International Agency for Research on Cancer

World Health Organization

Scientific Council
Sixty-second Session

SC/62/6
21/11/2025

Lyon, 11–13 February 2026
Auditorium

UPDATE ON DATA SCIENCE ACTIVITIES

1. Following the recommendation made by the Scientific Council (SC) at its 56th Session, on the Evaluation Report on IARC activities vis-à-vis its mandate (see Document SC/56/8)[1], and the request for biennial updates (see SC Report SC/58/8), the Secretariat is providing an update on the capacity for data science activities including: 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 aims to promote the open-science concept, in particular the FAIR principles for open data, open-source code sharing and reproducible research in line with the IARC Medium-Term Strategy (MTS) 2021-2025. The DSSC also contributes to stimulating best practices for data management, data security and data safety, liaising with the IARC Committee for Information Security Oversight (CISO), 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. Meetings include internal and external seminar series and are complemented by informal community discussions among bioinformaticians 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 30 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 increasing availability of complex data, including multi-omics datasets, in both epidemiological and cancer biology studies calls for stronger 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 trainings have been organized in 2024 and2025, attended by 278 IARC participants in total (plus 81 external participants). The training programmes were predominantly delivered in person and covered a wide range of topics, including statistical and machine learning methods, programming, data collection, genomics as well as open science, such as best practices for ensuring code and research reproducibility according to open-science standards. The high participation rates (153 participants in 2024 and 225 participants in 2025) reflect a strong commitment to ongoing professional development within the field of data science.

8. The IT WG identifies IT needs and develops solutions to support IARC staff and collaborators. IARC's main resource, the Scientific IT (SIT) platform, offers cost-effective, centralized data storage and computing for scientific analysis. From 2023 to 2025, SIT saw a 122% rise in computing time utilisation (9.8 million CPU hours), a 14% increase in stored data (1.3 PB), 30% more users (197), and 55% more hosted projects (350 total, including 7 Consortiums and 134 Subprojects).

9. In 2024, the SIT Primary Storage underwent an upgrade, providing 40% more capacity, enhanced performance, built-in data encryption, and malware detection capabilities. All IARC scientific data has been successfully migrated from the previous system, with very limited disruption. The SIT Long Term Storage system that utilizes cost-effective magnetic tape for non-active data was also expanded to a capacity of 1.5PB.

10. Following the request for support from the Governing Council Special Fund in 2025 (see document GC/67/9-Rev1 and Governing Council Resolution GC/67/R8), the upgrade of the Computing Infrastructure began in 2025 after the different architectures, technologies, and systems were evaluated, leading to the publication of a Request for Proposal. The proposal offering the best value was chosen, and implementation is scheduled for the first half of 2026.

11. The SIT platform continues to evolve to support external collaborators, reinforcing IARC's capacity to serve as an Open Science data hub for collaborative projects. Significant progress has been made in maturing both the administrative processes and the technical solutions required to support this expansion securely and sustainably.

12. While the extension of the SIT platform to external collaborators remains in a pilot phase, two dedicated WGs were established to improve the operational framework. The first group developed a financial sustainability model, which has been approved by the Senior Advisory Team (SAT) and is pending implementation. The second group defined the requirements for a centralized "back-office" management tool to streamline contracts, user management, and software licensing. The necessity of these solutions is underscored by the rapid growth in

uptake: the number of external collaborators accessing the platform rose from 15 in 2023 to 134 in 2025.

13. Reflecting the growing demand from external partners for a broader range of analysis tools, a project has been initiated to align the capabilities of the external platform with those already available to IARC staff. By 2026, the environment dedicated to external collaborators will include support for JupyterLab, RStudio, and Microsoft Visual Studio Code alongside standard command-line interfaces, while maintaining strict data confinement policies, in line with worldwide data protection standards.

14. Over the 2024–2025 biennium, data science activities at IARC have aligned closely with major emerging trends in the field. Teams have continued to develop and implement innovative statistical, machine learning, and Artificial Intelligence (AI) methods, ranging from outlier detection for GLOBOCAN and clustering of countries in cancer screening programmes to advanced supervised and unsupervised learning applied in multi-omics, diagnostic imaging, and tumour classification. Important methodological advances have also been made in causal inference and survival analysis, including towards the emulation of target trials using electronic health records and the study of time-dependent risk factors to identify windows of susceptibility—periods in life when exposures may have the strongest impact on cancer development. Work has progressed on AI-driven tools such as LLM-based agents to support systematic reviews, while recognising ongoing challenges related to legal, trust, and transparency issues surrounding generative AI. To address confidentiality and privacy constraints while allowing international data to be analysed jointly, the Agency is also developing distributed and federated analytic frameworks.

15. Significant progress has also been made in bioinformatics and computational pathology, driven by the increasing adoption of high-resolution sequencing technologies, including single-cell and spatial approaches. Emerging foundation models for whole-slide imaging now enable scalable feature extraction and more systematic integration of histopathology with genomic, transcriptomic, and spatial data. New analytical workflows and tools have been developed to process, integrate, and interpret these complex data types, several of which have been released as open-source software packages to strengthen computational capacity both within IARC and across the global cancer research community.

16. Substantial efforts have been devoted to improving reproducibility and standardization through version control, containerization, and the harmonization of data management pipelines, including for major platforms such as the Global Cancer Observatory.