Original Research Article

Health Informatics Journal

Health Informatics Journal
1–14

DOI: 10.1177/14604582241235632
journals.sagepub.com/home/jhi

How can current oncological datasets be adjusted to support the automated patient recruitment in clinical trials?

Maria-Luisa Marino, Lara Kazmaier and Antonia Krendelsberger
Comprehensive Cancer Center (CCC Munich LMU), LMU University Hospital, Munich, Germany

Silvia Müller
Comprehensive Cancer Center (CCC Munich LMU), LMU University Hospital, Munich, Germany; Comprehensive Cancer Center, Technical University of Munich Hospital Rechts der Isar, Munich, Germany

Sabine Kesting
Preventive Pediatrics, Department of Sport and Health Sciences, Technical University of Munich, Munich, Germany; Department of Pediatrics and Children's Cancer Research Centre, TUM School of Medicine, Kinderklinik München Schwabing, Technical University of Munich, Munich, Germany

Theres Fey and Daniel Nasseh
Comprehensive Cancer Center (CCC Munich LMU), LMU University Hospital, Munich, Germany

Abstract

Objectives: This study aims to identify necessary adjustments required in existing oncological datasets to effectively support automated patient recruitment. Methods: We extracted and categorized the inclusion and exclusion criteria from 115 oncological trials registered on ClinicalTrials.gov in 2022. These criteria were then compared with the content of the oBDS (Oncological Base Dataset version 3.0), Germany's legally mandated oncological data standard. Results: The analysis revealed that 42.9% of generalized inclusion and exclusion criteria are typically present as data fields in the oBDS. On average, 54.6% of all criteria per trial were covered. Notably, certain criteria such as comorbidities, pregnancy status, and laboratory values frequently appeared in trial protocols but were absent in the oBDS. Conclusion: The omission of criteria, notably

Corresponding author:
Daniel Nasseh, Comprehensive Cancer Center (CCC Munich LMU), LMU University Hospital, Pettenkoferstraße 8a, Munchen 80336, Germany.
Email: daniel.nasseh@med.uni-muenchen.de

comorbidities, within the oBDS restricts its functionality to support trial recruitment. Addressing this limitation would enhance its overall effectiveness. Furthermore, the implications of these findings extend beyond Germany, suggesting potential relevance and applicability to oncological datasets globally.

Keywords
clinical study, clinical trial recruitment, oncology, tumor data

Introduction

Over the past few decades, rapid advancements have been witnessed in the field of oncology, leading to significant improvements in cancer patient treatments. Clinical research has played a pivotal role in driving this progress.[1] Participating in oncological studies offers patients the opportunity to receive early treatment using new and innovative methods. However, the current approach of relying on doctors, tumor boards, or interdisciplinary oncological councils to recruit patients for study participation poses numerous challenges. Even well-informed doctors typically lack knowledge of all ongoing studies and their inclusion and exclusion criteria. Therefore, a limited availability of potential participants does not necessarily stem from a lack of interest or refusal on the part of patients.[2] Other factors contributing to low recruitment numbers may include inadequate organizational structures and ethical conflicts, which may also vary depending on the locality.[3–6] Despite advancements in medical and information technology, patient recruitment in Germany continues to rely heavily on direct patient contact, with limited utilization of available technologies.[7]

Though, the concept of automating patient recruitment to alleviate the associated challenges has been contemplated for some time.[7] For instance, the German Medical Informatics in Research and Care in University Medicine (MIRACUM) consortium[8] is developing an automated recruitment tool that utilizes patient data to determine their eligibility for clinical trials.[9] However, a significant hurdle lies in the increasing complexity of patient disease profiles, requiring more comprehensive documentation. In the field of oncology, initiatives such as the German Network for Personalized Medicine (DNPM) strive to provide personalized therapy for complex medical cases.[10] This paper examines the relevant inclusion and exclusion criteria in current oncological studies and evaluates the extent to which existing enforced oncological standards, in particular, the German oncological base data set - version 3.0 (oBDS), can support automated patient recruitment.[11] The selection of the oBDS dataset for analysis is based on its status as a legally mandated dataset in Germany and its integral role as a major inspiration in numerous large-scale oncological projects. As an example, the oBDS serves as the foundation for the clinical data catalogue of the National Network Genomic Medicine (nNGM), which draws extensively from the oBDS.[12] Additionally, the oBDS serves as a blueprint for the expansion module of the medical informatics core dataset, aiming to connect various medical domains, including e.g. oncology and radiology, with the oBDS dataset as its cornerstone within the oncology domain.[13] While the oBDS may not encompass all necessary data categories for all clinical trials, it remains the nationwide standard for tumor documentation in Germany.[11] Similarly, oncological datasets exist in other countries like the United States, England or others.[14,15] While their structures differ, these datasets are similar in complexity and general data contents among each other, suggesting that the findings in this study may have relevance beyond Germany.

