ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ESMO REAL WORLD DATA AND DIGITAL ONCOLOGY

ORIGINAL ARTICLE

Opportunities and challenges in pooling health-related quality-of-life data for prediction modeling in breast cancer across Europe: lessons from the EORTC BALANCE project

T. G. W. van der Heijden[1*†], N. J. Hubel[2†], K. M. de Ligt[1], I. R. Kist[3], V. Arndt[4], H. M. Verkooijen[5,6], G. Velikova[7,8], M. Hoedjes[9], B. H. de Rooij[9], B. Holzner[2] & L. V. van de Poll-Franse[1,9], on behalf of the EORTC QLG

[1]Department of Psychosocial Research and Epidemiology, Netherlands Cancer Institute, Amsterdam, The Netherlands; [2]University Hospital of Psychiatry II, Medical University of Innsbruck, Innsbruck, Austria; [3]Department of Board of Directors, Netherlands Cancer Institute, Amsterdam, The Netherlands; [4]Unit of Cancer Survivorship, Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany; [5]Division of Imaging and Oncology, University Medical Centre Utrecht, Utrecht; [6]Utrecht University, Utrecht, The Netherlands; [7]Leeds Institute of Medical Research, University of Leeds, Leeds; [8]Leeds Cancer Centre, St James's University Hospital, Leeds, UK; [9]Center of Research on Psychological Disorders and Somatic Diseases, Department of Medical and Clinical Psychology, Tilburg University, Tilburg, The Netherlands

Available online 7 August 2025

Background: Health-related quality of life (HRQoL) is a crucial outcome for cancer patients, providing a comprehensive measure of patient well-being beyond traditional clinical endpoints. While HRQoL data are increasingly available from real-world data (RWD), randomized controlled trials (RCTs), and observational studies, they remain fragmented, limiting their utility for large-scale analysis. The European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group's BALANCE project aims to address this by pooling and harmonizing international HRQoL datasets for breast cancer patients.

Materials and methods: This article describes the challenges of pooling international HRQoL datasets, including the process of dataset identification, acquisition, and harmonization within the BALANCE project.

Results: We successfully pooled and harmonized six datasets, representing 6500 patients and over 30 000 observations from diverse RCTs, observational studies, and RWD sources. The resulting database includes 142 variables across demographic, clinical, and HRQoL domains. Challenges included various interpretations of the General Data Protection Regulation across Europe, related to data protection and ownership. Furthermore, inconsistent data collection and resource limitations (e.g. funding or personnel) required iterative negotiations and customized harmonization. This led to the exclusion of 17 datasets containing an estimated number of 20 000-22 500 patients.

Conclusions: The BALANCE project demonstrates the feasibility of pooling international HRQoL data by overcoming key barriers and creating one of the largest HRQoL datasets for breast cancer. It lays the groundwork for upcoming publications focused on developing and validating prediction models. Future research should focus on adopting standardized data models, including secondary use clauses in consent forms, and establishing RWD registries to facilitate data sharing.

Key words: health-related quality of life, patient-reported outcomes, breast cancer, data pooling, data harmonization

INTRODUCTION

Many cancer patients suffer from the effects of cancer and its treatments, which can cause physical and psychosocial morbidity and result in impairments of health-related quality of life (HRQoL).[1] In cancer research and care, the collection of HRQoL data using patient-reported outcome measures (PROMs) has increased in recent years.[2,3] This expansion aligns with the broader trend in health care, where digitalization is driving rapid growth in data availability.[4-6] The growing availability of HRQoL data facilitates big data analysis[7]: machine and deep learning (ML/DL) can uncover clinically relevant information in massive amounts of data, by discovering new connections between the available variables beyond classical analysis. For example, predicting the next HRQoL measurement regardless of the time between measurements creates the flexibility needed when applying these models to clinical practice. Yet, these self-learning algorithms require large amounts of data.[4,8]

*Correspondence to: Thijs G. W. van der Heijden, Department of Psychosocial Research and Epidemiology, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. Tel: +31 20 512 9111
E-mail: t.vd.heijden@nki.nl (T. G. W. van der Heijden).

†These authors share first authorship.

Therefore, the BALANCE (Big data in patients with breast cancer) project [funded by the European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group (QLG)] aims to bring together HRQoL data from multiple sources in Europe to support the development of such algorithms.

Bringing together HRQoL data is essential: despite the substantial volume of HRQoL data in oncology, most datasets remain too small for ML/DL and, in some cases, even for traditional prediction methods.[7,9] Existing HRQoL data are scattered over many smaller datasets, with a variety of PROMs and measurement strategies collected in many initiatives.[10-13] Pooling existing datasets from various sources could enable robust HRQoL predictions using sufficiently large samples.[9,14] Unlike hypothesis-testing studies, prediction models do not rely on traditional sample size calculations; instead, the goal is to optimize model performance.[15] A larger sample size generally provides more precise and reliable results, with effective sample sizes often determined by the number of outcome events.[16]

Furthermore, the growing availability and volume of HRQoL data necessitate sustainable, future-proof management to achieve sufficient sample sizes to unlock the potential of HRQoL prediction modeling. Secondary use of data—repurposing them beyond their original intent—offers a cost-effective alternative to setting up new large-scale studies, which can be expensive, labor-intensive, and burdensome for patients.

While pooling data overcomes some barriers and creates opportunities for more analysis types, challenges are introduced for identification, acquisition, and harmonization of the data. To inform future data pooling projects, it is important to share our experiences. Therefore, we discuss the challenges of HRQoL data pooling encountered in the BALANCE project. Firstly, the BALANCE project will be outlined, followed by the steps of identification, acquisition, and harmonization, highlighting the challenges encountered and the solutions implemented. We then discuss key lessons, solutions, and alternative solutions not taken.

MATERIALS AND METHODS

Setting

The objective of the EORTC QLG BALANCE project is to develop comprehensive and accurate risk prediction models for HRQoL for breast cancer patients. The combined effects of treatment, lifestyle, demographics, comorbidities, and psychosocial factors on HRQoL outcomes will be investigated.

A sample size of 10 000 patients was aimed for to exceed typical patient numbers in artificial intelligence and classical modeling studies[9,17-20]; initial analyses were deemed feasible with 5000 patients. No formal sample size calculations were conducted, as this exceeded conservative estimates using a rule of thumb. The rationale for the large sample was based on the necessity to adjust for population heterogeneity, to account for missing data, and to enable robust subgroup analyses.

