https://doi.org/10.1038/s41591-023-02715-8

A federated learning system for precision oncology in Europe: DigiONE

Piers Mahon, Ismini Chatzitheofilou, Andre Dekker, Xosé Fernández, Geoff Hall, Aslaug Helland, Alberto Traverso, Cedric Van Marcke, Janne Vehreschild, Gennaro Ciliberto & Giovanni Tonon

DigiONE is a pilot European learning health system in precision oncology that aims to identify optimal cancer treatments by learning from every patient, not just those in trials, through privacy-preserving interrogation of their standardized routine electronic health records.

CREDIT: DRAFTER123/ DIGITALVISION VECTORS / GETTY

Every cancer is different, depending on its particular molecular subtype. Cancer treatments will only be effective on certain subtypes. The optimal matching between patients and treatments will require granular and comparable information from routine care. This includes the clinical molecular profile of the cancer from next-generation sequencing (NGS), the treatments given to patients and the response of each patient to therapy. These data can be used to create a learning health system for precision oncology and interrogated to identify what treatments work best, for whom.

The Digital Oncology Network for Europe (DigiONE) aims to achieve this goal by using an innovative approach that is easy to integrate into routine care and is agnostic with respect to the electronic health record (EHR) system. Because it takes such a scalable approach, the technology is appropriate for adoption by broad national care systems beyond elite academic centers. This will prevent widening inequalities in cancer care outcomes between different care settings. Federation technologies will preserve privacy and allow the interrogation of learnings without centralizing data. This approach requires straightforward technological and organizational adaptations. DigiONE is aligned with Europe's Beating Cancer Plans and the Cancer Mission and could be integrated with the United States's Biden Cancer Moonshot to provide global benefit to cancer patients.

All cancer is now rare cancer

Precision oncology is the application of tests to define which therapies to give to which patient. It has profoundly transformed the treatment of cancer patients, as well as research and development in oncology, encompassing over 40% of current oncology pipelines. However, the partitioning of the cancer patient population into smaller and smaller subgroups is a major challenge. Oncology is moving toward a situation in which every cancer is characterized as a unique rare cancer with unique molecular specifications, such that each cancer will require a different therapeutic algorithm.

This characterization of cancer into rare subgroups is forcing innovation in trial and research methods globally[1]. This includes three European innovations: the drug rediscovery protocol, which aims to explore new indications for approved targeted and immunotherapies in clinical trials[2]; the European Organisation for Research and Treatment of Cancer, which merges prospective randomized trials and real-world evidence in trials-within-cohorts[3]; and the European Medical Agencies review of the use of external control methods[4]. Precision cancer is also driving the creation of large discovery -omics programs such as Europe's 1+ Million Genome/Genomic Data Infrastructure and International Cancer Genome Consortium–Accelerating Research in Genomic Oncology.

Care and guidelines are becoming increasingly complex

This subdivision of cancer into rare subtypes also has profound implications for routine clinical management because of its impact on treatment complexity—a challenge that is often overlooked. The average National Comprehensive Cancer Network guideline (one of the major clinical guidelines for cancer care in the United States) swelled, on average, from 26 pages and 30 clinical decision points in 1996 to 198 pages and 111 clinical decision points in 2019[5]. Care systems must accommodate increasingly complex testing requirements, with the number of cancer drugs approved by the US Food and Drug Administration with a companion diagnostic growing from approximately 10 in 2012 to 46 in 2021[6]. Such intricacy will only increase as more cancer subtypes are identified.