This paper analyzes the potential utilization of enforced oncological datasets, specifically the oBDS dataset, to support automated patient recruitment in clinical trials. It assesses the extent to which enforced oncological datasets can fully cover studies and discusses the need for dataset expansion.

Methods

The first part of this work was the detection of relevant studies. The source of the study collection is the website ClinicalTrials.gov. It is provided by the U.S. National Library of Medicine and serves as the world's most important registry which provides access to information about clinical trials.[16]

The initial detection of studies was performed on January 31 of 2022. The first filter set was the subject area of studies; in this case, oncological studies were required. There was no further specification regarding the cancer variations or the organ entities. To narrow it down it was necessary that the study was either recruiting or not yet recruiting. Additionally, it needed to be an interventional study, which contained a study protocol. The study protocol was fundamental since it holds information about the inclusion and exclusion criteria. Furthermore, the studies must have taken place between January 1 of 2019 and the day of the initial detection, January 31 of 2022.

After a thorough selection of the studies, the review of each study protocol with regard to their inclusion and exclusion criteria began. The review of the study protocols started out with the identification of each inclusion and exclusion criteria of every study. Using Microsoft Excel 2016, the transcription of the study protocol's criteria into a database was carried out. During the transfer, no attention was paid to whether it was an inclusion or exclusion criteria, but rather that it was a criteria in the general sense as all exclusion criteria can be formulated as negated inclusion criteria.

This process was followed by a generalization of each study criteria. The generalization serves to group multiple criteria into fitting categories. To demonstrate the idea of the generalization the following example can be given: Extracted criteria regarding the age, in this case for example '18+ (NCT04342429)' and 'Adult 21+ (NCT04745754)' were transcribed into a newly generalized category called 'age'. This process served to generalize criteria that fundamentally share a common goal. The new categories were not established based on existing standards, for example, those derived from the EHR4CR project or other sources.[17–19] Instead, they were empirically formed by extracting all inclusion and exclusion criteria in their raw form and grouping them based on their similarities. This approach was chosen to remain open to potential new categories, such as COVID-19, and also to provide a specific focus on oncology, as exemplified in the case of molecular markers.

While the primary author undertook the direct extraction of raw criteria from the trial protocols, discussions on generalization were conducted in collaboration with senior staff at the local Comprehensive Cancer Center. Additionally, insights were sought from the center's physicians, tumor documentalists, and medical informatics practitioners to ensure a comprehensive and well-informed approach. After these preparatory steps, a comparative data analysis was carried out with the goal to identify the previously generalized categories in data fields of the oBDS dataset. This comparative data analysis additionally aimed to identify the extent to which the oBDS dataset can provide information usable in an automated patient recruitment process. It also allows identifying those frequently occurring generalized categories that are missing in the oBDS.

To alleviate the effort of interpretation, a descriptive analysis was conducted using the free software programming language R (version 4.2.1), developed by the R Core Team.[20] The goal was to evaluate which of the categories have the highest impact when it comes to patient recruitment, either in terms of individual studies or more generally the overall impact.

Results

On the day of the initial detection, 402,624 studies were registered on the clincaltrial.gov website. First, the filter "oncology" was set, which resulted in an inclusion of 98,880 studies. Then, another filter was set regarding the start date, which reduced the number of studies to 23,985. The filters "recruiting and not recruiting", "interventional" and "study protocol" reduced the number of studies to 506. The last filter was the end date, which was the date of the initial detection, and reduced the size to 145 studies. Later on, another 30 studies were excluded. This was due to some trials being labeled as non-oncological trials as well as some trials being deleted from the clinialtrial.gov website. The final study collection contains 115 studies. The exact number of studies selected after each subsequent filter can be seen in Figure 1.