Once the first datasets were merged, sample size calculations for classical models with continuous outcomes were carried out using R's 'pmsampsize' package.[21] Depending on the selected R^2 (0.4-0.9), the minimum required sample size ranged from 537 to 2455 patients, assuming all possible predictors of the BALANCE project and a varying model intercept (10-90) with a standard deviation of 10.

Data source identification

To identify potential data sources, the scope of the search was limited to datasets about adult women diagnosed with non-metastatic breast cancer (stage I-IIIA), who had completed at least two PRO assessments using the EORTC Quality of Life Core questionnaire (QLQ)-C30 and/or breast cancer-specific QLQ-BR23 instruments[22,23] before, during, or after treatment. Additionally, each dataset had to include dates of HRQoL measurements, date of diagnosis (DoD), treatments received, and disease characteristics like staging and TNM (tumor—node—metastasis) status.

The first datasets were identified through the EORTC QLG network, a European consortium of researchers focused on improving HRQoL in cancer care. A collaboration was established with the EORTC statistics department to identify and access relevant EORTC trials.

Additionally, we explored data availability through Project Data Sphere (PDS; https://data.projectdatasphere.org) and Vivli (https://vivli.org). While differing slightly in scope, both platforms are key facilitators of clinical research data sharing. They provide access to anonymized participant-level data from completed clinical trials, either through direct download or upon request.

Lastly, we conducted a literature search to ensure comprehensive data identification.[7,24,25]

Data acquisition

After identifying the datasets, the principal investigators and/or the corresponding authors of the original studies were contacted. Once an agreement was reached, the data-sharing process was initiated. Depending on institutional requirements, additional approval from the institutional review board (IRB) or ethics committee (EC) was sometimes necessary before data were acquired.

The data-sharing agreements (DSAs) were drafted by the data owner; often, a template from the EORTC legal department provided a starting point for contract negotiations. The DSAs specified ownership of the datasets and rights to the resulting output. For the EORTC BALANCE database, a 'data processor' model was adopted, wherein ownership remains with the original institute, while data stewardship and processing rights were granted to the applicants' institutions—Netherlands Cancer Institute (NKI) and Medical University Innsbruck (MUI). Intellectual property (IP) rights for the research output were assigned to NKI, MUI, and EORTC, while data contributors were credited for authorship and data. General Data Protection Regulation (GDPR) Article 26 'Joint controllers', which

governs data use and processing, was either addressed in separate agreements or the main DSAs. This model was chosen as other alternatives were practically impossible, including owning all data outright, as some studies have clauses against this in their informed consent, while others were technically or financially infeasible for the project (e.g. federated learning, see 'Discussion' section). For the resulting models and other research output, the NKI, MUI, and the EORTC are the joint owners of the end product.

Data harmonization

Without a uniform standard for included variables, a codebook was devised to encompass all pertinent variables for our project (see Supplementary Table S1, available at https://doi.org/10.1016/j.esmorw.2025.100172). The variable definitions are based on the guidelines of the Dutch Federation of Medical Specialists[26] for disease and treatment characteristics of breast cancer, the European Network of Cancer Registries,[27] and the Observational Medical Outcomes Partnership (OMOP) common data model[28] for data standardization and quality checks. We harmonized all variables that were acquired into the database. To ensure alignment in the HRQoL assessment timeline and to allow further analysis, the DoD was selected as the independent anchor for all date variables. DoD is both available for real-world data (RWD) as well as trial data and therefore suitable to align the differing HRQoL timelines.

All HRQoL scales were calculated using the EORTC scoring manual,[29] and the C30 domain scores were dichotomized according to the clinically relevant cut-offs of Giesinger et al.[30] These thresholds classify patients' domain scores based on whether the impairment is clinically relevant.

Ethics

Ethical approval for the EORTC BALANCE study was obtained in 2022 from the IRB of Antoni van Leeuwenhoekziekenhuis/Nederlands Kanker Instituut (IRBd22-179). Ethical approval for the OPTIMUM study was obtained from the Medical Research Ethics Committee Brabant, The Netherlands (reference number: NL66913.028.18). The VERDI study was approved by the ECs of the University of Heidelberg and the Medical Association of Saarland. Written informed consent was obtained from all participants. The study protocol for UMBRELLA was approved by the Institutional Review and Ethics Board of the University Medical Center Utrecht, The Netherlands.

RESULTS

Data identification

Twenty-five datasets with breast cancer patients who completed HRQoL questionnaires were identified. Key sources were the EORTC QLG ($n = 9$), EORTC Quality of Life department (QLD) ($n = 1$), and the authors' networks ($n = 6$) (Figure 1). Six more were found via data-sharing platforms, and one registry contacted us through word-of-mouth communication. A literature search (including the results from van der Heijden et al.[7]) yielded two additional datasets.

Data acquisition

Outreach and screening. For datasets from EORTC QLD ($n = 1$), Vivli ($n = 5$), and PDS ($n = 1$), documentation was requested to assess eligibility. For the remaining studies ($n = 18$), we contacted the principal investigators to screen for eligibility.

Data-sharing negotiations were started with nine dataset holders, while 17 datasets (estimated 20 000-22 500 patients) were excluded (Figure 1). Three datasets (∼2000 patients) were excluded due to missing or unclear formulation of patient consent for data sharing or secondary data use. In these cases, the lawful basis of explicit consent (pursuant to Article 6 (1) (a) and 9 (2) (a) together with Article 9 (2) (j) and 89 (1) of the European Data Protection Regulation 2016/679 (GDPR[31]) was not applied entirely. Two datasets were excluded due to a lack of monetary resources, as no budget was available for data purchase or reimbursing data preparation (∼3500 patients). Two datasets were excluded due to competing projects or conflicting interests of the principal investigators (∼1000 patients). Four additional datasets were excluded for various reasons, like lack of perceived benefits for contributors (∼1200 patients). All datasets from Vivli and PDS were excluded due to missing HRQoL assessment dates ($n = 3$, ∼1600 patients) or key variables like treatments and age ($n = 3$, ∼12 500 patients).

Data-sharing negotiations. Negotiations for seven out of nine datasets have been completed, and six sets have been included in the BALANCE database; we are currently waiting on data transfer for the seventh dataset. Other datasets are in data protection review ($n = 1$) or awaiting steering committee approval ($n = 1$).

Table 1 shows the time from first contact to data receipt; the average was 18.2 months (range 2-32 months, $n = 6$). On average, each data-sharing contract underwent three reviews by legal or data protection officers.

