ACTA ONCOLOGICA
2026, VOL. 65, 159–163
https://doi.org/10.2340/1651-226X.2026.45120

NPCM 2025

MJS PUBLISHING

LETTER

‘Crossing borders’ in data standardisation: application of OMOP CDM in an international clinical trial network in precision cancer medicine

Maria Martin Agudo[a], Henk van der Pol[b,c], Gabriel Bratseth Stav[a], Tina Kringelbach[d], Katarina Puco[a], Åsmund Flobak[e], Hans Gelderblom[b], Kjetil Taskén[a], Gro Live Fagereng[a], Eivind Hovig[a]; on behalf of the PRIME-ROSE Consortium

[a]Institute for Cancer Research, Oslo University Hospital, Oslo, Norway; [b]Department of Medical Oncology, Leiden University Medical Center, Leiden, The Netherlands; [c]Mathematical Institute, Leiden University, Leiden, The Netherlands; [d]Department of Oncology, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark; [e]Department of Oncology, Trondheim University Hospital, Trondheim, Norway

ARTICLE HISTORY
Received 1 December 2025
Accepted 30 January 2026
Published 23 February 2026

KEYWORDS
Precision cancer medicine (PCM) clinical trials; data sharing network; standardisation; OMOP Common Data Model (CDM); ETL pipeline; evidence generation

Introduction

The PRIME-ROSE initiative is a European collaboration involving 28 countries and 11 national precision cancer medicine (PCM) trials that are ongoing or starting soon [1]. It combines data from trials with similar designs using an umbrella–basket approach and has shown that PCM is feasible and beneficial in European countries [2–4]. Patients with advanced cancer are enrolled into cohorts defined by tumour type, molecular alteration and assigned drug. However, recruitment is slow because these alterations are rare [2].

The PRIME-ROSE main objective is to demonstrate the effectiveness and safety of expanding the indication, and pooling trial data accelerates evidence generation [5].

Several approaches can be applied to standardise the structure of the incoming data within a common framework. Widely used strategies include HL7 Fast Healthcare Interoperability Resources [6], Phenopackets [7] or the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [8, 9]. The PRIME-ROSE consortium has adopted the OMOP CDM because it reduces variation across multisite data and supports the generation of reliable evidence [9–10, 11,12] in life sciences.

OMOP CDM allows harmonisation of the data from the different PCM clinical trials and retains the original values in the dedicated source fields. Additionally, the standardisation to OMOP CDM supports the usage of various Observational Health Data Sciences and Informatics (OHDSI) tools such as Usagi [13] for semantic mapping or Data Quality Dashboard (DQD) [14, 15] for quality check. Lastly, the standardisation to the OMOP CDM enables federated analysis in large precision oncology networks [16, 17].

The aim is to build an automated extract, transform and load (ETL) pipeline for rapid extraction of statistical outcomes from standardised, aggregated cohort patient data. PRIME-ROSE aims to establish a blueprint for sharing and pooling data between PCM trials.

Methods

Data sharing

Clinical trial data from ongoing trials are uploaded to and shared in the Service for Sensitive Data (TSD) after cohorts are completed. TSD is a secure environment managed by the University of Oslo. The consortium partners have agreed on sharing 41 variables (Table S1) featuring primary and secondary endpoints including progression-free survival [18].

ETL pipeline

Data controllers of each trial submit the respective datasets according to the variables included in the common dataset. Subsequently, these data will be standardised to the OMOP CDM v5.4 using the ETL pipeline (see Figure 1A and Supplementary Material for a detailed description of the pipeline).

ETL implementation and deployment

A pre-processing workflow was developed for sources where eCRF variables must be combined in datasets at the trial level.

CONTACT Eivind Hovig ehovig@ifi.uio.no Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Ullernchausseen 70, 0379 Oslo, Norway

Supplemental data for this article can be accessed online at https://doi.org/10.2340/1651-226X.2026.45120