After the generalization, 56 different criteria groups were created regarding the inclusion and exclusion criteria. The five most common criteria groups in study protocols are the age of the patient (n = 115/100%), the oncological diagnosis (n = 105, 91.3%), a patient's comorbidities (n = 90, 78.3%), a signed declaration of consent (n = 77, 67.0%) and the assessment if there has been a previous surgery (n = 66, 57.4%). In certain instances, a single criterion extracted from a study protocol may be classified into multiple categories. The following Table 1 shows the prevalence of each criteria group in the given study protocols.

The color difference on the bar chart in Figure 2 is intended to show coverage by the oBDS dataset in addition to the prevalence of the criteria. The lighter bars show the criteria found in the oBDS dataset while the darker bars show which criteria were not found in the oBDS dataset.

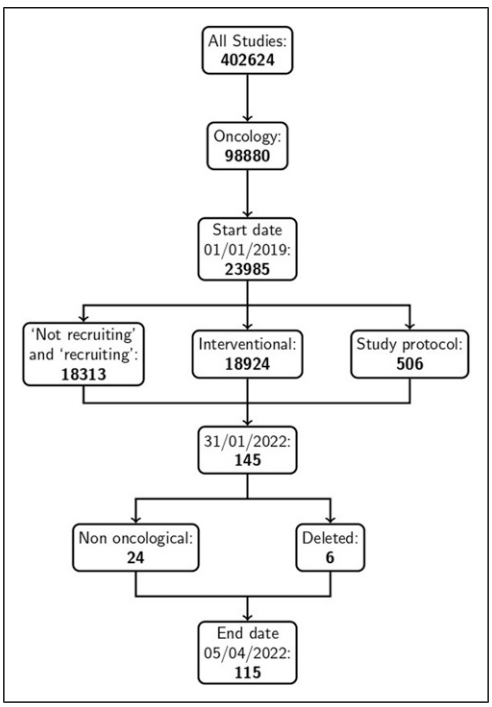

Figure 1. Process of study selection.

The results of the comparative analysis show that 24 out of 56 identified criteria groups from the ClinicalTrial.gov study protocols are listed in the oBDS dataset. Out of the five most common required criteria three were covered by the oBDS dataset.

The 24 generalized criteria groups which are listed in the oBDS dataset are listed in Table 2. Therefore, this means that the remaining 32 criteria groups are not listed in the oBDS dataset. Table 2 also includes information under which term it is noted in the oBDS dataset, as well as, the indication number of the oBDS for reference.[11]

Another focus of this investigation was how many criteria, regardless of whether they were inclusion or exclusion criteria, were required by the study protocol for each study and how many of these could be supported by the oBDS dataset.

Two studies were fully covered by the oBDS dataset. The study with the ClinicalTrials.gov study ID NCT04424758 requires three criteria which were included by the oBDS dataset and the study with the ClinicalTrials.gov study ID NCT04743999 requires six criteria which were included by the oBDS dataset. The study protocol that required the most criteria has the ClinicalTrials.gov study ID NCT04357873. This protocol requires 35 (out of 56) criteria with 15 (42.9%) being covered by the oBDS dataset. The bar chart in Figure 3 visualizes the number of required criteria per study protocol and whether they are covered by the oBDS dataset.

Discussion

At the outset of the study, three primary questions were posed: To what extent can the oBDS dataset comprehensively cover various studies? Which categories prove useful to expand the oBDS in context of automated trial recruitment? To what degree can tumor documentation data already be used to support automated trial recruitment?

Regarding the first question about a possible coverage by the oBDS dataset: It can be stated that barely any study can be fully covered by the oBDS, in fact only 2 (1.7%) studies can. Also only 24 of the 56 categories are supported by the oBDS dataset (42.9%). The listed names of the criteria in the study protocols displayed high heterogeneity but could be generalized into similar groups.

