BMJ Health & Care Informatics

How to organise a datathon for bridging between data science and healthcare? Insights from the Technion-Rambam machine learning in healthcare datathon event

Jonathan Sobel ,[1] Ronit Almog,[2] Leo Celi,[3] Michal Yablowitz,[4] Danny Eytan,[2] Joachim Behar[1]

To cite: Sobel J, Almog R, Celi L, *et al*. How to organise a datathon for bridging between data science and healthcare? Insights from the Technion-Rambam machine learning in healthcare datathon event. *BMJ Health Care Inform* 2023;**30**:e100736. doi:10.1136/bmjhci-2023-100736

Received 31 January 2023
Accepted 25 July 2023

[1]Biomedical Engineering, Technion Israel Institute of Technology, Haifa, Israel
[2]Epidemiology and Pediatric Critical Care, Rambam Health Care Campus, Haifa, Israel
[3]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[4]TIMNA- Israel's Ministry of Health Big Data Platform, State of Israel Ministry of Health, Jerusalem, Israel

Correspondence to
Dr Jonathan Sobel;
jsobel@campus.technion.ac.il

INTRODUCTION

A datathon is a time-constrained information-based competition involving data science applied to one or more challenges.[1–7] Datathons and hackathons differ in their focus, with datathons prioritising data analysis and modelling, while hackathons concentrate on building prototypes. Furthermore, hackathons can encompass a broad range of topics, spanning from software development to hardware design, whereas datathons are more narrowly focused on data analysis. In-person datathons offer the unique opportunity to learn alongside a community of fellow students and researchers, as well as to directly interact with clinicians and medical professionals. This is in contrast to Kaggle like competitions, which are often self-learning experiences.

Context of the event

A joint event organised by the Technion, Rambam Healthcare Campus and the MIT Critical Data group in March 2022 provided a unique opportunity to understand the challenges faced by leading researchers and clinicians working in the field of medical data science. The Technion is a leading science and technology research institutes and Rambam is the largest hospital in the north of Israel. It was organised as the inaugural event of a new joint Technion-Rambam initiative in medical AI (TERA), which aims to serve as an academic centre for medical AI committed to advanced medical and clinical research, with significant and actionable benefit to patient care.[3] The initiative opening event entitled 'Technion-Rambam Hack: Machine Learning in Healthcare,' was attended by about 250 people. The first two days consisted of a collaborative information-based competition that focused on solving real-world clinical problems through interdisciplinary teams and access to real data.[1–7] The datathon was followed by a one day conference with lectures delivered by researchers from the Technion, Rambam, MIT, the Israeli Ministry of Health (MOH), Clalit Health Services, GE Healthcare, and Roche.

The datathon days

The planning of the datathon and the conference began approximately six months before the event. After an initial brainstorming between the scientific committee, which included Technion principal scientists, Rambam clinicians and MIT scientists, a fundraising campaign was launched as list of potential speakers for the conference day was drawn up and invitations were extended. Communication around the event was initiated in November 2021 via social media platforms (Twitter, LinkedIn and Facebook). Students interested in the datathon were asked to apply to the event and were asked to complete a survey about their skills, their interests and their level of education (Bsc, Msc, Ph.D, alumni) and specialty (engineering or bio/med). We accepted approximately 70% of the applicants and the participation rate exceeded 95%. To ensure commitment from registrants to participate in the datathon, we required a registration fee of $25. In parallel, we contacted clinicians from Rambam and asked them to propose projects consisting of a medical question and to provide a relevant dataset to research the question. Four challenges

proposed by clinicians who had collected large datasets in recent years and who presented challenging scientific questions which could be tackled by ML were selected. The projects were (1) Prediction of newborn birth weight by maternal parameters and previous newborn siblings birthweights,[8] (2) ML-based predictive model for bloodstream infections during hematopoietic stem cell transplantation,[9] (3) Prediction of recurrent hospitalisation in heart failure patients[10] and (4) Risk factor and severity prediction in hospitalised COVID-19 patients.[11 12] Project leaders were required to provide an agreement for their dataset, following the standard Hospital Institutional Review Board (IRB) process.