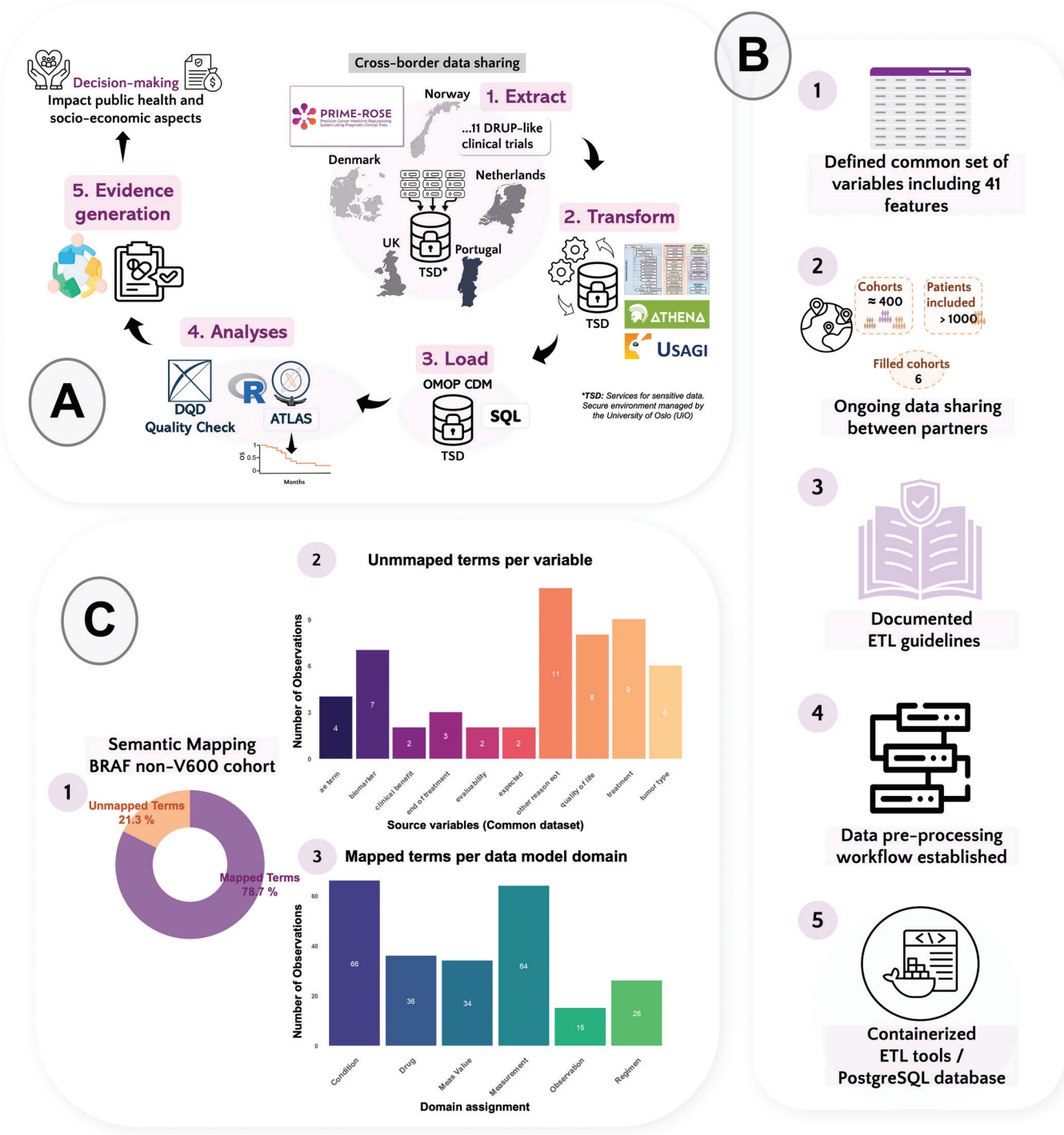

Figure 1. (A) The diagram shows the flow of the data within PRIME-ROSE. Sharing data between countries across Europe and the application of an extract, transform and load (ETL) process for data transformation and statistical analyses are key elements to accelerate the analysis of clinical trial data and evidence generation by enabling rapid patient cohort enrolment. (B) Summary with the major advances of the ETL development in PRIME-ROSE. The common dataset is only shared once the patient cohorts are filled (1 and 2). The ETL logic (3) guides the coding process for transforming the data and connecting the variables to the CDM Database. Currently, a pre-processing step prepares the data acquired from multiple sites (4). A minimal test for setting up software containers in the sensitive area TSD has been developed (5). (C) Semantic mapping was tested with the OHDSI tool Usagi in a BRAF-nonV600 cohort. (1) 306 source observations (terms) were analysed with Usagi and 78.7% (241) were mapped to a standard concept within the OHDSI standardised vocabulary catalogue. (2) The unmapped terms are confided to one of the nine variables displayed in the x axis of the bar plot, refining the semantic mapping process is crucial for the study. (3) Majority of the mapped terms connects to condition_ocurrence or measurement tables within the OMOP CDM version 5.4.