The most commonly asked for criteria, such as age or diagnosis, are indeed present in the oBDS dataset, however, other criteria such as pregnancy, informed consent, and comorbidities are not present in the oBDS dataset, despite frequent demand. Age was the only criteria that was needed in every analyzed study, thus it was requested 115 (100%) times. This is not an unexpected result, as the age information is already mandatory when submitting the study protocol to the ethics committee for the assessment of the clinical trial.[21] A question about possible comorbidities appeared in 90 (78.3%) study protocols. Despite its high frequency in the trial protocols, this field is yet not included in the oBDS dataset. This field provides a good example that regardless of a possible automated patient recruitment, it would be useful for research intentions to include comorbidities in the oBDS dataset. In fact, some cancer related comorbidities might potentially be available alongside some of the international tumor set equivalents like the British Cancer Outcomes and Services Data (COSD). In this case the National Disease Registration Service (NDRS) compiled comorbidity data by using Hospital Episode Statistics (HES) and other data sources but the usage of this data is not recommended.[14,22] Similarly, comorbidities might also be abundant in other datasets like the German MI core dataset, documentation of this data is not enforced yet, hence, its completeness might be lacking and lead to new issues.

Incorporating laboratory values and pregnancy status into the oBDS dataset could also offer improvements. However, it's crucial to note that these values may fluctuate between the time of

Table 1. Generalized criteria groups, short description and their prevalence.

Criteria category	Short description	Prevalence (n = 115)
Age	Patient's age	115 (100.0%)
Oncological diagnosis	Specific oncological diagnosis, including both free-text and coding systems like ICD-10, ICD-O-3, or snomed	105 (91.3%)
Comorbidities	Presence of non-oncological comorbidities	90 (78.3%)
Consent	Signed declaration of consent	77 (67.0%)
Surgery	Surgical procedures specifically related to the tumor diagnosis	66 (57.4%)
Laboratory	Existence of specific laboratory values for blood, kidney, liveretc.	62 (53.9%)
Pregnancy	Pregnancy status including breastfeeding	59 (51.3%)
Pathology	Information present in the pathology report	53 (46.1%)
Disease severity	Measurement based on specific standards like UICC, BCLC, or ASA	50 (43.5%)
System therapy	Previous systemic therapy in the context of the tumor diagnosis	48 (41.7%)
Other criteria	Miscellaneous highly individual criteria, such as querying if the patient was in specific hospitals, family status, conducted counseling sessions, ethnic minorities, or completion of a patient questionnaire	47 (40.9%)
Secondary tumor	Existence of a secondary tumor	46 (40.0%)
Metastases	Existence of metastases	45 (39.1%)
Patient status	Measurement of the well-being of a patient, e.g., ECOG performance scale/Karnofsky	45 (39.1%)
Treatment of comorbidities	Specific treatment of a comorbidity	41 (35.7%)
Radio therapy	Previous radiotherapy in the context of the tumor diagnosis	38 (33.0%)
Other diagnosis	Other diagnoses studied in connection with an oncological condition	37 (32.2%)
Gender	Patient's gender	36 (31.3%)
Imaging	Availability of specific findings from imaging, such as CT, MRI, endoscopy, or scan	35 (30.4%)
Allergies	Presence of allergies	34 (29.6%)
Tumor information	Information about tumor size, diameter, or spread (e.g., based on the TNM)	33 (28.7%)
Participation in other studies	Participation in other studies	33 (28.7%)
Medication	Medication other than chemotherapy	31 (27.0%)
Compliance	Compliance with specific study conditions	28 (24.3%)
Chance of pregnancy of women	Contraception, possibility of pregnancy in women of reproductive age, menstruation	24 (20.9%)
Recurrence	Recurrence and risk of recurrence	23 (20.0%)
Contraindications	Specific contraindications, e.g., related to implants versus MRI	21 (18.3%)
Language	Patient's language	21 (18.3%)
Start and end of therapy	Start or endpoint of previous therapies	19 (16.5%)
Therapy goal before therapy	Curative, adjuvant, or palliative therapy before the current treatment	19 (16.5%)
Biomarkers	Biomarkers and molecular diagnostics	19 (16.5%)
HIV, HBV, HCV	Presence of HIV, HBV, HCV infections	18 (15.7%)

(continued)

Table 1. (continued)