Two competing teams composed of 5–7 participants were assigned to each project. This approach was adopted for two reasons: first, to increase the likelihood of obtaining interesting results from at least one of the teams, and second, due to the resource-intensive nature of dataset creation, which involves extraction, curation, and anonymization processes. The projects were designed to have comparable difficulty in terms of the structured (tabular) medical data provided, and we intentionally limited the number of variables to prevent overwhelming teams with an excessive amount of data. We had participants from diverse fields, comprising 1/3 biologists/medical professionals and 2/3 engineers, computer scientists, statisticians, or mathematicians. Ethical agreement was requested from all participants during the subscription process. Each participant signed a consent and a non-disclosure agreement. Each team was assigned a clinical mentor from the Rambam and a data science mentor from either the Technion or the industry. Participants were selected based on their interests and competency (studies and skills). Our goal was to have mixed teams in terms of data analysis capacity and field knowledge to work on each challenge. Each team had a separate virtual machine with personal, secured access for each team member. During the 2 days of the event, the teams were split in several rooms at the Technion Faculty of Biomedical Engineering. Each team was asked to present its work at the end of the second day. Thereafter, using an external jury comprised of a principal investigator from the Technion, clinicians, Rambam epidemiology and IT department, and industrial partners, the three best teams were selected for the competition final, which took place on the conference day.

The conference day

The guest talks at the conference aimed to introduce clinical data science to a wide audience and provide a perspective on its future impact on medicine. There was a total of 12 lectures delivered. The lectures were divided into three thematic sessions which are: (1) current trends in machine learning in healthcare, (2) data stakeholders, (3) deployment of machine learning in medical practice. The full list of lectures and speakers is available on the event website for reference (https://technion-hack.github.io/).

HOW TO ORGANISE A DATATHON?

To organise a successful event, several important points should be well thought through before the event (figure 1). The checklist provided below should help any organiser in this process. We further elaborate on some of these key points reflecting on our more mature experience.

Datathon check list

- Venue: Physical/virtual/hybrid, dates, location.
- Logistics: catering, strong WIFI, rooms and amphitheatre.
- Partners: Industrial, NGO, clinicians and academic stakeholders, who may fund some part of the event (awards/venue/infrastructure) as well as deliver relevant talks during the event.
- Projects: Research project call for datasets with ethical consents (IRB) and specified aims/questions.
- IT support and secured computational infrastructure (for sensitive clinical data to be shared with participants).
- Mentors: Clinical and data science. Try to select senior mentors.
- Participants: Who is your targeted audience? (students/medical professionals/data scientists).
- Communication/PR: Information to participants and advertisement of the event (flyers/website/social media).
- Awards: Money for the winning teams or other gifts, and support for the continuation of the project (scientific publication and start-up/spin-off).

One of the first decisions is related to the place and dates of the future event, should it be virtual, in person or hybrid. On-site events offer the advantage of providing a face-to-face experience, facilitating networking opportunities, allowing for more immersive experiences, and creating a sense of community among attendees. Online events offer several advantages, including increased accessibility and convenience, the ability to reach a wider audience regardless of geographic location, reduced costs for both organisers and attendees, and the ability to easily collect and analyse data on attendee engagement and behaviour. Hybrid events offer the advantage of combining the best of both virtual and in-person events, allowing for a wider reach and increased engagement while still maintaining a personal touch. However, hybrid events tend to reduce in-person attendance because of the alternative online option, which may be more convenient. We preferred an in-person event since the main objective was to enable human interaction and initiate a professional local community interested in ML in medicine. Without a doubt, this was the right choice, and a virtual meeting would have had very limited impact. Additionally, two of the talks were delivered as recorded videos. It was noticeable that while these were projected on a large screen with high-quality resolution and sound, the audience did not focus on these presentations at all. Instead, they started consulting their emails or working