All records are validated against an internal schema and loaded into a harmonised data model. This ensures that downstream modules performing semantic and structural mapping to the OMOP CDM v5.4 target schema receive a consistent input.

For execution within the secure TSD environment, a prototype container-based deployment was tested, comprising a PostgreSQL database container (v15.0) and a separate ETL container with the Python application. Container images were built locally using Docker (v27.5.0) and executed using Podman (v5.6.0) [19].

Results

Common PRIME-ROSE variables

To date, the ongoing data sharing and manual aggregation between the different partners in PRIME-ROSE has led to merging of 396 cohorts with 1133 patients included. Six of the cohorts have been completed, and none has been stopped due to lack of efficacy (Figure 1B). A common dataset with 41 variables has been defined (Table S1).

Mapping to OMOP CDM version 5.4

A first iteration of the ETL logic serves as a reference for mapping the PRIME-ROSE source variables into the OMOP CDM v5.4 structured tables (Figure 1B). Similarly as reported by Ajmal et al. [17], we found that for some of the common PRIME-ROSE variables (e.g. *Concomitant medication* or *Dose delivered)* there is no straightforward mapping to the model and additional relationships should be established.

Semantic mapping challenges

Usagi (v1.4.3) was used to perform a semantic mapping test where source terms from aggregated data were mapped into standard concepts from the OHDSI standardised vocabularies (v5.0 27-FEB-25) (see Table S2). A BRAF-nonV600 cohort (Figure 1C) was selected, consisting of 53 patients recruited in three trials: (1) DRUP (Netherlands), (2) IMPRESS-Norway (Norway), and (3) FINPROVE (Finland).

In total, 306 source terms were processed. From those, 78.7% (241) were successfully mapped (Figure 1C1) to a standard concept within one of the six domains in Figure 1C3. Mostly, the mapped terms belonged to condition_ocurrence or measurement tables within the OMOP CDM v5.4. As shown in Figure 1C1, 21.3% of the terms (*n*=65) remained unmapped, which belonged to one of the nine variables displayed in the plot (Figure 1C2). This can happen in variables where the input is free-text and it contains a misspelling, as seen in one observation for our aggregated dataset. Also, we found that broad terms are employed for defining some of the biomarkers/targets, for example, fusions or BRAF activating mutations. These are not specifically mapped to the OHDSI vocabulary, OMOP Genomic, which contains 289889 concepts. Finally, most of the unmapped terms belonged to the free-text variable *other reasons for end of treatment (eot).*

Discussion and conclusion

Developing an ETL, such as the one presented here, is a dynamic and malleable process that should accommodate the different needs of the partners who are sharing the data. To ease versioning, reproducibility, stabilisation and individualisation of processes, we utilise Docker containers. From our experience, semantic mapping of certain variables such as *biomarker* or even *adverse events* is complex, and probably requires a higher level of data granularity or standard terms to input in the mapping and/or more extensive vocabularies (e.g. more concepts in OMOP genomics OHDSI vocabulary). Also, other variables with free-text observations such as *other reasons end of treatment* are difficult to map, as textual similarity comparisons with Usagi has some limitations. We are refining the process by performing data validation, maintaining multidisciplinary collaboration with experts and investigating alternative approaches, including machine learning-based tools to enhance semantic mapping.

PRIME-ROSE implements FAIR (Findable, Accessible, Interoperable, and Reusable) principles [20]. We are commited to open science, within the limits of sensitive patient data protection, as much as possible, in order to benefit the scientific and public community. FAIRification and standardisation to OMOP CDM prepares PRIME-ROSE for the implementation of the European Health Data Space (EHDS) [21, 22], as a key ecosystem to harbour large-scale evidence networks such as EHDEN [10].

This work with ongoing PCM trials in Europe showcases how standardised and structured data in PCM may facilitate cross-border data sharing to influence the development of PCM, particularly for rare cancer trials. It can serve as a blueprint for similar or expanded initiatives such as Joint Action on Personalised Cancer Medicine (JA PCM) [23]. PRIME-ROSE aims to increase the number of partners, and implementation of standardisation to OMOP CDM is key to enable federated analysis in PCM with similar trials outside of the European consortium.