Criteria category	Short description	Prevalence (n = 115)
Therapy recommendation	Therapy recommendations, e.g., from a tumor board	17 (14.8%)
Implants/prosthetics	Presence of specific implants	17 (14.8%)
Prognosis	Estimation of life expectancy	17 (14.8%)
Contraception men, sperm donation	Male contraception, current sexual status, and state of semen	13 (11.3%)
Untreated	Previously untreated in relation to the oncological diagnosis	12 (10.4%)
Organ transplant	Previously performed organ transplantations	11 (9.6%)
Toxicity	Toxicity assessment	10 (8.7%)
Date of diagnosis	Date of cancer diagnosis	10 (8.7%)
Living conditions	Living conditions, incarceration, belonging to a vulnerable population	10 (8.7%)
Complications in operation	Complications from prior surgeries	9 (7.8%)
Vaccines	Receipt of vaccinations	9 (7.8%)
Ability to swallow	Ability to swallow tablets/capsules and oral medication	9 (7.8%)
Drug and alcohol consumption	Reporting on drug and alcohol consumption	9 (7.8%)
Family history	Family history of diseases or conditions that pose specific risks	7 (6.1%)
RECIST	Patients meeting RECIST criteria	6 (5.2%)
Palliative treatment	Palliative treatment	4 (3.5%)
BMI, weight	Body Mass index (BMI) and weight	4 (3.5%)
Sars-CoV-2	Current or past sars-CoV-2 infection	4 (3.5%)
Health insurance	Presence of health insurance	4 (3.5%)
Supplements	Consumption of specific supplements, such as vitamins	3 (2.6%)
Digital equipment	Presence of digital equipment, e.g., smartphones	3 (2.6%)
Sport activity	Inquiry about physical activity	2 (2.6%)
Nicotine consumption	Smoking status	2 (2.6%)
Residence	Patient's place of residence	2 (2.6%)

documentation and the screening phase for an oncological trial, thus, their inclusion presents only potential benefits depending on the time of documentation.

The frequency with which consent was listed as an inclusion criterion (77/67.0%) was noteable, given that consent is inherently obligatory, particularly in interventional studies. Consequently, including this criterion in the oBDS, even if it appeared frequently, does not enhance the dataset's content.

Table 3 shows the possible differences that could be achieved if the oBDS dataset was expanded, using the fields laboratory, pregnancy and comorbidities. The study that can be covered the least by the oBDS dataset is currently only 10% coverable with data. If the oBDS dataset is expanded to include the 3 categories mentioned above, the coverage of this study by the oBDS dataset increases to 30%. At the median, 54.6% of all categories can be covered in the current oBDS dataset. If the oBDS dataset is expanded 62.3% will be covered at the median. After potentially adding the 3 categories, trials in the fourth quartile would even have a coverage of over 75.7%.

