International Agency for Research on Cancer

World Health Organization

Scientific Council
Fifty-eighth Session

SC/58/3
19 November 2021

Lyon, 9–11 February 2022
By Web conference

UPDATE ON DATA SCIENCE ACTIVITIES

1. Following Recommendation 5 from the Evaluation Report on IARC activities vis-à-vis its mandate (see Document SC/56/8)[1], below is an update of the capacity for data science activities including: computational biology, bioinformatics, biostatistics, and (supporting these areas) Information Technology (IT).

2. The IARC Computational Biology, Bioinformatics and Biostatistics Committee (C3B) has continued to oversee these activities. C3B is composed of three working groups (WG): the bioinformatics WG, the biostatistics WG, and the IT WG. C3B's structure has been streamlined to increase the interaction between WGs: responsible officers in bioinformatics (Dr Matthieu Foll, GEM), biostatistics (Dr Vivian Viallon, NME), and scientific IT (Mr Christopher Jack[2], ITS) have been nominated, each of them chairing their respective WG, and rotating to chair the overall C3B.

3. Bioinformaticians and computational biologists are spread across IARC scientific branches, and comprise PhD students, postdoctoral scientists, research assistants, and professional staff members. The IARC bioinformatics WG is composed of approximately 20 active members. Its main objectives are to facilitate interaction and knowledge sharing, and to organize training in these areas. Regular internal seminar series are organized, complemented by biweekly informal community discussions among bioinformaticians (including early-career scientists), to discuss common technical aspects, new methodologies, and promote the use of best-practice tools.

4. Statisticians and biostatisticians are also spread across IARC scientific branches. The IARC biostatistics WG is composed of approximately 20 active members. Its overarching objective is to familiarize non-statistician colleagues with both standard and more advanced statistical tools, methods, and concepts. Seminars are organized every six weeks (approximately), on topics such as modern survival analysis, machine learning and dimension reduction techniques for high-dimensional (e.g. -omics) data, and causal inference. Training in these areas and ad-hoc technical support to non-statistician and early-career colleagues are provided.

---

[1] Recommendation 5: "*As computational biology needs an increasingly important component of laboratory capacity, IARC should regularly update the SC and the GC on capacity for computational biology in the future*".

[2] Following the departure of Mr Christopher Jack (ITS), a new responsible officer is to be nominated following recruitment.

5. The rapid growth of complex data (e.g. -omics data), available in both epidemiology and cancer biology calls for deeper interactions between bioinformaticians and biostatisticians. Accordingly, the bioinformatics and biostatistics WGs have enhanced their collaborations by organizing joint seminars and training (with support from the Human Resources Office/IARC Learning and Capacity Building (LCB) Branch), with the aim to also increase interactions between these communities at IARC.

6. During the COVID-19 lockdown, an innovative hybrid online course was organized in September 2020 entitled "Tidyverse Fundamentals with R: modern data manipulation and visualization in R", made up of five courses of four hours each. Thirty-three IARC participants followed the self-paced online courses, while dedicated internal online discussion groups and virtual meetings were organized to discuss specific points.

7. An advanced course called "Multivariate analysis for -omics data integration: principles and applications" will be held at IARC in November-December 2021 (3.5 days). The course is organized jointly with the Swiss Institute of Bioinformatics (SIB), with 21 participants from IARC and ten from SIB. The number of participants from IARC confirms the interest and need for such advanced courses at the interface between bioinformatics and biostatistics, as well as the increasing use of -omics techniques across all scientific branches.

8. The IT WG oversees the informatics hardware and software required to support IARC scientific activities. In the past two years (2021 vs 2019), we have seen a 95% increase in computing time on the IARC high-performance computing platform (3.5 million CPU hours in 2021), and an 84% increase in data stored (now totaling 725 TB).

9. The major achievement of the IT WG has been the creation and deployment of the IARC scientific IT platform, following a request for funds from the Governing Council Special Fund in 2020 (see Document SC/56/4). The platform provides access to cost-effective shared centralized computing resources for scientific data storage and analysis, based on modern tools and best practices to facilitate collaborative work. Of note, this system has also been crucial throughout the pandemic to allow IARC personnel to work remotely.

10. In an environment with increasing data protection regulations, this project has been conducted with support from the Committee for Information Security Oversight (CISO), and has been designed to provide security and protection of data at every step of its lifecycle.

11. The scientific IT platform was presented at an IARC-wide seminar and has been open to all IARC personnel since October 2021, following multiple testing phases. More than 100 projects from all IARC scientific branches are already hosted on the platform, and the remaining projects will be migrated before the move to the Nouveau-Centre in 2022. Detailed documentation is available to facilitate migration, and various resources including training and interactive and dynamic tutorials will be offered to facilitate and optimize the use of the modern computational tools available within the scientific IT platform at IARC.

12. We foresee scientific opportunities associated with further developments of the scientific IT platform, for instance, we anticipate IARC becoming a data hub for specific projects with external collaborators able to remotely use our platform to analyse data hosted on IARC premises.