Acknowledgements

This work is supported by 'Precision Cancer Medicine Repurposing System Using Pragmatic Clinical Trials' (PRIME-ROSE coordinated by Kjetil Taskén.) funded under the Horizon Europe programme (grant number 101104269). Key participants in the PRIME-ROSE consortium are named as consortium authors. We are also grateful to a number of individuals, institutions, industry partners, and grant agencies that are not listed, but contribute to building national initiatives and DRUP-like clinical trials in each country. We would like to acknowledge the advice and guidance from Lars Halvorsen and Emma Gesquiere (EdenceHealth). Special thanks to the NOR-OMOP members for sharing their OMOP knowledge with us. This work was performed on the TSD (Tjenester for Sensitive Data) facilities, owned by the University of Oslo, operated and developed by the TSD service group at the University of Oslo IT-Department (UiO IT). (tsd-drift@usit.uio.no). Finally, this short report summarises the

work presented at the Nordic Precision Cancer Medicine Symposium 2025 (NPCM2025), taking place in Oslo on the 15th–17th of September 2025. NCPM2025 was financially supported by the Acta Oncologica Foundation.

Conflicts of interest

No competing interests to declare.

Data availability statement

The methodology described in this short report has been developed and tested using: (i) a synthetic dataset created for pipeline development purposes, and (ii) a small patient cohort dataset. Relevant code is publicly available in the GitHub repository https://github.com/pcm-primerose/omop_etl. However, for ethical reasons and in complicance with European GDPR regulations, to protect the patient information, datasets are not publicly available.

Ethics declarations and trial registry information

All aggregated data generated within the PRIME-ROSE are originated from individual datasets shared by partners with on-going DLCTs. All patients have signed informed consent and are informed about the data sharing in the PRIME-ROSE network. Each trial is approved by their ethics committee and registered in CTIS; DRUP study: 2023-509152-33-0. ProTarget: 2023-510527-29-00. IMPRESS-Norway: 2020-004414-35. FINPROVE: 2024-517478-68-01. megaMOST: 2019-001494-88. DETERMINE: NCT05722886.

Funding

The Precision Cancer Medicine Repurposing System Using Pragmatic Clinical Trial (PRIME-ROSE), funded under the Horizon Europe programme (grant: 101104269).

Authors' contributions

Maria Martin Agudo: Conceptualisation, Methodology, Investigation, Formal Analysis, Data curation, Writing – Original Draft, Writing – Review & Editing.
Henk van der Pol: Conceptualisation, Methodology, Investigation, Formal Analysis, Data curation, Writing – Review & Editing.
Gabriel Bratseth Stav: Conceptualization, Methodology, Software, Investigation, Formal Analysis, Data curation, Writing – Review & Editing.
Tina Kringelbach: Methodology, Investigation, Formal Analysis, Data curation, Writing – Review & Editing.
Katarina Puco: Methodology, Writing – Review & Editing.
Åsmund Flobak: Methodology, Writing – Review & Editing.
Hans Gelderblom: Supervision, Writing – Review & Editing.
Kjetil Taskén: Supervision, Writing – Review & Editing, Project administration, Funding acquisition.
Gro Live Fagereng: Conceptualisation, Methodology, Investigation, Resources, Data curation, Supervision, Writing – Review & Editing.
Eivind Hovig: Conceptualisation, Methodology, Resources, Supervision, Writing – Review & Editing.
All authors contributed to the article and approved the submitted version.

References

[1] Taskén K, Haj Mohammad SF, Fagereng GL, Sørum Falk R, Helland Å, Barjesteh Van Waalwijk Van Doorn-Khosrovani S, et al. PCM4EU and PRIME-ROSE: collaboration for implementation of precision cancer medicine in Europe. Acta Oncol. 2024;63:385–91. https://doi.org/10.2340/1651-226X.2024.34791

[2] Van Der Velden DL, Hoes LR, Van Der Wijngaart H, Van Berge Henegouwen JM, Van Werkhoven E, Roepman P, et al. The Drug Rediscovery protocol facilitates the expanded use of existing anticancer drugs. Nature. 2019;574(7776):127–31. https://doi.org/10.1038/s41586-019-1600-x

[3] Kringelbach T, Højgaard M, Rohrberg K, Spanggaard I, Laursen BE, Ladekarl M, et al. ProTarget: a Danish Nationwide Clinical Trial on Targeted Cancer Treatment based on genomic profiling – a national, phase 2, prospective, multi-drug, non-randomized, open-label basket trial. BMC Cancer. 2023;23(1):182. https://doi.org/10.1186/s12885-023-10632-9

[4] Puco K, Fagereng GL, Brabrand S, Niehusmann P, Støre Blix E, Samdal Steinskog ES, et al. IMPRESS-Norway: improving public cancer care by implementing precision medicine in Norway; inclusion rates and preliminary results. Acta Oncol. 2024;63:379–84. https://doi.org/10.2340/1651-226X.2024.28322