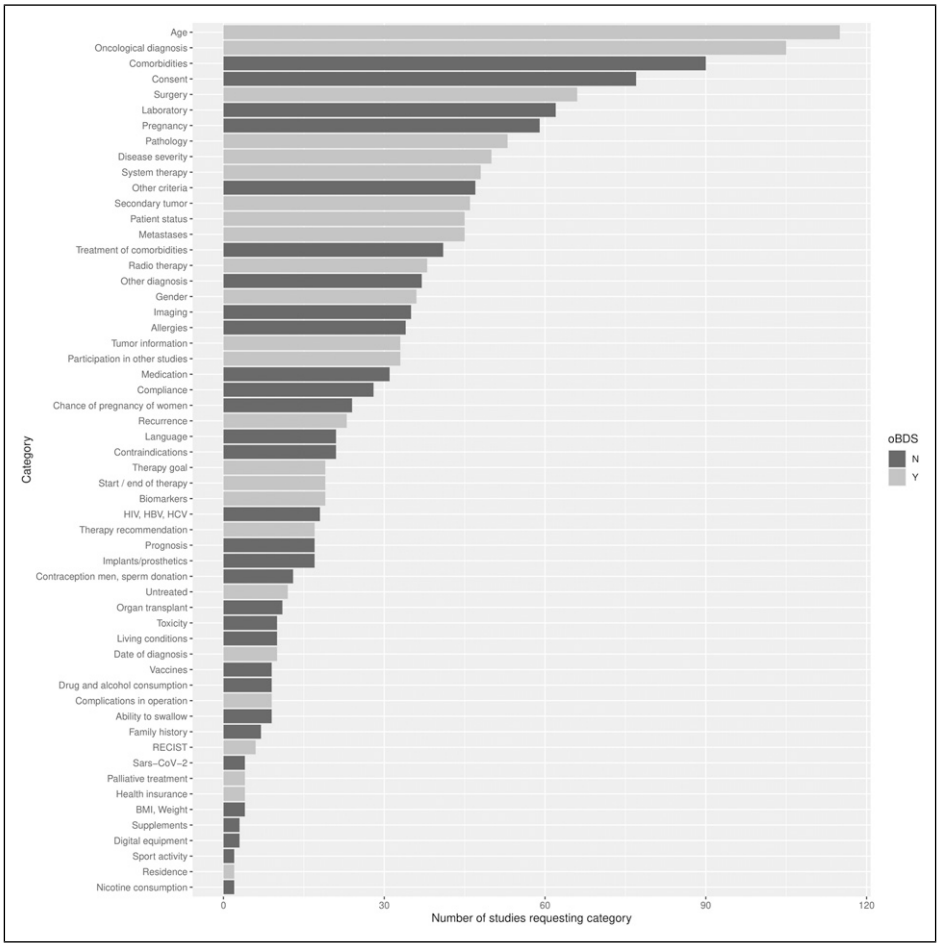

Figure 2. Prevalence of the generalized criteria groups and it's occurence in the oBDS dataset.

Certain studies exhibited an unusually high number of inclusion and exclusion criteria, exemplified by the clinical trial with the ID NCT04357873, which necessitated the inclusion of 35 (out of 56) distinct criteria. There is a trend that shows that studies that require a larger number of criteria are proportionally less likely to be covered by the oBDS dataset. In most cases, there are more than 50% of categories that are not included in the oBDS dataset. Though, 100% compliance with the oBDS dataset does not seem desirable, because extending the oBDS dataset with all these fields would move it away from its original purpose. An extension consisting of the most frequently occurring fields would be desirable in the sense of a beneficial extension in terms of patient recruitment. Striving for a 100% inclusion of trial criteria within the oBDS may not be essential, especially when dealing with time-sensitive variables like laboratory values or pregnancy status, as previously mentioned. Instead, a preliminary selection of studies can serve as a decision support system for physicians, underscoring that it complements rather than supplants the doctor's role by suggesting potential study candidates. In this context, prioritizing the most commonly used criteria can prove advantageous. This approach aligns with the findings of Gulden et al., which also

Table 2. Comparison of the generalized criteria groups and the oBDS dataset ordered by it's prevalence according to Table 1.

Generalized criteria categories from study protocols	oBDS indication (translation)	oBDS indication number
Age	Date of birth	3.10
Oncological diagnosis (ICD10, ICD-O-3 or snomed)	Primary tumor, tumor diagnosis, ICD code	5.1
Surgery	Surgery	13.0
Pathology	Histology	6.0
Disease severity (UICC, BCLC, ASA)	TNM-classification	8.0
	UICC stadium	8.17
System therapy	System therapy	16.0
Secondary tumor	Previous tumors	5.9
Metastases	Distant metastases	11.0
ECOG performance scale/karnofsky	Overall performance	12.1
Radio therapy	Radiation therapy	14.0
Gender	Gender	3.9
Details about tumor and tumor spread	TNM-classification	8.0
	Residuals	10.0
Participation in other studies	Study participation status	24.1
	Study participation date	24.2
Recurrence and risk of recurrence	Progress	17.0
	Overall assessment of tumor status	17.2
Start and end of therapy	Radiation therapy start	14.5
	Radiation therapy end	14.6
	Systemic therapy start	16.5
	Systemic therapy end	16.6
Therapy goal before therapy	Intention of the operation	13.1
	Intention of the radiation therapy	14.1
	Intention of the systemic therapy	16.1
Biomarkers and molecular diagnostics	Genetic variant	23.0
Therapy recommendation/tumor board	Tumor conference, therapy planning	18.0
	Therapy recommendation	19.0
Toxicity	Side effects of radiation therapy and systemic therapy	15.0
Date of diagnosis	Date of diagnosis of primary tumor	5.6
Complications in operation	Complications in operation	13.5
Palliative treatment	Intention of the operation	13.1
	Intention of the radiation therapy	14.1
	Intention of the systemic therapy	16.1
Health insurance	Health insurance number	3.1
Residence	Street	3.11
	House number	3.12
	Country	3.13
	Postal code	3.14

