HYPERION RESEARCH

HPC User Forum Update

AI Making a Difference at NYU Langone Health

Thomas Sorensen
June 2024

IN THIS UPDATE

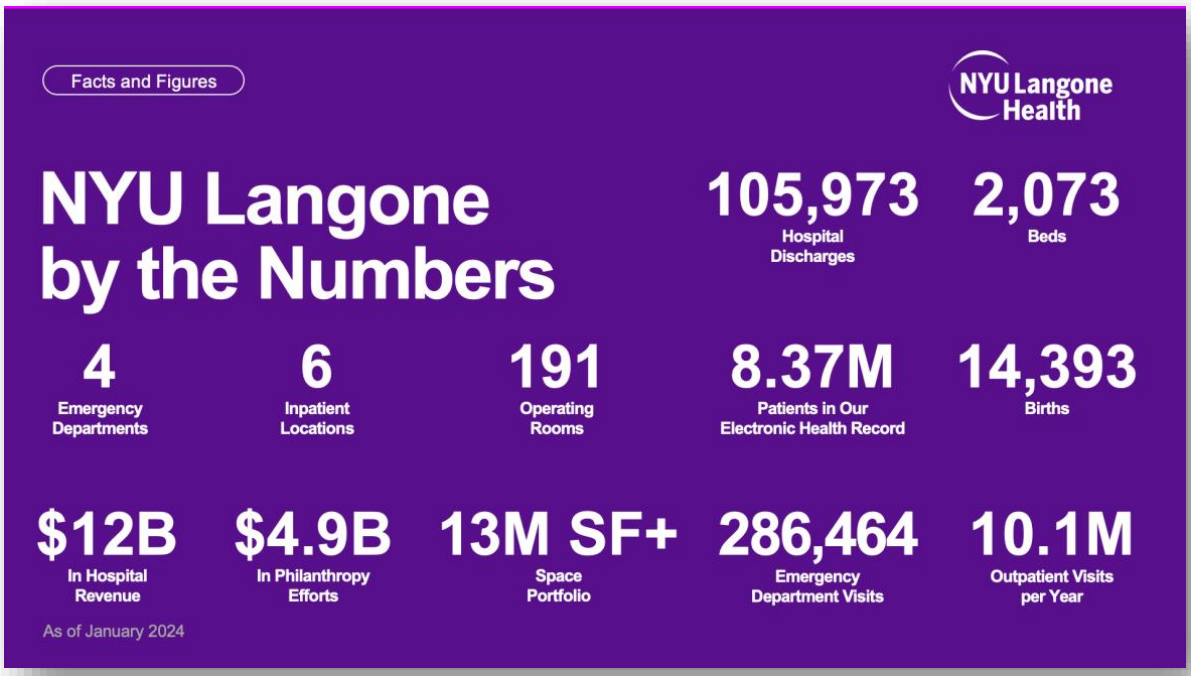
The HPC User Forum was established in 1999 to promote the health of the global HPC industry and address issues of common concern to users. In April 2024, the 84th HPC User Forum took place in Reston, Virginia. This update summarizes a presentation from that conference given by John Speakman, Assistant Vice President for Research and Education Information Technology at NYU Langone. He leads the provision of information and technology for the academic (scientific and educational) missions of NYU Langone Health, advancing the vision of the institute's digital transformation. His discussion at the April 2024 meeting covered the HPC resources built and leveraged at the university and healthcare system as well as details about their newly integrated AI tools.

Source: John Speakman, 2024

PRESENTATION: AI AND LLM AT NYU LANGONE BY JOHN SPEAKMAN, AVP RESEARCH AND EDUCATION INFORMATION TECHNOLOGY

Speakman characterizes NYU Langone Health as a relatively typical health system; they serve patients, they educate, and they discover. NYU Langone Health has health records of over nine million patients in Epic, their single healthcare database. Geographically, they cover the whole of New York City including Queens and Staten Island as well as much of Long Island. Their research has grown immensely over the past decade, with 260% growth over that time in NIH grant funding receipts. In research funding dollar per faculty, NYU tops NIH funding, edging out other systems such as Duke, Northwestern, and UCSD.