This complexity has troubling implications for the care of cancer patients. Already the quality of cancer care (as measured by 5-year survival) varies enormously throughout Europe. The recent Lancet Oncology Commission estimated that raising care levels in the countries with poorest survival to the European median range would avoid around 50,000 deaths annually[7]. The commission also emphasized the well-documented outcome gap between academic and non-academic healthcare settings. This gap is especially pronounced in Europe given the general lack of investment in molecular testing. As an example, high disparities are seen between European countries for molecular

profiling of NSCLC at diagnosis, with NGS used for only 50% of cases[8]. A major barrier is appropriate public test reimbursement for NGS panels or comprehensive genomic profiling that can simultaneously report all relevant biomarkers for on-label therapies. As a result, European molecular testing often relies on research cross-subsidy, which heavily favors academic healthcare settings.

Learning from real-world evidence

Tackling these mounting care inequalities is a policy priority in Europe's Beating Cancer Plans. Traditional approaches to improve quality (such as clinical audit) are not working fast enough, as shown by the ongoing differences in relative outcomes. Clinical auditing is also likely insufficient to cope with the increasing complexity of precision oncology, given the impact of the latter in fragmenting care pathways. The major implementation science challenge for precision oncology is how to bring such innovation equitably and efficiently into care systems and retire legacy models of care, regardless of the various care settings. This implementation challenge needs as much attention from the academic community and its funders as the molecular mechanisms of cancer.

Care quality can be improved in the precision era using digital methods and real-world evidence. A learning health system in precision oncology can connect routine NGS test results, clinical phenotype, treatment and outcome information[9]. Such a system could enable near-real-time care quality solutions and substantial real-world research on multimodal treatment optimization.

There are several implementation challenges in delivering this vision[10], given that EHR systems include highly heterogeneous data, often of low quality (especially for key clinical phenotype concepts); test data for biomarkers and somatic mutations are often in PDF format or in semi-structured pathology records; clinical abstraction in cancer means that major features such as iterative lines of therapy are hard to define digitally; lack of hospital information technology capabilities has resulted in a digital skills and resourcing gap both for clinical and IT teams; and there is considerable heterogeneity in privacy and data protection norms between nations (even within the European Union under the General Data Protection Regulation (GDPR)), with often additional complexity from risk-averse local interpretation of those norms.

Focus on clinical care outcomes

The Digital Institute for Cancer Outcome Research (DIGICORE) is a public–private partnership founded in 2020 to digitize cancer outcomes research. Today 40 cancer centers and two industry technology partners (IQVIA in clinical informatics and Illumina in genomics and bioinformatics) have joined DIGICORE. Over the past two years, DIGICORE has designed and is now prototyping a Digital Oncology Network for Europe (DigiONE), a federated learning health system that is compatible with the GDPR.

DigiONE has been designed to improve care delivery and quality. Under the GDPR, care delivery is considered the primary use case for health data, whereas research is the secondary use. Outcomes research sits at the overlap between research and care, with many ethics authorities considering it as clinical audit. DigiONE's focus on care improvement (rather than research) makes its implementation legally simpler for participating cancer centers under the GDPR, with legitimate interest a potential legal basis for data processing. The upcoming European Health Data Space regulation is expected to further strengthen the legal basis for such secondary data use for the public good.

The initial goal for DigiONE is the automation of outcomes research based on DIGICORE's collective experience in international real-world outcomes research studies[11]. Five of the authors (X.F., G.H., P.M., G.T. and J.V.) used their research experience and the OSIRIS minimum data set for data sharing and interoperability in oncology[12] to select key clinical concepts. These concepts describe cancer sufficiently for outcomes research, are clinically important and have reasonable EHR availability across Europe. The smallest possible data set was chosen to make technical implementation easier in less digitally mature cancer centers and to reduce the data-capture burden on frontline staff in non-academic settings.

A minimal essential description of cancer

A two-step open innovation competition was used to refine this overall concept and design a prototype network. Step 1 of the competition asked hospitals to engage with and improve the design through a series of surveys and a clinical consensus process. Sixteen hospitals in 13 countries contributed to the clinical consensus of a target dataset with fewer than 40 concepts, named MEDOC (Minimal Essential Description Of Cancer). MEDOC covers key demographic, clinical phenotype, biomarker, treatment and pragmatic outcomes (Table 1). To our knowledge, this is the smallest ever minimal cancer dataset; its detailed specification in Cancer OMOP (an open-source common data model developed by the Observational Health Data Science and Informatics (OHDSI) community) is underway. Other additional features of DigiONE's consensus design are summarized in Table 2.