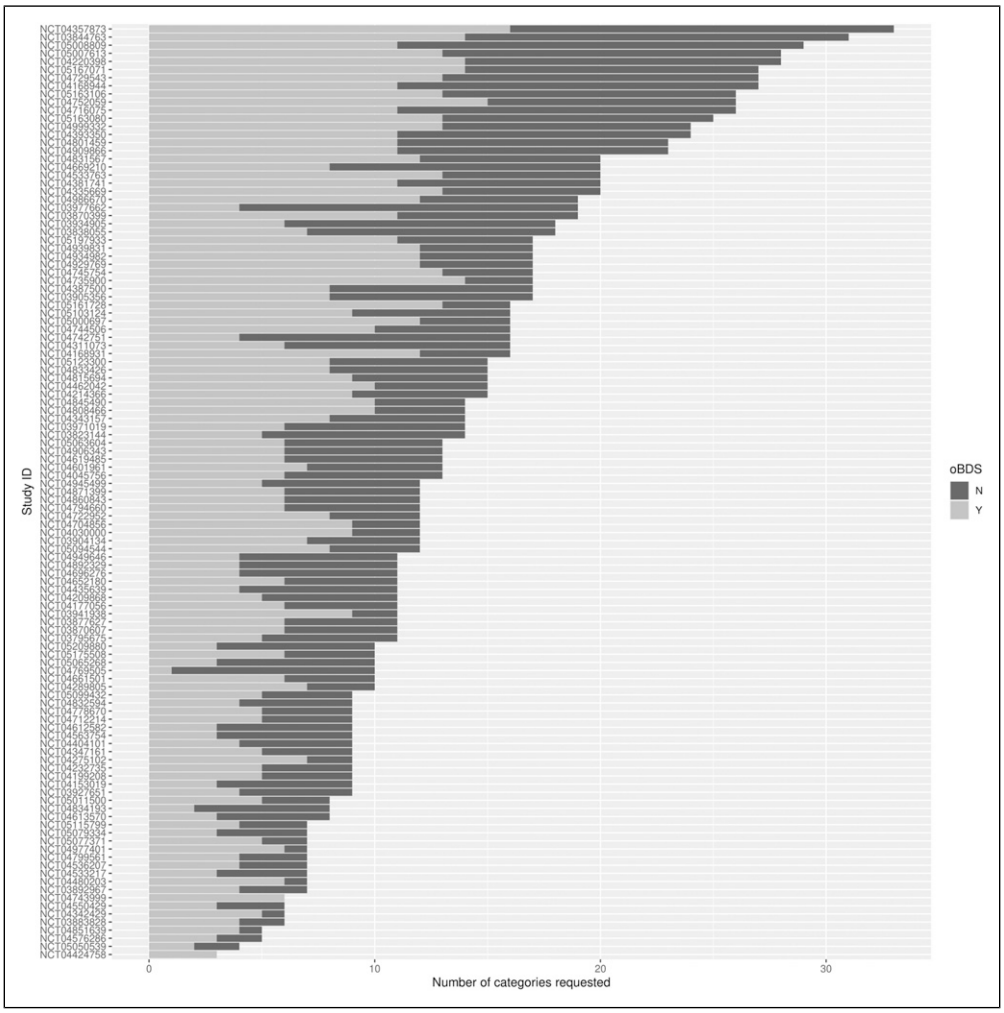

Figure 3. Required criteria in each study and to what extend they are covered by the oBDS dataset. (Y = yes, N = no).

Table 3. Quartiles before and after extending the ADT dataset.