Data harmonization. Harmonization began with creating the codebook, using the NKI data as a template, with adjustments made after consulting a breast cancer surgeon and the study team. Variable definitions followed the established guidelines (see 'Materials and methods' section). Missing variables were either derived from others or set to not applicable, while others were recoded to the BALANCE codebook format using R Studio (v4.3.2; Posit PBC, Boston, MA). Coding issues were resolved by two researchers (TGWvdH and NJH); if no agreement was reached, the study team decided (KMdL, BHdR, LVvdPF, BH). This led to, for example, having a binary comorbidity variable as well as some binary variables for categories of comorbidities. This resulted in a 'raw' dataset where both original variables and BALANCE variables were present; from this set, a 'cleaned' dataset was selected for merging.

Before merging, each patient was assigned a unique BALANCE ID, a dataset source indicator, and a data type (RWD/trial) indicator. Merging was done in stages, starting with the NKI and Kufstein data, followed by the AMAROS, UMBRELLA, OPTIMUM, and VERDI datasets.

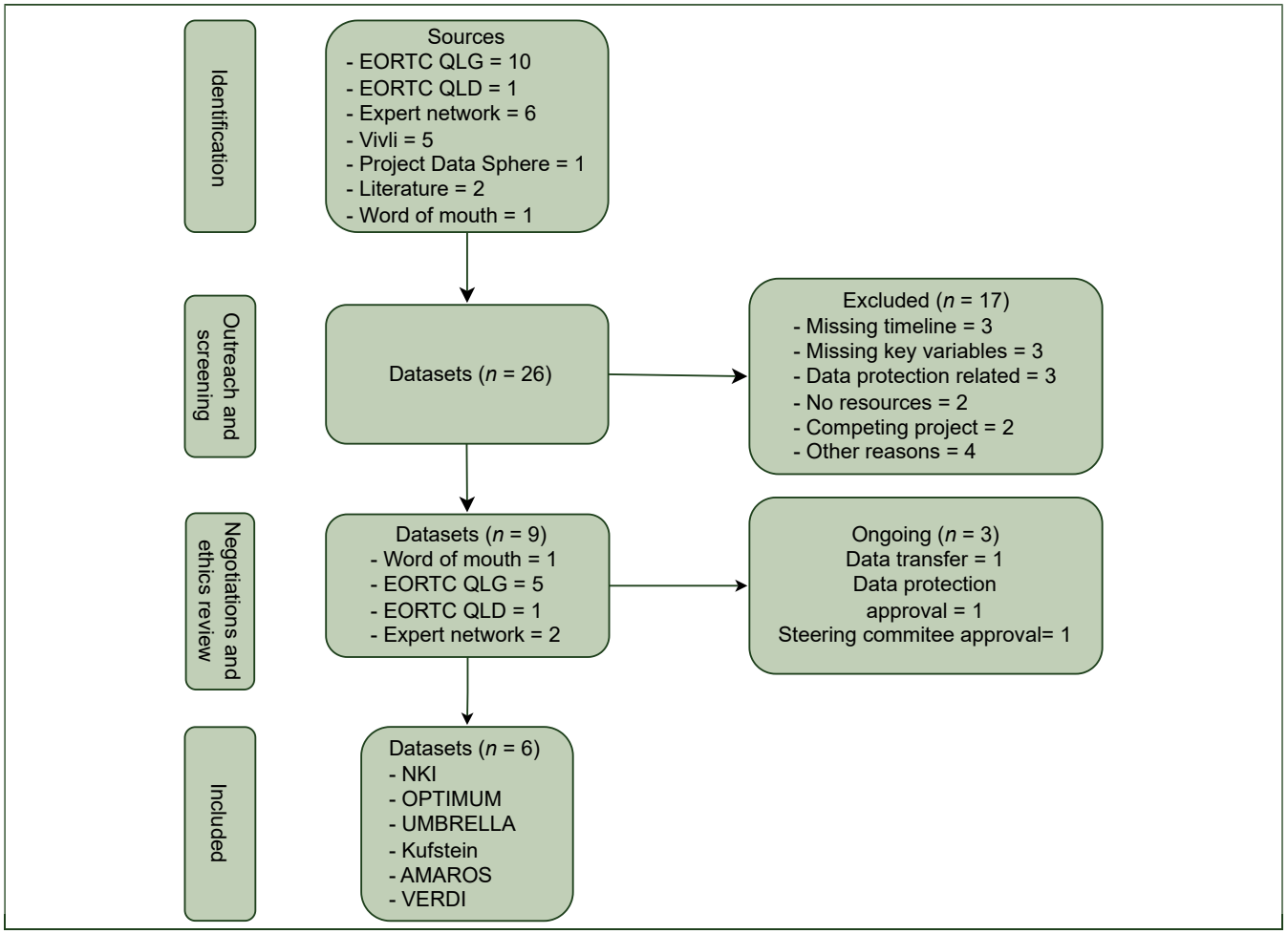

Figure 1. Flowchart of dataset inclusion into the EORTC BALANCE study. Only datasets that met the inclusion and exclusion criteria at the identification step are shown. EORTC, European Organisation for Research and Treatment of Cancer; QLD, Quality of Life department; QLG, Quality of Life Group.

Overview of datasets included. The BALANCE database currently includes six datasets with 6500 patients (Table 2), with >30 000 observations of varying assessment time points (Figure 2). These datasets include randomized controlled trial data (EORTC AMAROS), observational studies (UMBRELLA, OPTIMUM), and RWD (NKI, BKH Kufstein), representing a diverse breast cancer population with varying treatments, ages, and survival durations.

Harmonization of the six datasets yielded 157 common variables, including 53 HRQoL items, 23 HRQoL scales, 15 binary scales (QLQ-C30 thresholds), 15 demographic characteristics, 15 comorbidities, 20 disease characteristics (e.g. tumor staging), 14 therapy characteristics (e.g. type of therapy), and 2 time variables (time of HRQoL assessment in months/days since diagnosis; codebook in Supplementary Table S1, available at https://doi.org/10.1016/j.esmorw.2025.100172).

Table 1. Overview of duration between first contact, agreement between researchers, signed DTA, and receipt of the data

Dataset	Time between first e-mail contact and first meeting	Time between first meeting and starting contract negotiations	Time between starting contract negotiations and signing the contract	Time between signing the contract and receiving the data	Total time between first contact and receiving the data
1	<1 month	<1 month	1 month	5 months	6 months
2	1 month	1 month	29 months	<1 month	31 months
3	1 month	<1 month	1 month	1 month	3 months
4	<1 month	<1 month	7 months	<1 month	7 months
5	<1 month	<1 month	7 months	<1 month	7 months
6	12 months	<1 month	Pending (>9 months)	—	—
7	<1 month	5 months	9 months	<1 month	15 months
8	<1 months	15 months	9 months	<1 month	Pending
9	1 month	Pending	—	—	—

Last updated in March 2025. Anonymized for privacy.