In clinical informatics terms, MEDOC operates at the level of data concept, not data item. To illustrate this, although the biomarker section has only three terms (name, measure, sample ID), the ingestion process covers all routine biomarkers across blood tests, immunohistochemistry, germline or somatic mutation analysis and other specialist tests. Redundancy has been built in to allow for different data availability. As an example, treatment fields can be filled in from claims procedure data or by natural language processing (NLP) applied to clinical notes. Likewise, some countries prevent routine access to national death data, so vital status can be inferred from date of last visit or via death linkage.

Four data concepts must adhere to local or national rules on privacy and data protection: healthcare ID as a local patient identifier; legal basis for data processing; date of death; and derived vital status. Data concepts marked as derived are not ingested from source data into OMOP, but rather defined and calculated at the study level.

In step 2 of the competition, the hospitals that supported this design work were then able to bid for funding to build a local node according to the collectively agreed design. The scheme was run with open innovation rules approved by DIGICORE's independent board under DIGICORE's president, Gennaro Ciliberto, and was funded by industry partners IQVIA and Illumina. Technical solutions to generate cancer data in OMOP structure could be from any vendor or open source. Where health data in EHRs or other source systems is stored in an unstructured format, such as in a PDF or free text notes, hospitals need to structure the data to the OMOP common data model format to allow multicenter analysis. Multiple hospitals in DigiONE are using supervised NLP, which after iterative training and quality assurance steps may be migrated to unsupervised automated NLP if sufficient quality standards are reached.

Data science for care quality improvement

Learning health system solutions have been integrated into oncology EHR systems in the United States, for example by Flatiron Health[13].

Table 1 | Data concepts in MEDOC, a Minimal Essential Description Of Cancer

Area	MEDOC concept
1. Demographics	1.1 Date of birth (month) 1.2 Sex 1.3 Weight (with timestamp) 1.4 Height 1.5 Healthcare ID (or other unique identifier) 1.6 Legal basis for data processing
2. Clinical phenotype	2.1 Primary cancer diagnosis and comorbidities, typically in International Classification of Disease standards such as ICD10, ICD9 or ICD-O-3 2.2 Charlson comorbidity index (derived from 17 comorbidities in 2.1) 2.3 Date of primary cancer diagnosis 2.4 Method of primary cancer diagnosis 2.5 Performance status, for example coded by ECOG or Karnofsky standards 2.6 Disease stage in a recognized standard such as TNM 2.7 Histological cell type, typically in ICD-O-3 standards 2.8 Menopausal status, for breast cancer patients
3. Biomarkers	3.1 Biomarker name 3.2 Biomarker measure 3.3 Biological sample ID
4. Treatment	4.1 Line of therapy (derived algorithmically within each cancer type) 4.2 Anti-cancer treatment name, including systematic treatment and supportive therapy 4.3 Molecule generic name 4.4 Start date for drug treatment 4.5 Treatment dose 4.6 End date for drug treatment 4.7 Radiotherapy type 4.8 Radiotherapy start date 4.9 Radiotherapy dose 4.10 Radiotherapy end date 4.11 Surgery type 4.12 Surgery date 4.13 Participation in clinical trial 4.14 Date of trial consent
5. Outcomes	5.1 Date of death, at any location 5.2 Time to next treatment (derived from treatment start dates) 5.3 Metastasis presence/absence 5.4 Metastasis location 5.5 Date of clinical visits (with cancer related visits separated from other visits) 5.6 Vital status (derived from visits or death linkage) 5.7 Extent of debulking surgery, for gynecological cancer patients

Note: implementation of 1.5 and 5.1 is influenced by national regulations; 2.8 and 5.7 are essential only in some cancers for which risk-normalized audit or research cannot take place without their capture.

