgreSQL, Microsoft SQL Server Management Studio (SSMS)
- **Toolkits:** MS Office software, Tableau, Google Cloud Platform (GCP), Amazon Web Services (AWS), Spark, Latex, Docker, Container, R Shiny app, Git, Numpy, Pandas, sklearn, Pytorch, Keras, nltk

## EXPERIENCE

**Columbia University, *Research Assistant, Center of Patient Safety Research*** New York, US, 10/2022-now

- Established new relational databases in MS SQL Server to store electronic medical records and keeping track of wrong patient errors generated by clinicians in daily use
- Built user-friendly python-based SQL database interface to extract measurements of wrong patient errors

**Elevance Health, *Health Data Analytics Intern, AIM Specialty Health*** Remote, US, 06/2022-08/2022

- Built a machine learning model (LightGBM) on Lumber MRI medical claim data to predict case approval status and reached goals of automated approval for 10% of cases in Python; improved performance and get 86% in precision
- Merged and pulled >60k records of raw claim data from multiple datasets using SQL in the Microsoft SQL Server
- Refined a Natural Language Processing (NLP) Model by communicating with medical doctors for model validation

**Columbia University, *Research Assistant, Data Science Institute*** New York, US, 03/2022-08/2022

- Created an automated and highly efficient statistical analyzing pipeline in R and Python for genetic analysis
- Embedded entire scripts into a docker image to allow large-scale computing in cloud computing platforms

**GlaxoSmithKline (GSK), *Digital Analyst Intern, R&D Tech*** Shanghai, China, 04/2021-07/2021

- Initiated two projects in applying Natural Language Processing (NLP: NER, RE) to extract and compare features of clinical trials for competitor identification, and to achieve auto-revision of medical writing
- Reduced clinical teams' time and effort on medical document translation and revision by 50%
- Presented 6-8 clinicians on principles of NLP and AI to extract disease-related information from > 10 million research articles and FDA documents in the Neo4j-based Medical Knowledge Graphs (KG) interface
- Designed a database to support automated drug pharmacokinetics (PK) analysis and visualized in R shiny app
  - Tidied semi-structured JSON data of > 400k FDA clinical trials by Elasticsearch in Python
  - Relieved manual workload from 2 months to 2 days by R and Python

## PROJECTS

**Cloud Computing Project: *Building a Full-Stack Website App for Badminton Appointment System*** 09/2022-12/2022

- Established a website with Angular-based typescript-HTML UI interface (frontend), Flask based Python microservices frameworks and API development (backend), and MySQL databases to build a badminton court reserving system on Amazon Web Services (AWS: RDS, EC2, Beanstalk, CSS, API Gateway)

**Humana-Mays Healthcare Analytics Case Competition: Analysis of Housing Insecurity** 08/2022-10/2022

- Led a team of four to establish a machine learning model (XGBoost) to predict the housing insecurity status based on >40k Humana medical claim data with 865 variables; applied parameter tuning to achieved 73% in AUC score
- Authored a business analytic report to transform model findings into business insights

**CodaLab Competition: Covid-19 Infection Percentage Estimation** 01/2022-03/2022

- Established a Convolutional Neural Network (CNN) model to estimate the Covid-19 infection percentage from the CT scans of Covid-19 patients; Achieved a 7.2 MAE score and ranked in the top 20% among all participants

**Deep Learning Project: *Applying Masked Token Transformer (MaskGIT) in Audio Generation*** 01/2022-05/2022

- Applied a Computer Vision Model (Google MaskGIT) to audios utilizing the Pytorch package in Python, built the model by combining an audio-pre-trained VQGAN model and the MaskGIT transformer; trained on GCP
- Authored a research article and a conference-style poster as final reports; received 90/100 (first class)