FIGURE 1



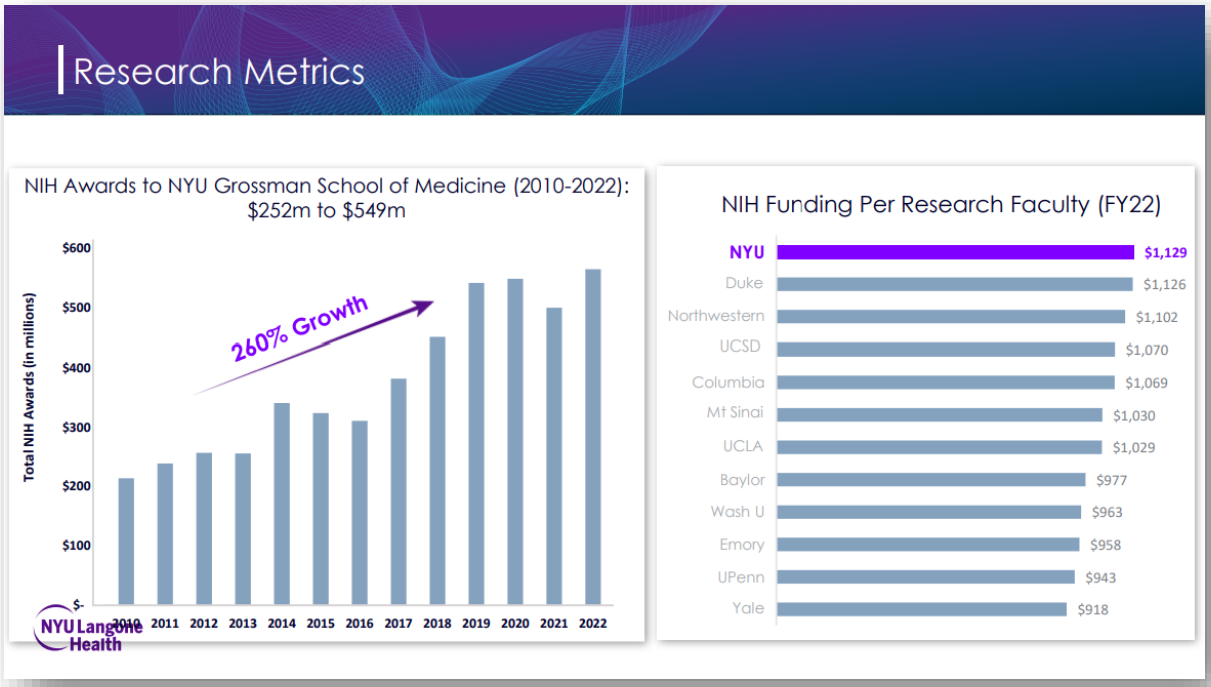
Source: John Speakman, 2024

NYU Langone's HPC system, the Ultraviolet Cluster, ranked #362 in a recent Top500 list, is expected to climb further up the list with the addition of their new SuperPOD. This makes them #16 among US academic sites in HPC system rankings. The Ultraviolet HPC Cluster is the only system in the Top500 fully dedicated to healthcare and life sciences. This quality is not simply an accolade; as Speakman notes, it allows the system to be a fully HIPAA-compliant insular system, with no considerations having to be made on health or patient data being stored or sent to an unsecure or non-compliant location.

At NYU Langone Health, they believe healthcare is an optimal location for the transformative power of AI. Demand for healthcare is "getting worse", patients want care that is both more data and

scientifically driven as well as more “human.” For Speakman, AI has the potential to handle the massive data and administrative demands of achieving this goal. In this way, it could enable clinicians to focus on interacting with patients, providing valuable time for offering personal care. These tools can easily identify medications, their side-effects and uses, and cut down on many of the time consuming or even harmful potentials in provisioning care.