Table 2 | Features of DigiONE's consensus design

Feature	Details
Minimal Essential Description Of Cancer (MEDOC)	Consensus selection of <40 concepts with high availability to describe cancer using Findable, Accessible, Interoperable and Reusable (FAIR) data principles.
Near-real-time frontline feedback loops	Gives value to frontline clinicians and incentivizes improving data quality, in the form of clinical dashboards, audit tools or training log automation.
Pan-format cancer data ingestion	Solutions to incorporate both structured and unstructured data, including natural language processing (NLP) and optical character recognition (OCR).
Privacy-preserving solutions for NGS panels and biomarkers	Derived either from Portable Document File (PDF) or Variant Call Format (VCF) with no nucleic acid features (and so such biomarkers are not considered personal and sensitive under GDPR) and in data models that are technically appropriate for a network with heterogeneous testing workflows.
Full privacy-by-design, and federated infrastructure[14]	Allows statistical analysis equivalent to centralized data without data centralization. All patient-level data remains under local hospital control, with no transfers of pseudonymized patient-level data during research.
Modular, protocolized implementation plans	Drives international data normalization in digitally less mature settings, all in vendor agnostic open standards designed on open innovation principles

However, implementation under the US Health Insurance Portability and Accountability Act (HIPAA) regulations is legally simpler than it is under the EU's GDPR. Flatiron provides outcomes research mostly from its proprietary OncoEMR® network, with some specific integrations into EHRs at academic health centers. DigiONE provides a prototype for an international learning health system based on European oncology EHRs under the GDPR, which could be applied in other jurisdictions.

Implementation is underway in NSCLC, breast cancer and ovarian cancer in 2023. Other cancer subtypes will follow in 2024 and beyond, through a learning-by-doing approach to implementation. A repeatable series of OMOP protocols will be developed for each cancer, starting from simple modules and then building upon these to address more complex research questions, with a focus on care quality improvement. The most comprehensive studies will include studying contemporary disease natural histories with detailed retrospective biomarkers in those three cancers. Those studies will provide a proof of concept demonstrating that care quality monitoring can be automated, for instance by developing algorithms to measure key features of guideline compliance. The simpler studies provide a modular series of sprints for any hospital to start to mobilize its data towards research and care. This modular implementation approach builds from the experience of the Personal Health Train in international research in NSCLC to optimally predict prognosis[15]. Other cancer centers are welcome to join our protocols by contacting the authors.

The underlying architecture of DigiONE could provide a ready-made solution for multiple European initiatives, including the European Cancer Open Science Cloud, the European Health Data Space and the cancer arm of the 1+ Million Genomes. A key goal is that this project could eventually be applied to non-academic centers, where it could digitize national Beating Cancer Plans.

DIGICORE now has funding to add another 15 centers from the European Regional Development Fund. Care quality solutions that comply with the GDPR will also work in more permissive legal systems globally. They will also work in low- and middle-income countries, something that remains a global health priority. We invite hospitals and research funders across the world to join us to digitally transform cancer care and research for the precision oncology era.

Piers Mahon[1,2] ✉, Ismini Chatzitheofilou[3], Andre Dekker[4], Xosé Fernández[2], Geoff Hall[5,6], Aslaug Helland[7], Alberto Traverso[1,4,8], Cedric Van Marcke[9], Janne Vehreschild[10], Gennaro Ciliberto[11] & Giovanni Tonon[8,12]