Table 2. Overview of datasets and their main characteristics that make up the BALANCE database

Dataset	Number of patients	Country	Data origin	Maximum number of HRQoL assessments	Questionnaires used	Time period of data collection
AMAROS[32]	1274	Multiple[a]	RCT	5	C30	2001-2010
UMBRELLA[33]	3108	The Netherlands	Observational cohort	22	C30 and BR 23	2013-present
OPTIMUM[34]	664	The Netherlands	Observational cohort	2	C30	2019-2023
NKI[35]	1018	The Netherlands	RWD	6	C30 and BR 23	2021-present
BKH KUFSTEIN[36]	130	Austria	RWD	58	C30	2005-2022
VERDI[37]	306	Germany	Observational cohort	4	C30 and BR 23	1997-2009

RCT, randomized controlled trial; RWD, real-world data.
[a]France, Israel, Italy, The Netherlands, Slovenia, Switzerland, Turkey, UK.

Issues regarding harmonization. During harmonization, several challenges emerged due to the nature of the project. Firstly, data quality varied across datasets, requiring standardization at the lowest complexity level. For example, treatment variables were standardized into dichotomous variables for each treatment category (surgery, chemotherapy, etc.), whereas datasets like NKI contain more detailed treatment information, including initiation and termination dates.

A second challenge involved the completeness of the data, including data missing either at random (e.g. inconsistently recorded in RWD) or systematically (e.g. not being available/measured across entire datasets). This heterogeneity may influence subsequent analyses.

Thirdly, to ensure patient privacy, all time points were converted to days since diagnosis, making this variable pseudonymized. Diagnosis was set as day 0, and a timepoint, e.g. 3 months, is considered 91 days. For AMAROS, where DoDs were unavailable, randomization dates were used to align datasets.

Resources needed for pooling

Two PhD students contributed to the project—one for 3 years and the other for 1.5 years. Securing the study contract and passing ethics checks took 9 months. Data recruitment was a part-time task over 2 years (6-30 months) (Table 1).

Data cleaning was the most resource-intensive phase, involving codebook creation and dataset standardization. Standardization took 1-1.5 weeks per dataset, but ongoing updates, missing variables, and corrections extended cleaning to 3-4 weeks per dataset, totaling 4-6 months.

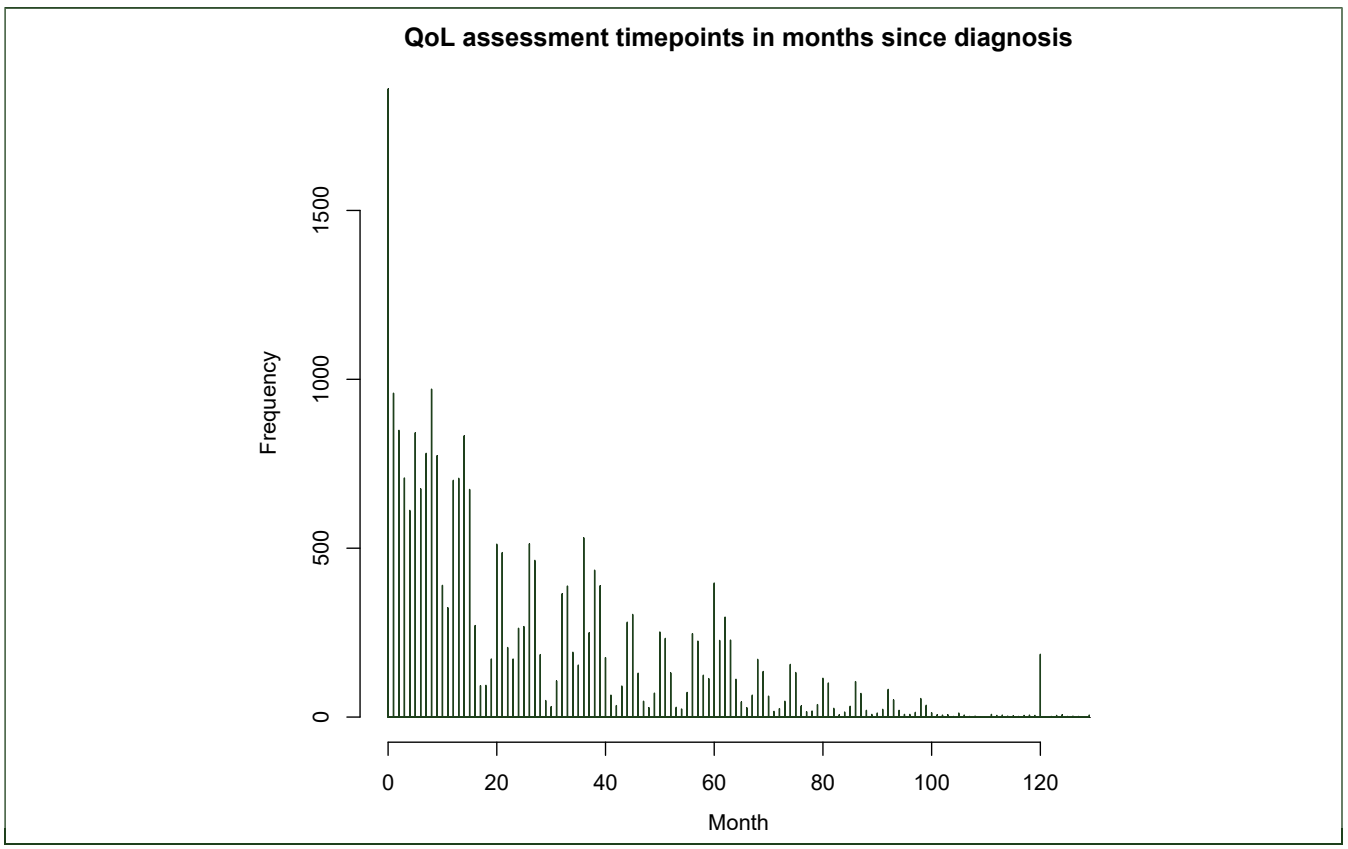

Figure 2. Distribution of HRQoL measurements over time in the database of 6500 patients. HRQoL, health-related quality of life.