FIGURE 2



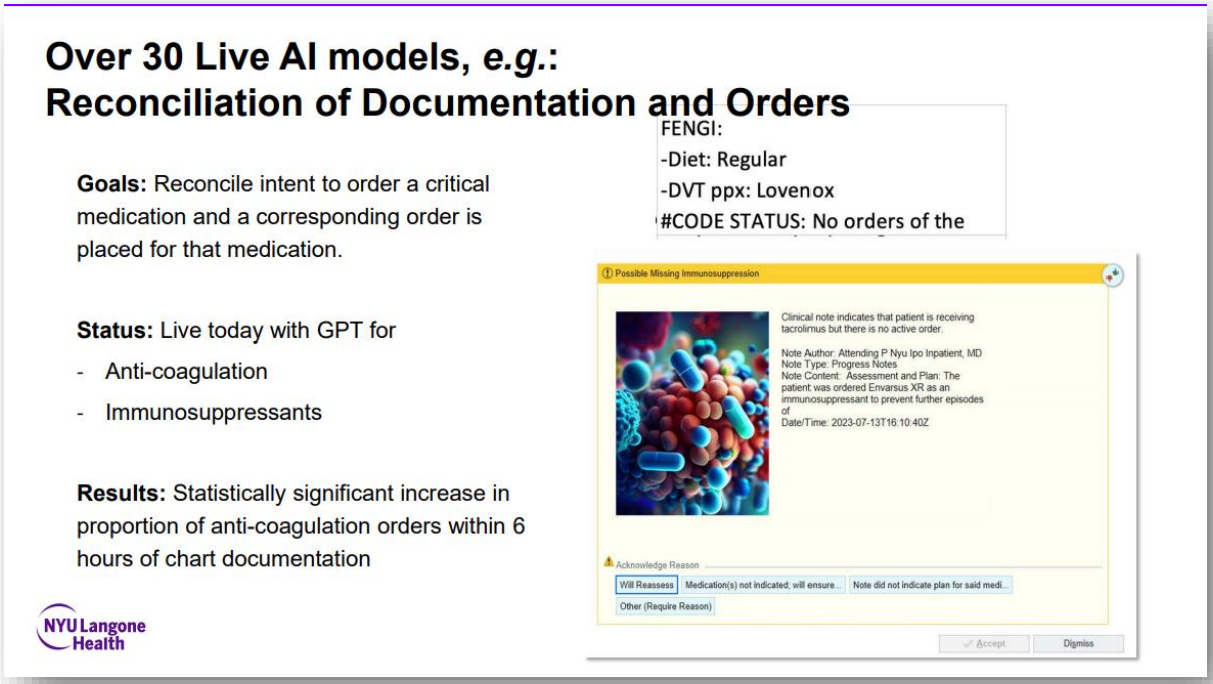
Source: John Speakman, 2024

Speakman admits that the typical considerations that must be made with AI tools must also be in sharp focus when applied to a healthcare environment, perhaps more so. Hallucinations, lack of sufficient data, automation bias, and non-generalizable models can be obstacles or blind spots when providing patient care with an AI tool.

The MCIT Department of Health Informatics at NYU Langone Health has the mission of developing, translating, and applying AI/ML and predictive models to support the highest quality of care and operational advancement. To Speakman, these tools are “not a means of writing nice papers or getting grants, it’s trying to move the needle on outcomes for our own patients.” It’s not about academic glory, it’s about actually trying to make a difference. The MCIT Department has been building AI models for a long time, that, for Speakman, gives them a significant leg up. For instance, one of their over 30 live integrated AI models is designed to reconcile the intent to order a critical medication by creating a corresponding order for that medication. This has resulted in a statistically significant increase in the

proportion of anti-coagulation orders within 6 hours of chart documentation. Simply put, these patients are getting important healthcare intervention faster and more reliably.