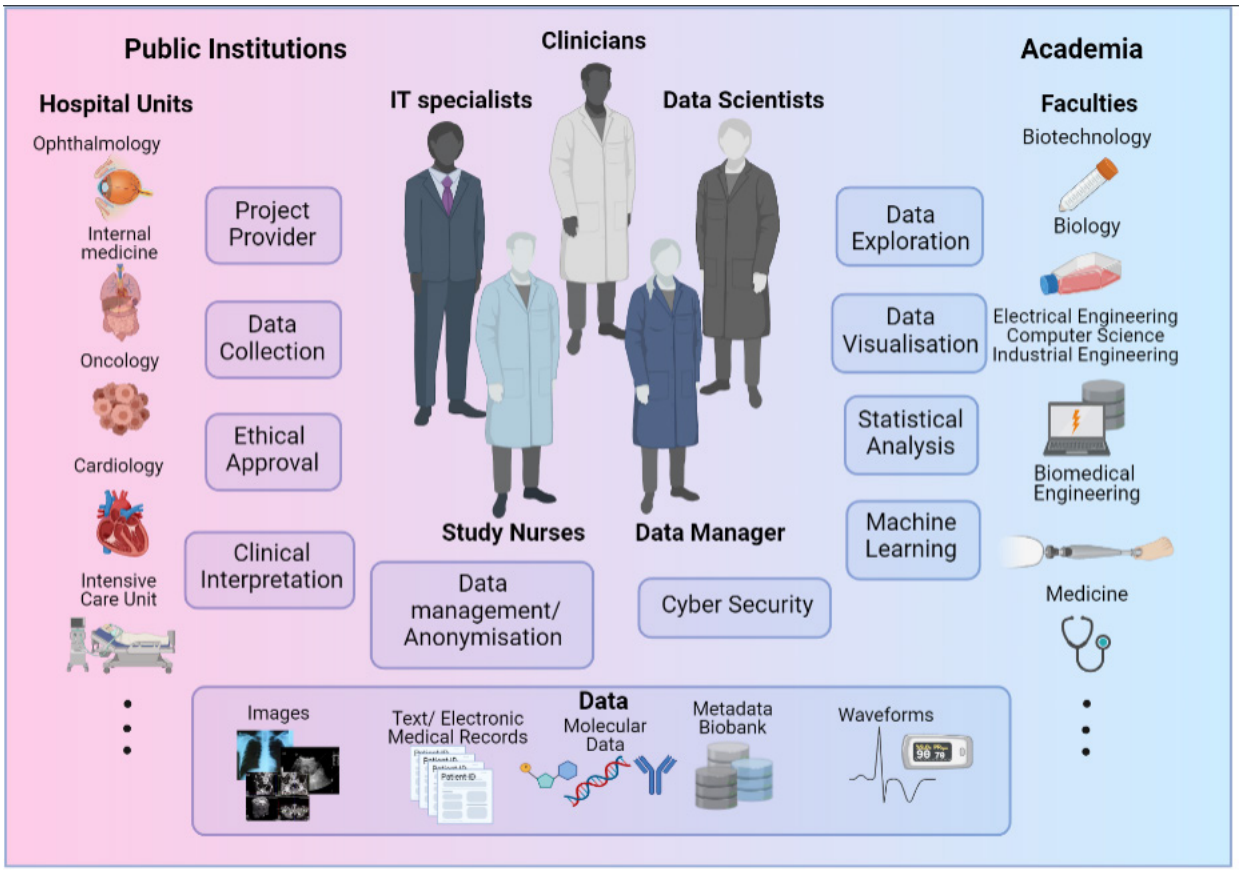

Figure 1 Datathon partners and essential components for sucessful projects. This figure was made with *BioRender*.

on their laptops. Thus, based on our experience, we strongly recommend an in-person meeting for such an event. Finally, we suggest that the lectures can be recorded and later released on YouTube if the organisers choose to widely distribute them – something we did – in order to maximise the impact of the event.

Recruiting sponsors was not the easiest task. Primarily, this was due to reaching out to local industries that had connections with some of the conference organisers. We presented the sponsors with various options. The generous partners had the opportunity to deliver a lecture and thereby promote their own research activities in the field of medical AI. This aligns well with the scientific programme and also helped in financing a portion of the event's cost. The lower-tier sponsorship option involved featuring their logo on our communications such as the website and flyers. However, this option did not attract any sponsors.

