A conversation with Kashyap Tumkur: The promise and challenges of precision medicine

INNOVATION LEADERS

by Bushra Anjum

Editor's Introduction

Ubiquity's senior editor Dr. Bushra Anjum chats with Kashyap Tumkur, a software engineer at Verily Life Sciences, the healthcare and life sciences arm of Alphabet. They discuss how the notion of "precision medicine" has gained popularity in recent times. Next, the focus turns to Tumkur's work, where he, along with his team, is working on collecting and integrating continuous time-series data to create a map of human health.

A conversation with Kashyap Tumkur: The promise and challenges of precision medicine

INNOVATION LEADERS
by Bushra Anjum

Kashyap Tumkur is a software engineer at Verily Life Sciences, the healthcare and life sciences arm of Alphabet. Previously, he was a graduate student researcher at the Department of Bioinformatics at the University of California, San Diego, where he earned his master's in computer science with a specialization in artificial intelligence. Kashyap seeks to promote equitable computing applications and broader access to computing opportunities and is a member of the ACM's Future of Computing Academy and US Technology Policy Committee, and a Global Shaper, Oakland Hub, of the World Economic Forum. Kashyap can be reached on Twitter [@nonlogic](https://twitter.com/nonlogic).

What is your big concern about the future of computing to which you are dedicating yourself?

We're currently in the middle of a computing revolution for healthcare and clinical research. Driven by the promise of big data, the notion of "precision medicine" has gained popularity in recent times. Precision Medicine considers a patient's medical or test history in addition to their genetic, environmental, and lifestyle data. This helps to obtain a fuller understanding of the patient's condition and to provide the best-personalized treatment to them. The availability of large, rich datasets and advances in machine learning and artificial intelligence also herald a new era of computer-aided diagnostics, motivated by early successes like [AI that screens for diabetic retinopathy](#).

We're not quite there yet, though. Data must be digitized, cleaned, integrated, and labeled to enable this kind of AI - the diabetic retinopathy team specifically called out how much effort they spent in curating their dataset. Even data collected routinely in hospitals, like your medical record, is generated in one of several formats via systems that don't necessarily integrate well. Chances are that this data stays in your hospital's data warehouse or in that of an intermediate medical data aggregator. There, it's usually siloed and doesn't get integrated with other potentially useful data streams. If you're talking about a smaller hospital or test laboratory, the situation gets worse: e.g., the *New York Times* did a [recent feature](#) on how vast numbers of Covid-19 test results are being transmitted from laboratories to hospitals over fax, generating reams of paper that may not be processed in time to be actionable.

However, the industry is making progress in policy and technology. The adoption of electronic health records (EHRs) was driven in part by the HITECH act, and technical standards like [FHIR](#), [aided by the MyHealthEData initiative](#), simplify EHR interoperability. Still, inconsistent data collection and data fragmentation remain significant obstacles to delivering on the promise of precision medicine.

Consumers also need to have confidence in the privacy safeguards of these systems before adopting them. This is also true for developers building new tools and applications that interact with these systems. Some of the high-stakes tasks to be tackled here are securing data, tracking data history or provenance, audit logging accesses, and tracking user consent for sharing. We also need standardized implementations of interfaces that appropriately limit applications to subsets of data, such as aggregates, models, or anonymized data based on context across treatment, support, and research.

How did you get introduced to the obstacles of data collection and fragmentation in healthcare systems?

We've all noticed that healthcare systems aren't the most user-friendly, like when we've switched doctors and found our medical history didn't switch with us, or when we've tried to obtain the results of a lab test.

During my work as a graduate student at UCSD, I realized this situation wasn't any better on the clinical research side of things either. Genome-Wide Association Studies (GWAS) are population studies to find associations between genetic variations and specific phenotypes or perceivable traits in a person. This helps understand the path of a disease or determine an individual's risk factor for it. The [NHGRI GWAS Catalog](#) is a public database of known associations for a specific population, genetic variation, or disease, but is maintained manually by epidemiologists who curate association information from GWAS publications and add to the catalog. To speed up this process while simultaneously requiring less time and attention from these highly qualified people, we built a natural language processing system that used [weakly supervised learning to automatically extract information about associations](#) from GWAS publications. Along the way, we realized many of these publications were only available as PDFs, and a significant chunk of our work was focused on a machine-learning pipeline to extract text from them, and this data extraction turned out to be an equally challenging problem.

Of course, it also helps that my father, a practicing doctor, regularly reminds me that he finds it easier to look up CPT codes—which represent diagnostic or treatment procedures in a medical record—via Google Search instead of his medical system. He also wonders when Alphabet will build a medical system he can use instead! [laughing]

What novel healthcare applications are you currently working on to realize some of the promises of precision medicine?

A key component of precision medicine is collecting and organizing environmental and lifestyle data, which makes new datasets available for evidence generation, leading to a better understanding of the path of disease and potential treatments. This data may originate from health apps on your phone or a wearable device like a smartwatch. This fundamentally novel health data is rich, continuous (and hence very dense), and specific in structure and content to each application.

My primary work focuses on this continuous time-series data. Being novel, they don't necessarily fit into existing standards like FHIR. So, one technical challenge has been developing interfaces and schemas that support time series data for many different sources, formats, and destinations. For example, how can this streaming data be efficiently stored in a database and then quickly represented in tools like BigQuery for analysts? At the same time, these interfaces and schemas must be standardized enough to be meaningfully combined with other data streams to support evidence generation and clinical trials, as in Verily's [Baseline Platform](#). My team's work on resolving data integration challenges like these is enabling us to create "a map of human health" in [Project Baseline](#) and features like [irregular heartbeat monitoring](#) in the Study Watch.

In the long term, this kind of continuous health measurement has the potential to revolutionize medicine from a reactive approach, where you go to the doctor when you perceive a symptom, to a more proactive approach, where you become aware even before you start feeling ill. With a large amount of activity and several players [working in this domain](#), I'm optimistic for a drastically changed healthcare paradigm over the next decade.

If you're working on novel healthcare applications or looking to learn more about this messy, regulated, but rewarding intersection of computing and health, and I can help in any way, please reach out at kashyaptumkurATacmDOTorg.

Biography

Bushra Anjum is a software technical lead at Amazon in San Luis Obispo, CA. She has expertise in Agile Software Development for large scale distributed services with special emphasis on scalability and fault tolerance. Originally a Fulbright scholar from Pakistan, Dr. Anjum has international teaching and mentoring experience and has served in academia for over five years before joining the industry. In 2016, she has been selected as an inaugural member of the [ACM Future of Computing Academy](#), a new initiative created by ACM to support and foster the next generation of computing professionals. Dr. Anjum is a keen enthusiast of promoting diversity in

the STEM fields and is a mentor and a regular speaker for such. She received her Ph.D. in computer science at the North Carolina State University (NCSU) in 2012 for her doctoral thesis on Bandwidth Allocation under End-to-End Percentile Delay Bounds. She can be found on Twitter [@DrBushraAnjum](https://twitter.com/DrBushraAnjum).

DOI: 10.1145/3466883