FIGURE 3

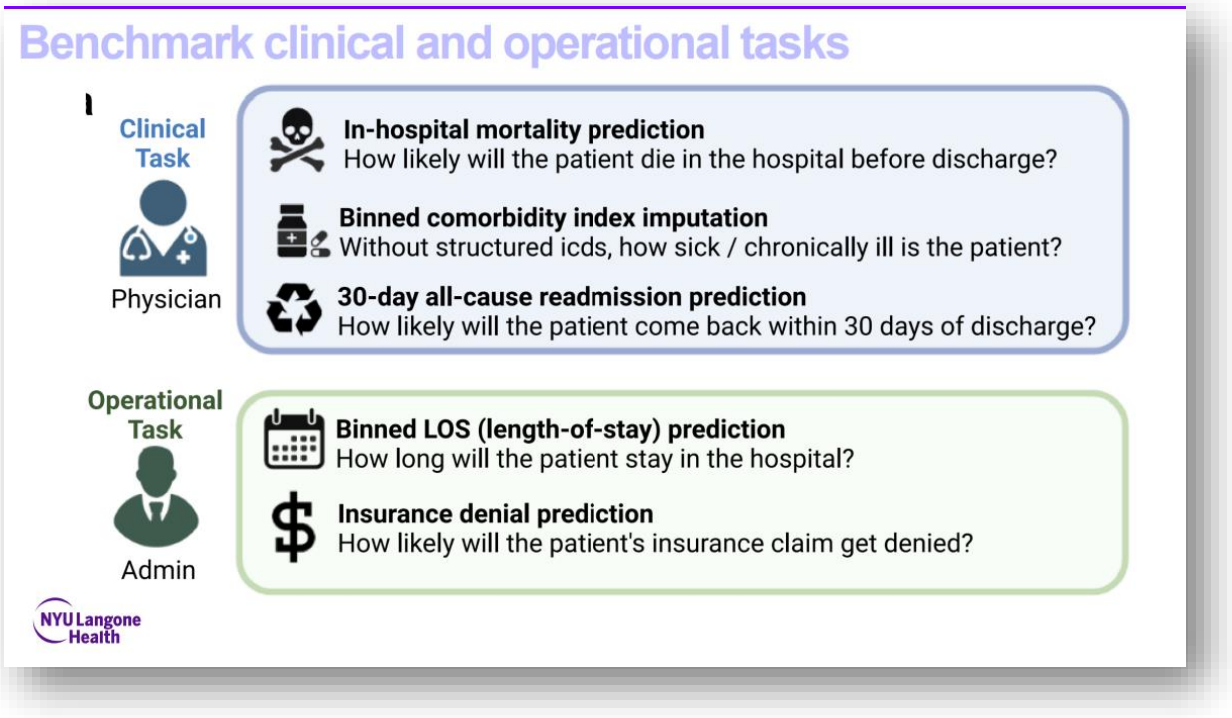


Source: John Speakman, 2024

Another tool that is integrated into their Epic health information system sorts patient messages via a triage model, putting those on the top which have identified the most urgent needs as opposed to a chronological ordering. This allows healthcare providers a greater opportunity to provide help or advice to patients facing the most critical needs in a speedier fashion. A high acuity request in this manner might result in a recommendation by the doctor to visit the emergency room. Conversely, a low acuity request, like a request for a prescription refill, might be taken out of the hands of the doctor entirely and entered in through an automated system which can handle the administrative burden of a message to the pharmacy.

In terms of their exploration into new tools, natural language search embedded in analytics tools is a capability currently being tested. There is also a proposed tool which provides short summaries of all patient conditions for clinicians and healthcare providers for end-of-shift hand-off reporting, a practice which can be a potential communication weak point. Similarly, an effort in partnership with Nuance hopes to provide ambient documentation powered by GPT, which are reports constructed through patient data that are accessible and easily readable by clinicians. All of these tools are aimed at freeing up valuable time for healthcare providers to focus on non-administrative, patient-focused tasks.