Merging datasets was time-consuming, with each update requiring a new database version. Though merging each dataset took <1 week, the total effort required at least 1 month of full-time work with all checks.

Summary of challenges

A summary of challenges is provided in Table 3.

DISCUSSION

The EORTC BALANCE project showed that integrating international HRQoL datasets is feasible but challenging due to issues like missing patient consent for secondary use, limited resources for data acquisition, and data heterogeneity. Despite these obstacles, the database currently includes >30 000 observations from 6500 patients across a diverse breast cancer population (2001-2024). If ongoing negotiations succeed, we aim to expand the dataset by 4500 patients, reaching 11 000 patients and 50 000-60 000 observations.

Data identification

A comprehensive overview of available datasets is crucial for data pooling projects. For clinical trial data and large registries, platforms like PDS and Vivli, along with literature searches, provide well-indexed resources. However, for RWD data, the situation is different.

The systematic collection of PROMs in RWD is a trend of the past decade, primarily driven by care centers instead of (inter)national organizations. While some centers collect RWD for academic and clinical purposes, much of them remains unreported in literature as they are often only collected for clinical use. Without a comprehensive registry of HRQoL RWD, identifying relevant data through literature remains challenging. In BALANCE, for example, RWD was identified using researcher and funder networks (NKI and BKH Kufstein datasets).

An (inter)national registry of centers that collect PROMs would greatly support projects like BALANCE in identifying suitable RWD. The PROFILES registry, integrated with the Netherlands Cancer Registry, is an example where centralized HRQoL data led to multiple pooled analyses and yielded an HRQoL prediction model.[11,38,39] Through this registry, we identified the OPTIMUM and UMBRELLA studies. Expanding efforts like PROFILES (inter)nationally could further enhance the secondary use of HRQoL data. International organizations (e.g. EORTC) are well equipped to lead such initiatives. The European Health Data Space (EHDS) is another key initiative, aiming to establish a framework for the exchange of electronic health data within the European Union (EU), specifically for secondary data use.[40] Ideally, HRQoL data registries will be integrated into health data access bodies to create a European network of easily accessible HRQoL data registries.[41]

Data acquisition

Intellectual property. Ownership and IP rights to data present another significant barrier to data sharing. Several approaches exist for structuring data pooling projects to navigate this barrier.

Table 3. Overview of encountered challenges and proposed solutions

Challenge	Potential solution(s)	Solution chosen	Reasoning for solution chosen
Data identification			
Finding suitable data	• Comprehensive registry of real-world HRQoL data • Literature search • Leveraging researcher networks	• No • Yes • Yes	• Currently not available • Feasible within the project • Feasible within the project
Data acquisition			
Intellectual property and data ownership	• Ownership remains with the original data holders, contributors act as processors only • Consortium structure with shared ownership • Federated learning, no changes in ownership • Purchase of all ownership	• Yes • No • No • No	• Standard procedure for data sharing under GDPR • Required a different project setup and funding • Required a different project setup, missing expertise, and funding • Not budgeted
Various interpretations of data protection	• Separate joint controllership agreement • Additional clauses within data-sharing agreements • Federated learning, no data sharing	• No • Yes • No	• Required a different project setup • Different interpretation chosen by leading legal counsel • Required a different project setup
Patient consent for secondary use	• Approval of data sharing by medical ethics boards • Re-consenting of all patients	• Yes • No	• When needed, the easiest solution • Time-consuming, (probably) impossible to track all patients
Resource limitations	• Financial compensation • Joint efforts by international research organizations	• No • Yes	• Not budgeted • EORTC QLG promotes international collaboration and data sharing
Data harmonization			
Heterogeneity and missing data	• Common data models and frameworks used by original data collection • Common data models and framework used by pooling study	• No • Yes	• Not possible; no original data collection • Standardizing available data to a certain common level

EORTC, European Organisation for Research and Treatment of Cancer; GDPR, General Data Protection Regulation; HRQoL, health-related quality of life; QLG, Quality of Life Group.

The first approach, used by BALANCE, involves contributors acting as data processors. In this model, all project-generated outputs are the IP of the funder and the processors, while the raw data remain the IP of the original institutions. While effective for this project, this model may not be suitable for follow-up work using the pooled data. Within the DSA, the purpose of data processing is explicitly defined, often restricting its use to a specific research question or focus. As a result, any follow-up work that falls even partially outside this scope cannot be conducted using the data, unless you renegotiate a new DSA for the new research question.

A second option is to establish a consortium structure, in which data contributors receive 'a seat at the table' in exchange for contributing data to a large pooled dataset. All partners of the consortium share rights over the generated IP and raw data.

A third option would be to purchase all ownership and IP rights of all existing data. While theoretically this solves all issues, it seems uncommon in the HRQoL space with only one dataset encountered for sale.

Data protection. A key challenge in data sharing is the various interpretations of the European GDPR 2016/679, effective from 2018.[31] The GDPR aims at promoting both the free flow of personal data within the EU and beyond, and protecting the individual and their personal data.[42] Thus, individuals have the right to supervise their data and be informed about their use, but this does not mean that the individual can exercise complete control over their data.[43]

The lawful basis of consent is only one of several legal grounds for processing personal data for secondary research purposes. In the EU, member states have adopted various implementation laws and the interpretation of the GDPR framework differs as well.[44] The GDPR provides for six lawful bases for the processing of personal data (Article 6 GDPR) and several exemptions for the processing of health data for scientific research purposes (Articles 9 and 89 GDPR). In multicenter research, member states use different legal bases for the processing of health data or they have adopted other consent mechanisms.[45,46]

While crucial for data protection, the use of different legal grounds can delay agreements and data transfers. Furthermore, parties in multicenter studies often have different views on the roles of the data controller (Article 24 GDPR), the data processor (Article 28 GDPR), and the joint controllers (Article 26 GDPR). For instance, some legal departments require a separate joint controllership agreement under Article 26 of the GDPR, which defines the responsibilities of both parties for data collection, storage, and use, leading to prolonged negotiations, while others rely solely on additional clauses within the DSA to regulate controllership of the data.