Another critical point for a datathon involves finding questions of clinical importance that can be addressed using previously collected data such as in.[13] There are several options here as some events may be more flexible in the sense that participants/projects leader can come with their data and questions or a more directed approach with defined datasets and questions. We chose projects from various medical fields rather than focusing on a specific problem area. This decision was made to foster the development of a professional community in the field of medical AI, promoting diversity in terms of the represented clinical specialties. Additionally, due to the time-constrained nature of the datathon, we aimed for students to work with real hospital data. Therefore, we sought datasets that had been developed by hospital researchers specifically for research purposes. In all cases, the data should be properly anonymized and the ethical statements from the IRB should be provided before the event. Mentors, with a clinical or a data science expertise, who can follow each team during the whole event are necessary to ensure the success of each project. The goal of a datathon is to demonstrate the effectiveness of a multidisciplinary approach where each team member can provide different expertise (medical, technical, social, legal and business). For that reason, we decided, as organisers, to create teams from the pool of participants prior to the datathon event.

We would like to add some additional recommendations based on things we would have done differently in retrospect. These include avoiding recorded lectures entirely. Additionally, it is important to ensure that speakers adhere to their allocated presentation time and to seek permission in advance for the use of pictures and videos from the event for marketing purposes. Finally, after the event, it would be beneficial to request feedback from participants through an online form in order to assess the impact of the event.

Twitter Jonathan Sobel @jonathansobel1

Acknowledgements We acknowledge the financial support of the Technion-Rambam Initiative in Artificial Intelligence in Medicine (TERA), the Technion Machine

Learning and Intelligent Systems (MLIS) Center, the Technion Human Health Initiative (THHI) and the MISTI MIT - Israel Zuckerman STEM Fund. This research was partially supported by The Milner Foundation, founded by Yuri Milner and his wife Julia. We are grateful to the Placide Nicod Foundation for their financial support (J.S.). We are grateful to the medical and technical staff from the IT and epidemiology departments at Rambam HCC. We are thankful to the Technion administrative staff for supporting the organization and communication of the event.

Contributors JS, JB, RA, LAC, and DE were involved in the conception and design. JS: was the coordinator. JS, JB, LAC, MY and DE wrote the first draft. All authors critically reviewed the first draft, and approved the final version.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iD

Jonathan Sobel http://orcid.org/0000-0002-5111-4070

REFERENCES

1 Aboab J, Celi LA, Charlton P, *et al*. “A "Datathon" model to support cross-disciplinary collaboration”. *Sci Transl Med* 2016;8:333ps8.
2 Anslow C, Brosz J, Maurer F, *et al*. Datathons: an experience report of data Hackathons for data science education. Proceedings of the 47th ACM Technical Symposium on Computing Science Education (SIGCSE ’16); New York, NY, USA: Association for Computing Machinery, 2016:615–20
3 Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health Informatics students: systematic review. *JMIR Med Educ* 2020;6:e19285.
4 Lyndon MP, Cassidy MP, Celi LA, *et al*. Hacking Hackathons: preparing the next generation for the Multidisciplinary world of Healthcare technology. *Int J Med Inform* 2018;112:1–5.
5 Serpa Neto A, Kugener G, Bulgarelli L, *et al*. First Brazilian Datathon in critical care. *Rev Bras Ter Intensiva* 2018;30:6–8.
6 Pathanasethpong A, Soomlek C, Morley K, *et al*. Tackling regional public health issues using mobile health technology: event report of an mHealth Hackathon in Thailand. *JMIR Mhealth Uhealth* 2017;5:e155.
7 Li P, Xie C, Pollard T, *et al*. Promoting secondary analysis of electronic medical records in China: summary of the PLAGH-MIT critical data conference and health Datathon. *JMIR Med Inform* 2017;5:e43.
8 McCowan LME, Harding JE, Stewart AW. Customised birthweight centiles predict SGA pregnancies with perinatal morbidity. *BJOG* 2005;112:1026–33.
9 Gupta V, Braun TM, Chowdhury M, *et al*. A systematic review of machine learning techniques in hematopoietic stem cell transplantation (HSCT). *Sensors (Basel)* 2020;20:6100.
10 Ouwerkerk W, Voors AA, Zwinderman AH. Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure. *JACC Heart Fail* 2014;2:429–36.
11 Reiner Benaim A, Sobel JA, Almog R, *et al*. Comparing COVID-19 and influenza presentation and trajectory. *Front Med* 2021;8.
12 Sobel JA, Levy J, Almog R, *et al*. Descriptive characteristics of continuous Oximetry measurement in moderate to severe COVID-19 patients. *Sci Rep* 2023;13:442.
13 Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:160035.