FIGURE 4



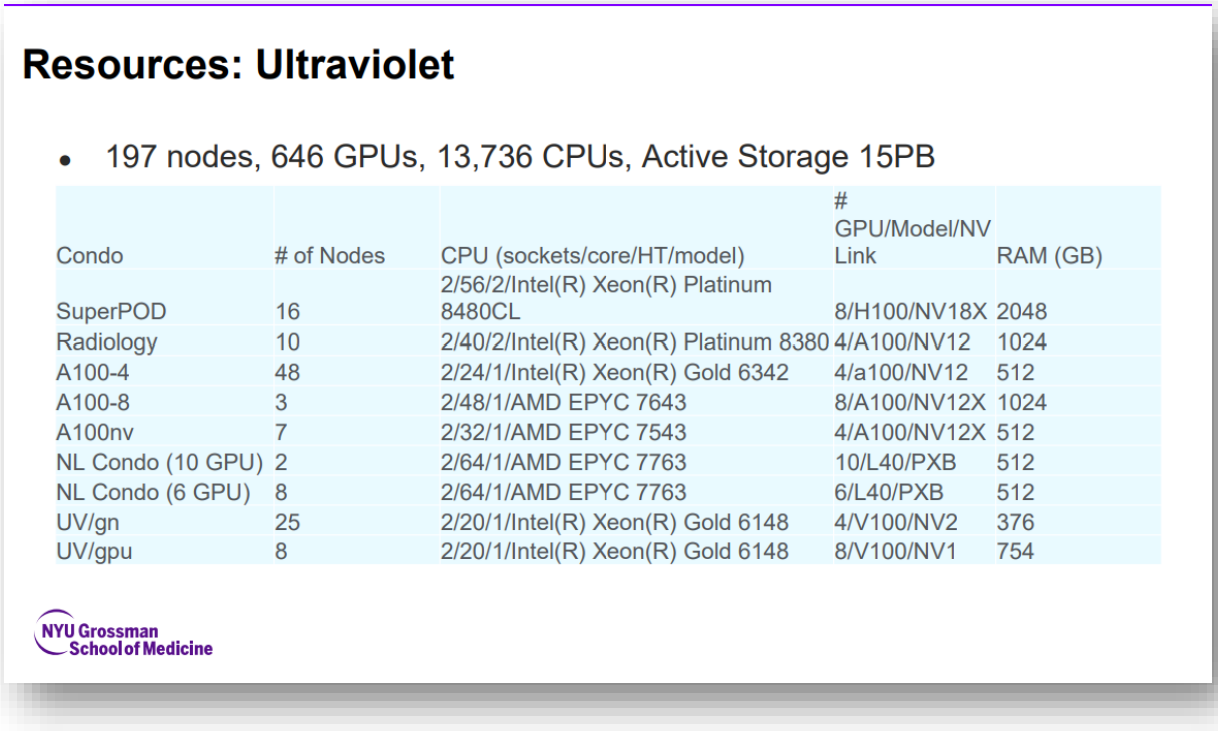
Source: John Speakman, 2024

NYU Langone Health has their own Large Language Model, NYUTron, developed by Eric K. Oermann, MD, neurosurgeon at NYU Langone. The first iteration of NYUTron required 24 NVIDIA A100s for 3 weeks. The hypothesis was to, instead of “drips” of data, provide “health-scale data”, using 7.25M notes, 4.1B words, 9.5 years’ worth of data into a 109M parameter BERT-like model with the goal of creating a truly predictive model of healthcare analytics. The model was pretrained with the records of 336,000 patients and further fine-tuned with 413,845 patient records from three different hospitals spanning the entire healthcare system. This segment still only represents less than 10% of the patients as the massive compute required in even this fractional train imposes major time constraints. There were five predictive outcomes tasked to the model: 1) in-hospital mortality, 2) comorbidity index, 3) readmission likelihood within 30 days, 4) length of stay prediction, and 5) insurance denial prediction.

Speakman describes NYUTron's performance with, for example, readmission prediction, as somewhat typical of AI models, its better than the average physician but is beat out by a highly-skilled physician. This is a pattern he recognizes in other areas like radiology. One of the key takeaways from this is that 30-day all-cause readmission risks can be predicted at the moment the clinician signs their notes. This seamless, low-latency integration is critical in a healthcare environment when information is needed quickly for decision making. For Speakman, another critical insight is that the big bottlenecks for this capability are hardware, engineering expertise, and datasets. Most academic medical institutions don’t invest in significant HPC capacity and, as such, don’t pretrain large LLMs on optimized in-house

software stacks. Furthermore, the lack of massive un-labelled and large, well labelled datasets is perhaps the biggest barrier to research.

FIGURE 5



Condo	# of Nodes	CPU (sockets/core/HT/model)	# GPU/Model/NV Link	RAM (GB)
SuperPOD	16	2/56/2/Intel(R) Xeon(R) Platinum 8480CL	8/H100/NV18X	2048
Radiology	10	2/40/2/Intel(R) Xeon(R) Platinum 8380	4/A100/NV12	1024
A100-4	48	2/24/1/Intel(R) Xeon(R) Gold 6342	4/a100/NV12	512
A100-8	3	2/48/1/AMD EPYC 7643	8/A100/NV12X	1024
A100nv	7	2/32/1/AMD EPYC 7543	4/A100/NV12X	512
NL Condo (10 GPU)	2	2/64/1/AMD EPYC 7763	10/L40/PXB	512
NL Condo (6 GPU)	8	2/64/1/AMD EPYC 7763	6/L40/PXB	512
UV/gn	25	2/20/1/Intel(R) Xeon(R) Gold 6148	4/V100/NV2	376
UV/gpu	8	2/20/1/Intel(R) Xeon(R) Gold 6148	8/V100/NV1	754

Source: John Speakman, 2024

The Ultraviolet Cluster supports these efforts as well as other compute-hungry projects with suites of A100 GPUs and largely other Nvidia hardware, and they have recently added their SuperPOD with16 nodes of H100s to boost this capability. With over 2,000 users with disparate projects and computing needs, fair share resource scheduling is not sufficient to support the quantity of resources needed to support LLMs. They face challenges with high demand and limited resources including checkpointing, education, and optimization.

With NYUTron and their electronic health record-integrated AI tools, NYU Langone Health is exploring the path towards AI capable healthcare. With their focus on patient outcomes and clinician empowerment, their use of this transformative technology serves to mirror the ideals of public health provisioning. NYU Langone Health's in-house capabilities can serve as an example of how greater HPC allocations in other healthcare systems can serve to grow administrative, academic, and medical efficacy.

For more information or to view this and other presentations given at HPC User Forums dating back to 2008, visit www.hpcuserforum.com.

About Hyperion Research, LLC

Hyperion Research provides data driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com