Federated learning is a tool to overcome obstacles due to various interpretations of data privacy, as this enables researchers to train a shared model over decentralized servers while keeping data localized.[47,48] It can be an effective solution to these various interpretations of the GDPR as keeping data localized mitigates data protection concerns. Additionally, it can reduce costs and is scalable; however, issues with heterogeneity, data standardization, and resource-intensiveness persist.[47,48] Sharma and Guleria recently provided a review of federated learning-based models in health care.[49]

Patient consent. Some studies lack explicit clauses on data reusability or have geographical restrictions in patient consent forms. For example, the Netherlands and Austria were not included in the approved countries for data sharing in UK patients' consent forms before Brexit, as EU law regulated data sharing. After Brexit, the UK adopted its own GDPR, causing additional delays as medical ethics boards had to approve data sharing. While we resolved some issues, in other cases, the medical ethics board's decision and data protection officer's advice led to data exclusion, as patients needed to be re-consented for data use, further prolonging the process.

Resource limitations. Contributing departments often face resource constraints, both financial and in personnel, for preparing data for research projects, partly hindering timely progress. Some potential collaborators decided not to participate due to these resource limitations. Additionally, the high volume of contracts requiring review by legal departments at research institutions further prolonged the timeline between research agreements and data transfer. This backlog of contracts highlights that many institutions are under-equipped to handle the growing number of requests to share data.

Providing financial compensation for data preparation and legal reviews could streamline research and support data accrual by funding dataset purchases or reimbursing prior data collection. In hindsight, the BALANCE project was underfunded for data acquisition, with no budget for reimbursing collaborators. Additionally, obtaining funding for secondary data use projects is challenging, despite initiatives like the EU Innovative Medicines Initiative 2's 'Big Data for Better Outcomes'.[50] We urge funding agencies to increase support for secondary data usage projects to allow data (preparation) reimbursement. Paying hospitals/research centers, e.g. €10 000, to make data accessible for secondary use will always be cheaper than starting novel data collection for a big data project.

In addition to funding agencies, international research organizations (e.g. EORTC) play a key role due to their global reach and centralized structure. We urge these organizations to lead efforts in data sharing and provide frameworks for acquisition for secondary use across borders.

Data harmonization

The main challenge in harmonization was not merging the data but dealing with the heterogeneity and missing values. We believe that in every data pooling project, there is an inverse relationship between data quality and volume. For BALANCE, the focus was on a large sample size, so data quality was traded off for volume while ensuring key outcomes (e.g. QLQ-C30, BR23, and minimal clinical data) were included.

Heterogeneity in secondary data pooling arises from varying data collection standards. The first level of heterogeneity stems from whether a variable is included in a dataset, while the second level stems from the differences in detail of the coding of variables. Both affect the analysis: the first by having variables only present in certain datasets, and the second leads to data loss due to standardization to the lowest complexity. To address this, detailed variables were retained alongside simpler ones, like binary event coding (e.g. some datasets contained information about the type of chemotherapy given, whereas others only had binary information). Another challenge is missing data due to the first level of heterogeneity, which cannot be imputed as these variables do not exist in certain datasets.

Data heterogeneity has no simple solutions. To address these challenges, the FAIR data principles (Findable, Accessible, Interoperable, Reusable) should be applied.[51] Common data models like the OMOP model, which is an open community data standard framework for the collection of data, ensure interoperability and reusability of data. Additionally, a core outcome set should be used to standardize data collection of HRQoL in oncology, e.g. the International Consortium for Health Outcomes Measurement (ICHOM) item sets.[52-54] While core outcome sets and common data models may suffice for trial data, RWD require additional tools for structuring and sharing. The Fast Healthcare Interoperability Resources framework can improve this for RWD by creating a seamless electronic exchange of health care data using modern web technologies.[55] Other solutions for tackling messy, non-standardized RWD are natural language processing and large language models, which can extract standardized data from clinical notes from electronic health records[56]

Limitations of our approach

All reported challenges are part of the limitations of this article, affecting data collection and (possible) analysis. As described, interpretation of applicable legislation and insufficient patient consent represent limitations, leading to missing data due to restrictions on data access and usage. Both constraints hindered our ability to gather comprehensive datasets, particularly concerning detailed patient information.

Another limitation is the competition among numerous projects seeking datasets, which can overwhelm potential data contributors and slow down the sharing process. Greater impact could be achieved through collaborative efforts or equitable access to large, shared databases.

Conclusion and outlook

This article outlines the process of pooling and harmonizing international HRQoL data from breast cancer patients, culminating in the creation of one of the largest multicenter breast cancer HRQoL databases to date. It demonstrates that while pooling fragmented HRQoL data across international sources is achievable, it involves overcoming significant challenges, including data privacy regulations, resource constraints, and data heterogeneity. In the next phase of the EORTC BALANCE project, we will use this database for testing various prediction methods, including joint modeling and landmarking. ML efforts will target both regression and classification algorithms to predict multiple time horizons and future questionnaire responses.

The lessons learned from these hurdles and the solutions implemented provide a blueprint for future data pooling projects, not only in HRQoL research but also for similar projects involving other tumor types or diseases.

ACKNOWLEDGEMENTS

We thank Iris van der Ploeg for guidance on which variables to include when designing the study and codebook; Corneel Coens for his help with identifying EORTC breast cancer trials; Pablo Reja for all his advice surrounding the data-sharing agreement and creating a draft data-sharing agreement for us; Coralie Poncet for preparing the EORTC AMAROS data for this project; the NKI Scientific data administration, especially Melanie Singer-van den Hout and Tony van de Velde for their data preparation of the NKI data; Henrike Bretveld and Ghita Puts from the Dutch cancer registry for clinical data preparation of UMBRELLA and OPTIMUM; Nicole Horevoorts and the team of the PROFILES registry for the HRQoL data preparation of UMBRELLA and OPTIMUM; the research team from the UMCU Division of Imaging for providing the key to couple NKR and PROFILES data for UMBRELLA; August Zabernigg from BKH Kufstein for providing the BKH Kufstein data; the NKI knowledge transfer and contracting department for all the help with the data-sharing agreements; and Lonneke van de Poll-Franse, Jaap C. Seidell, Floortje Mols, Sandra van Cappellen-van Maldegem, Janneke van den Boom, and Judith van Valenberg for the execution of the OPTIMUM study.

FUNDING

This work was supported by the European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group [grant number 007-2022] (EORTC 2052). The EORTC QLG business model involves license fees for the commercial use of their instruments. Academic use of EORTC instruments is free of charge. The founding party played no role in analyzing or interpreting the data. The OPTIMUM study was supported by a personal grant from the Dutch Cancer Society [grant number 10960] awarded to MH. The VERDI study was supported by two grants from the German Cancer Foundation (Deutsche Krebshilfe) [grant numbers 70-1816, 70-2413]. The AMAROS study was funded by EORTC with financial support [grant numbers 2U10 CA11488-28 to 5U10 CA011488-38] from the United States National Cancer Institute (Bethesda, MD, USA) and by a donation from the Kankerbestrijding/KWF from the Netherlands through the EORTC Charitable Trust.

