International Agency for Research on Cancer

World Health Organization

Scientific Council
Sixtieth Session

SC/60/6
28 December 2023

Lyon, 7–9 February 2024
Auditorium

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 Data Science Steering Committee (DSSC) oversees the data science activities, including bioinformatics, biostatistics, computational biology, and scientific information technology. It is composed of three working groups (WG): the bioinformatics WG, the biostatistics WG, and the IT WG.

3. The DSSC builds upon the work of the former Computational Biology, Bioinformatics and Biostatistics Committee (C3B), with adjustments introduced to ensure closer alignment between the committee's objectives and the anticipated data science activities within IARC. Besides its new name, the DSCC now also aims to promote the concept of Open Science, in particular the FAIR principles (Findable, Accessible, Interoperable, Reusable) for open data, open-source code sharing and reproducible research in line with the IARC Medium-Term Strategy 2021–2025. The DSSC now also contributes to stimulate best practices for data management, data security and data safety, liaising with the IARC Committee for Information Security Oversight (CISO), the IARC Laboratory Steering Committee, IARC Data Managers, the IARC Data Protection Officer and the IARC Information Security Officer.

4. 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. Weekly meetings alternate between internal and external seminar series and are complemented by informal community discussions among bioinformaticians, including early-career scientists, to discuss common technical aspects, new methodologies, and to promote the use of best-practice tools.

5. 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 on a broad range of topics covering survival analysis, causal inference, as well as machine learning and high-dimensional statistical methods for the analysis of omics data. 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*".

6. The rapid growth of complex data, including omics data, available in both epidemiology and cancer biology calls for deeper interactions between bioinformaticians and biostatisticians. Accordingly, the bioinformatics and biostatistics WGs organize 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.

7. Multiple training events have been organized in 2022-2023, which were undertaken by 151 IARC participants in total. Statistics and programming with R continues to be an active area of interest for IARC personnel (49 participants in total across hybrid or online training), with training including a new course about interactive data visualization with R Shiny (30 participants, 6 hours). An online course on "Introduction to multiple imputation for missing data" was held in 2022 (39 participants, 7 hours), highlighting the interest and need for advanced statistical courses. Following its open science strategy, IARC also organized its first training session on "FAIR data principles in practice" in 2023 (16 participants, 4 hours) to equip personnel with the necessary skills for data management and sharing according to the FAIR principles.

8. The IT Working Group's responsibility is to identify current and future IT needs and develop IT solutions to facilitate scientific activities of IARC personnel and, possibly, its network of collaborators. The central resource in this regard is the IARC Scientific IT (SIT) platform providing 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. In the past two years (2021 vs 2023), IARC has seen a 22% increase in computing time on the SIT platform high-performance computing cluster (4.3 million CPU hours in 2023), a 39 % increase in data stored (currently 1.2 PB), 16% more users and 74% more projects hosted.

9. One of the most promising opportunities associated with the development of the SIT platform is to make it accessible to external collaborators. This would allow IARC to act as an Open Science data hub for some projects and ensure IARC scientists can host and share data with their partners in line with increasingly strict worldwide data protection standards. In 2022 a working group was set up to run a pilot phase with external collaborators accessing and analysing data remotely on the SIT platform for four selected projects.

10. The pilot phase has led to the development of administrative processes, a template Data Use Agreement and technical documentations, and the evaluation and documentation of the SIT platform's future needs and gaps with the support of consultants from Do IT Now S.A.S. In particular, the need for a back-office management tool to manage contracts, users, licenses etc., and an established financial model enabling the sustainment of these activities in the long term has been pointed out to allow a smooth roll out of the platform to more external users.

11. Do IT Now S.A.S. also evaluated the SIT platform structure more globally. The strategy to host data on IARC premises rather than relying on third-party cloud infrastructure has been considered fit for purpose based on the evaluation of the current and future IARC specific usage and needs and has led to specific recommendations for the storage and computing infrastructures.

12. Based on these recommendations, the renewal of the storage system started in 2023 with the evaluation of different architecture, technologies, and systems. This was followed by the publication of a Request for Proposal. The proposal offering the best value for money was selected. The implementation will take place during the first semester of 2024 and the renewal of the computing infrastructure is planned for 2025.

13. The pilot opening of the SIT platform to external collaborators and the evaluation of the SIT platform by Do IT Now S.A.S. also highlight the increasing needs for the development of the platform around data management tools, and registry to efficiently manage datasets, projects and consortium metadata.

14. As Artificial Intelligence and Machine Learning attract more and more attention in cancer research, an informal working group has been set up with monthly meetings to assess the interest of deep learning and other modern machine learning techniques for the analysis of data analysed at IARC. These methods are usually computationally intensive so their evaluation and application at IARC heavily relies on the SIT platform. IARC scientists recently developed a new machine learning method, based on optimal transport, to automatically align untargeted metabolomics data acquired across multiple studies and illustrated its interest for the identification of metabolic biomarkers of alcohol intake using data from several untargeted metabolomics studies nested within the European Investigation into Cancer and Nutrition (EPIC) study (Breeur *eLife* 2024). The working group recently initiated a project aiming at assessing the interest of auto-encoders and other non-linear unsupervised dimension reduction methods for the analysis of omics data, including transcriptomics and metabolomics data, in cancer epidemiology. Deep learning is also transforming histopathological image analysis. IARC scientists have recently developed HaloAE (Mathian *VISIGRAPP* 2023), a local version of the Transformer architecture, known to achieve state-of-the-art performance in natural language processing (GPT-4), allowing for the first time its application on histopathological whole slide images.