[5] Tasken K, Fagereng GL, Abel E, Baltruskeviciene E, Carlsson KS, Falk RS, et al. Single point of entry to the European precision cancer medicine trial network PRIME-ROSE. J Clin Oncol. 2024;42(16_suppl):e23024. https://doi.org/10.1200/JCO.2024.42.16_suppl.e23024

[6] Bender D, Sartipi K. HL7 FHIR: an Agile and RESTful approach to healthcare information exchange. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems [Internet]. Porto, Portugal: IEEE; 2013 [cited 2025 Nov 30]. p. 326–31. Available from: http://ieeexplore.ieee.org/document/6627810/

[7] Jacobsen JOB, Baudis M, Baynam GS, Beckmann JS, Beltran S, Buske OJ, et al. The GA4GH Phenopacket schema defines a computable representation of clinical data. Nat Biotechnol. 2022;40(6):817–20. https://doi.org/10.1038/s41587-022-01357-4

[8] Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54–60. https://doi.org/10.1136/amiajnl-2011-000376

[9] Reisinger SJ, Ryan PB, O'Hara DJ, Powell GE, Painter JL, Pattishall EN, et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. J Am Med Inform Assoc. 2010;17(6):652–62. https://doi.org/10.1136/jamia.2009.002477

[10] Voss EA, Blacketer C, Van Sandijk S, Moinat M, Kallfelz M, Van Speybroeck M, et al. European Health Data & Evidence Network – learnings from building out a standardized international health data network. J Am Med Inform Assoc. 2023;31(1):209–19. https://doi.org/10.1093/jamia/ocad214

[11] Kent S, Burn E, Dawoud D, Jonsson P, Østby JT, Hughes N, et al. Common problems, common data model solutions: evidence generation for health technology assessment. PharmacoEconomics. 2021;39(3):275–85. https://doi.org/10.1007/s40273-020-00981-9

[12] Trinh NT, Houghtaling J, Bernal FL, Hayati S, Maglanoc LA, Lupattelli A, et al. Harmonizing Norwegian registries onto OMOP

common data model: mapping challenges and opportunities for pregnancy and COVID-19 research. Int J Med Inform. 2024;191:105602. https://doi.org/10.1016/j.ijmedinf.2024.105602

[13] Usagi [Internet]. [cited 2025 Dec 1]. Available from: https://ohdsi.github.io/Usagi/

[14] Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. eGEMs. 2016;4(1):18. https://doi.org/10.13063/2327-9214.1244

[15] OHDSI/DataQualityDashboard [Internet]. 2025 [cited 2025 Dec 1]. Available from: https://github.com/OHDSI/DataQualityDashboard

[16] Mahon P, Chatzitheofilou I, Dekker A, Fernández X, Hall G, Helland A, et al. A federated learning system for precision oncology in Europe: DigiONE. Nat Med. 2024;30(2):334–7. https://doi.org/10.1038/s41591-023-02715-8

[17] Ajmal A, Bouissou O, Brash J, Cheeseman S, Banduge PG, Gomes AL, et al. Establishing standards: harmonising coding principles for a minimal cancer dataset in the OMOP Common Data Model. ESMO Real World Data Digitl Oncol. 2025;9:100179. https://doi.org/10.1016/j.esmorw.2025.100179

[18] Van Der Pol H, Kringelbach T, Martin Agudo M, Bratseth Stav G, Fagereng GL, Fiocco M, et al. Procedures of data merging in precision cancer medicine: the PRIME-ROSE project. Acta Oncol. 2026;65:1–8. https://doi.org/10.2340/1651-226X.2026.44889

[19] Gantikow H, Walter S, Reich C. Rootless containers with Podman for HPC. In: Jagode H, Anzt H, Juckeland G, Ltaief H, editors. High performance computing [Internet]. Cham: Springer International Publishing; 2020 [cited 2025 Nov 27]. p. 343–54. (Lecture Notes in Computer Science; vol. 12321). Available from: https://link.springer.com/10.1007/978-3-030-59851-8_23

[20] Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3(1):160018. https://doi.org/10.1038/sdata.2016.18

[21] European Health Data Space Regulation (EHDS) – Public Health [Internet]. 2025 [cited 2025 Nov 30]. Available from: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en

[22] European Commission. Directorate General for Health and Food Safety. European health data space [Internet]. LU: Publications Office; 2024 [cited 2025 Dec 1]. Available from: https://data.europa.eu/doi/10.2875/269514

[23] About | japcm [Internet]. [cited 2025 Dec 1]. Available from: https://www-acc.japcm.eu/about