DISCLOSURE

The authors have declared no conflicts of interest.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

The authors declare that they have used generative artificial intelligence (ChatGPT, OpenAI) in the writing of this manuscript to improve the readability and language of the work. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

DATA SHARING

The data that support the findings of this study are available from each cohort study's principal investigator. Restrictions apply to the availability of these data, which were used under license for this study.

REFERENCES

1. Institute of Medicine (US) Committee on Psychosocial Services to Cancer Patients/Families in a Community Setting. The psychosocial needs of cancer patients. In: Adler NE, Page AEK, editors. *Cancer Care for the Whole Patient: Meeting Psychosocial Health Needs*. Washington, DC: National Academies Press (US). Available at https://www.ncbi.nlm.nih.gov/books/NBK4011/. Accessed November 13, 2024.
2. Balitsky AK, Rayner D, Britto J, et al. Patient-reported outcome measures in cancer care: an updated systematic review and meta-analysis. *JAMA Netw Open*. 2024;7(8):e2424793.
3. Basch E, Barbera L, Kerrigan CL, Velikova G. Implementation of patient-reported outcomes in routine medical care. *Am Soc Clin Oncol Educ Book*. 2018;38:122-134.
4. Jiang P, Sinha S, Aldape K, Hannenhalli S, Sahinalp C, Ruppin E. Big data in basic and translational cancer research. *Nat Rev Cancer*. 2022;22(11):625-639.
5. Efficace F, Cottone F, Sparano F, Caocci G, Vignetti M, Chakraborty R. Patient-reported outcomes in randomized controlled trials of patients with multiple myeloma: a systematic literature review of studies published between 2014 and 2021. *Clin Lymphoma Myeloma Leuk*. 2022;22(7):442-459.
6. Di Maio M, Basch E, Denis F, et al. The role of patient-reported outcome measures in the continuum of cancer clinical care: ESMO Clinical Practice Guideline. *Ann Oncol Off J Eur Soc Med Oncol*. 2022;33(9):878-892.
7. van der Heijden TGW, de Ligt KM, Hubel NJ, et al. Exploring the role of health-related quality of life measures in predictive modelling for oncology: a systematic review. *Qual Life Res*. 2025;34(2):305-323.
8. Cuocolo R, Caruso M, Perillo T, Ugga L, Petretta M. Machine learning in oncology: a clinical appraisal. *Cancer Lett*. 2020;481:55-62.
9. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *Br Med J*. 2020;368:m441.
10. Velikova G, Williams LJ, Willis S, et al. Quality of life after post-mastectomy radiotherapy in patients with intermediate-risk breast cancer (SUPREMO): 2-year follow-up results of a randomised controlled trial. *Lancet Oncol*. 2018;19(11):1516-1529.
11. van de Poll-Franse LV, Horevoorts N, van Eenbergen M, et al. The Patient Reported Outcomes Following Initial treatment and Long term Evaluation of Survivorship registry: scope, rationale and design of an infrastructure for the study of physical and psychosocial outcomes in cancer survivorship cohorts. *Eur J Cancer*. 2011;47(14):2188-2194.
12. Mierzynska J, Piccinin C, Pe M, et al. Prognostic value of patient-reported outcomes from international randomised clinical trials on cancer: a systematic review. *Lancet Oncol*. 2019;20(12):e685-e698.
13. Velikova G, Morden JP, Haviland JS, et al. Accelerated versus standard epirubicin followed by cyclophosphamide, methotrexate, and fluorouracil or capecitabine as adjuvant therapy for breast cancer (UK TACT2; CRUK/05/19): quality of life results from a multicentre, phase 3, open-label, randomised, controlled trial. *Lancet Oncol*. 2023;24(12):1359-1374.
14. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol*. 2014;14(1):137.
15. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med*. 2015;13:1.
16. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73.
17. Lee SK, Son YJ, Kim J, et al. Prediction model for health-related quality of life of elderly with chronic diseases using machine learning techniques. *Heal Inf Res*. 2014;20(2):125-134.
18. Arkin FS, Aras G, Dogu E. Comparison of artificial neural networks and logistic regression for 30-days survival prediction of cancer patients. *Acta Inf Med*. 2020;28(2):108-113.
19. Tseng YJ, Wang HY, Lin TW, Lu JJ, Hsieh CH, Liao CT. Development of a machine learning model for survival risk stratification of patients with advanced oral cancer. *JAMA Netw Open*. 2020;3(8):e2011768.
20. Hirasawa H, Murata H, Mayama C, Araie M, Asaoka R. Evaluation of various machine learning methods to predict vision-related quality of life from visual field data and visual acuity in patients with glaucoma. *Br J Ophthalmol*. 2014;98(9):1230-1235.
21. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276-1296.
22. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst*. 1993;85(5):365-376.
23. Sprangers MA, Groenvold M, Arraras JI, et al. The European Organization for Research and Treatment of Cancer breast cancer-specific quality-of-life questionnaire module: first results from a three-country field study. *J Clin Oncol*. 1996;14(10):2756-2768.
24. Krepper D, Cesari M, Hubel NJ, Zelger P, Sztankay MJ. Machine learning models including patient-reported outcome data in oncology: a systematic literature review and analysis of their reporting quality. *J Patient Rep Outcomes*. 2024;8(1):126.
25. Krepper D, Giesinger JM, Dirven L, et al. Information about missing patient-reported outcome data in breast cancer trials is frequently not documented: a scoping review. *J Clin Epidemiol*. 2023;162:1-9.
26. Federation of Medical Specialists. Richtlijnendatabase. 2024. Available at https://richtlijnendatabase.nl/. Accessed August 20, 2024.
27. Martos C, Giusti F, Van Eycken E, Visser O. *A Common Data Quality Check Procedure for European Cancer Registries*. Ispra, Italy: European Commission; 2023. JRC132486.
28. Observational Health Data Sciences and Informatics (OHDSI). OMOP Common Data Model. Available at https://ohdsi.github.io/CommonDataModel/index.html. Accessed August 20, 2024.
29. Fayers P, Aaronson NK, Bjordal K, Groenvold M, Curran D, Bottomley A. *EORTC QLQ-C30 Scoring Manual*. European Organisation for Research and Treatment of Cancer. Available at https://abdn.elsevierpure.com/en/publications/eortc-qlq-c30-scoring-manual. Accessed June 19, 2025.
30. Giesinger JM, Loth FLC, Aaronson NK, et al. Thresholds for clinical importance were defined for the European Organisation for Research and Treatment of Cancer Computer Adaptive Testing Core-an adaptive measure of core quality of life domains in oncology clinical practice and research. *J Clin Epidemiol*. 2020;117:117-125.
31. European Parliament, Council of the European Union. Regulation - 2016/679 - EN - gdpr - EUR-Lex. Available at https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng. Accessed June 19, 2025.
32. Straver ME, Meijnen P, van Tienhoven G, et al. Sentinel node identification rate and nodal involvement in the EORTC 10981-22023 AMAROS trial. *Ann Surg Oncol*. 2010;17(7):1854-1861.
33. Young-Afat DA, van Gils CH, van den Bongard HJGD, Verkooijen HM. The Utrecht cohort for Multiple BREast cancer intervention studies and Long-term evaLuAtion (UMBRELLA): objectives, design, and baseline results. *Breast Cancer Res Treat*. 2017;164(2):445-450.