[1]Digital Institute for Cancer Outcomes Research E.E.I.G, Brussels, Belgium. [2]IQVIA Cancer Research BV, Zaventem, Belgium. [3]Illumina Ltd, Cambridge, UK. [4]Department of Radiation Oncology (Maastro), GROW School for Oncology and Reproduction, Maastricht University Medical Centre, Maastricht, the Netherlands. [5]Leeds Teaching Hospital NHS Trust, Leeds, UK. [6]Health Data Research, <City>, UK. [7]Oslo University Hospital, Oslo, Norway. [8]IRCCS Ospedale San Raffaele, Milan, Italy. [9]Cliniques Universitaires Saint-Luc, Brussels, Belgium. [10]Goethe University Frankfurt, University Hospital, Center for Internal Medicine, Medical Department II, Frankfurt, Germany. [11]IRCCS National Cancer Center Regina Elena, Rome, Italy. [12]Università Vita-Salute San Raffaele, Milan, Italy.
✉e-mail: pmahon@iqvia.com

References
1. Madhavan, S. et al. *Nat. Cancer* **2**, 9–11 (2021).
2. van der Velden, D. L. et al. *Nature* **574**, 127–131 (2019).
3. Saesen, R. et al. *Eur. J. Cancer* **186**, 52–61 (2023).
4. Rippin, G. et al. *Drug Saf.* **45**, 815–837 (2022).
5. Kann, B. H. et al. *Cancer Treat Res. Commun.* **29**, 100492 (2021).
6. Lawler, M. et al. *Lancet Oncol.* **24**, e11–e56 (2023).
7. Normanno, N. et al. *Eur. J. Cancer* **176**, 70–77 (2022).
8. Mahon, P. et al. *J. Precision Med.* **1**, 39–45 (2015).
9. Barker, A. D. et al. *Cancer Res.* **82**, 2072–2075 (2022).
10. Cheeseman, S. et al. *Front. Oncol.* **13**, 1114435 (2023).
11. Guérin, J. et al. *JCO Clin. Cancer Inf.* **5**, 256–265 (2021).
12. Moncada-Torres, A. et al. *AMIA Annual Symposium Proceedings* **2020**, 870–877 (2020).
13. Deist, T. M. et al. *Radiother. Oncol.* **144**, 189–200 (2020).
14. Ma, X. et al. Preprint posted online (30 May 2020).

Q9

Competing interests
P.M. and X.F. are employees of IQVIA. I.C. is an employee of Illumina.

QUERY FORM

Manuscript ID	**[Art. Id: 2715]**
Author	**Piers Mahon**

AUTHOR:

The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query No.	*Nature of Query*
Q1:	Affiliations: Please provide city for affiliation 6 (Leeds? Birmingham?).
Q2:	Please confirm or correct the city name inserted in affiliation 6.
Q3:	Please check your article carefully, coordinate with any co-authors and enter all final edits clearly in the eproof, remembering to save frequently. Once corrections are submitted, we cannot routinely make further changes to the article.
Q4:	Note that the eproof should be amended in only one browser window at any one time; otherwise changes will be overwritten.
Q5:	Author surnames have been highlighted. Please check these carefully and adjust if the first name or surname is marked up incorrectly, as this will affect indexing of your article in public repositories such as PubMed. Also, carefully check the spelling and numbering of all author names and affiliations, and the corresponding author(s) email address(es). Please note that email addresses should only be included for designated corresponding authors, and you cannot change corresponding authors at this stage except to correct errors made during typesetting.
Q6:	Please check Figures for accuracy as they have been relabelled. Please markup minor changes in the eProof. For major changes, please provide revised figures. (Please note that in the eProof the figure resolution will appear at lower resolution than in the pdf and html versions of your paper.)
Q7:	In “Clinical auditing is also likely insufficient…,” correct to add “of the latter” (i.e., the original “its impact” referred to the impact of precision medicine, not the impact of clinical auditing)?
Q8:	Since the references were not cited in numerical order, they have been renumbered in the order of appearance. Please check.
Q9:	If ref. 13 (preprint) has now been published in final peer-reviewed form, please update the reference details if appropriate.