34. van Cappellen-van Maldegem SJM, Mols F, Horevoorts N, et al. Towards OPtimal TIming and Method for promoting sUstained adherence to lifestyle and body weight recommendations in postMenopausal breast cancer survivors (the OPTIMUM-study): protocol for a longitudinal mixed-method study. *BMC Womens Health*. 2021;21(1):268.
35. Boomstra E, Walraven I, van der Ploeg IMC, et al. Moving beyond barriers: a mixed-method study to develop evidence-based strategies to improve implementation of PROMs in clinical oncology care. *Qual Life Res*. 2025;34(1):173-188.
36. Wintner LM, Giesinger JM, Zabernigg A, et al. Evaluation of electronic patient-reported outcome assessment with cancer patients in the hospital and at home. *BMC Med Inform Decis Mak*. 2015;15:1-10.
37. Arndt V, Stürmer T, Stegmaier C, Ziegler H, Dhom G, Brenner H. Socio-demographic factors, health behavior and late-stage diagnosis of breast cancer in Germany: a population-based study. *J Clin Epidemiol*. 2001;54(7):719-727.
38. Van De Poll-Franse LV, Horevoorts N, Schoormans D, et al. Measuring clinical, biological, and behavioral variables to elucidate trajectories of patient-reported outcomes: the PROFILES registry. *J Natl Cancer Inst*. 2022;114(6):800-807.
39. Adiprakoso D, Katsimpokis D, Oerlemans S, et al. Development of a prediction model for clinically-relevant fatigue: a multi-cancer approach. *Qual Life Res*. 2025;34(1):231-245.
40. European Health Data Space Regulation (EHDS)—European Commission. 2025. Available at https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en. Accessed March 6, 2025.
41. Ministerie van Volksgezondheid Welzijn en Sport. *EHDS-factsheet—Publicatie—Data voor gezondheid*. Ministerie van Algemene Zaken. Available at https://www.datavoorgezondheid.nl/publicaties/publicaties/2024/12/02/ehds-factsheet. Accessed March 6, 2025.
42. Hoofnagle CJ, van der Sloot Bart, Borgesius FZ. The European Union general data protection regulation: what it is and what it means. *Inf Commun Technol Law*. 2019;28(1):65-98.
43. Opinion of Advocate General Campos Sánchez-Bordona delivered on 6 October 2022. UI v Österreichische Post AG. Request for a preliminary ruling from the Oberster Gerichtshof. Reference for a preliminary ruling — Protection of natural persons with regard to the processing of personal data — Regulation (EU) 2016/679 — Article 82 (1) — Right to compensation for damage caused by data processing that infringes that regulation — Conditions governing the right to compensation — Mere infringement of that regulation not sufficient — Need for damage caused by that infringement — Compensation for non-material damage resulting from such processing — Incompatibility of a national rule making compensation for such damage subject to the exceeding of a threshold of seriousness — Rules for the determination of damages by national courts. *Case C-300/21*. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=ecli%3AECLI%3AEU%3AC%3A2022%3A756. Accessed June 19, 2025.
44. Abboud L, Cosgrove S, Kesisoglou I, Richards R, Soares F. Summary of milestone 5.1 & 5.2. Annex A | Case studies: Different governance and health data systems in Europe. TEHDAS Consortium Partners. September 28, 2021. Available at https://tehdas.eu/app/uploads/2021/09/tehdas-annex-a-case-studies-different-governance-and-health-data-systems-in-europe-2021-09-28.pdf.
45. Kist I. Assessment of the Dutch rules on health data in the light of the GDPR. *Eur J Health Law*. 2022;30(3):322-344.
46. Abboud L, Bogaert P, Bowers S, et al. Report on secondary use of health data through European case studies. Towards European Health Data Space Consortium Partners. August 30, 2022. Available at https://tehdas.eu/app/uploads/2022/08/tehdas-report-on-secondary-use-of-health-data-through-european-case-studies-.pdf.
47. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. *Npj Digit Med*. 2020;3(1):119.
48. Sheller MJ, Edwards B, Reina GA, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep*. 2020;10(1):12598.
49. Sharma S, Guleria K. A comprehensive review on federated learning based models for healthcare applications. *Artif Intell Med*. 2023;146:102691.
50. BD4BO — Big Data for Better Outcomes. Available at https://bd4bo.eu/. Accessed June 19, 2025.
51. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3(1):160018.
52. ICHOM. Sets of Patient-Centered Outcome Measures. Available at https://www.ichom.org/patient-centered-outcome-measures/. Accessed June 19, 2025.
53. de Ligt KM, de Rooij BH, Hedayati E, et al. International development of a patient-centered core outcome set for assessing health-related quality of life in metastatic breast cancer patients. *Breast Cancer Res Treat*. 2023;198(2):265-281.
54. Ong WL, Schouwenburg MG, van Bommel ACM, et al. A standard set of value-based patient-centered outcomes for breast cancer: the international consortium for health outcomes measurement (ICHOM) initiative. *JAMA Oncol*. 2017;3(5):677-685.
55. Index—FHIR v5.0.0. Available at https://hl7.org/fhir/. Accessed June 19, 2025.
56. Osterman TJ, Ye J. The importance of studying the implementation of cancer data standards. *Cancer*. 2025;131(1):e35441.