	0%	25%	50%	75%	100%
Criteria coverage with the current oBDS dataset	0.1	0.448	0.546	0.639	1
Criteria coverage after adding laboratory, pregnancy and comorbidities	0.3	0.556	0.625	0.757	1

acknowledge the potential for errors in the formulation of inclusion and exclusion criteria or in the documentation of this information.[23]

In order to answer the question regarding the possible support of patient recruitment by tumor documentation, it is important to consider possible limitations. Overall, it should be noted that tumor documentation in Germany is usually delayed by multiple months and therefore data may not be

available in time.[24] This might prompt the question of why even consider tumor documentation data for trial recruitment. However, due to being manually curated, a tumor dataset inherently comprises structured information of comparatively high data and content quality, setting it apart from typical Electronic Health Record (EHR) entries. As an example, diagnosis codes found in medical billing are often inaccurate. It is worth considering whether interfaces to subsystems such as a chemotherapy system could reduce the necessary time for documentation. Some tumor documentation systems already do this for at least some subsystems (e.g., CREDOS directly pulls surgery data).[25]

A potential, forward-thinking approach to mitigate documentation time is the utilization of natural language processing (NLP) for the processing of medical reports. Given the substantial volume of data within unstructured free text in healthcare, employing machine-learning algorithms to cleanse and integrate this data into tumor documentation systems emerges as a possible solution.[26] In that regard, the commercial text mining system Averbis Health Discovery has a cooperation with the cancer registry of the federal state of Baden-Württemberg, the registry of Rhineland-Palatinate and the registry of Lower-Saxony. It aims to improve and speed up the data extraction from free text fields.[27] In general, natural language processing could also be a possible option to support digital trial recruitment by applying it on the study protocols and automatically parsing the inclusion and exclusion criteria towards a fixed catalogue.[17–19]

Perhaps other datasets (aside from the oBDS) could also be considered as the base for a similar project, such as the national core dataset of the Medical Informatics Initiative.[13] In this paper the oBDS dataset was used as an example for the previously mentioned reasons. Another option to focus on would, as an example, have been the Observational Medical Outcomes Partnership (OMOP) Common Data Model.[28] Though, while this model potentially covers all medical domains, it is, at least in Germany, not as widespread as compared to the oBDS dataset. It can be questioned whether there are oncological oriented datasets in other countries that could be considered for such projects.

As of the year 2023, the exclusive reliance on automated trial recruitment based on obligatory collected tumour data may prove unattainable until certain essential prerequisites are satisfied. These conditions, amongst others, include the necessity to expand existing datasets, for instance, by incorporating additional variables such as comorbidities, enhancing the quality of documented data in terms of completeness and accuracy, diminishing the temporal lag between the act of documentation and the occurrence of the associated events, such as laboratory values, and standardizing the criteria for inclusion and exclusion in alignment with an internationally or nationally accepted common framework, as established within study protocols and registries.

Conclusion

To assess how useful common oncology datasets are for automatically selecting patients for clinical trials, we examined 115 oncology trial protocols and compared them with the German oBDS dataset. Within the present oBDS dataset, on average, 54.6% of all investigated categories can be accommodated. However, it is worth noting that the majority of studies do not achieve full coverage. Moreover, the outcomes of this study underscore that the ability of the oBDS dataset to encompass an expanding array of inclusion and exclusion criteria in clinical studies is notably enhanced when certain categories, in particular comorbidities, are incorporated. Consequently, the current utility of the oBDS lies in its capacity to function as a tool for the preliminary selection of potential trial recruits. In this sense it might be viable as a minimum set for trial recruitment, aside from its lack of comorbidities. Nevertheless, it is important to emphasize that the complete automation of trial recruitment, as of the year 2023, remains an unattained objective.

Acknowledgements

We would like to express our gratitude to the staff of the Comprehensive Cancer Center Munich, supporting this project.

Author contributions

DN conceptualized the project, while MLM oversaw the data collection and preparation process. DN, SK, SM, and TF brought their senior expertise to the project, particularly in the realm of category generalization. AK played a crucial role in analytics and plot creation. All contributors collaborated on manuscript writing, with DN taking the lead in this aspect.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Ethical statement

Ethical approval

In accordance with the Ethics Committee of LMU Hospital, this work has received approval under the ethical board committee number 22-0879 KB. Given the nature of the utilized data, there is no cause for concern regarding ethical considerations in this study.

ORCID iDs

Maria-Luisa Marino https://orcid.org/0000-0001-7481-3264
Lara Kazmaier https://orcid.org/0000-0001-9110-1365
Antonia Krendelsberger https://orcid.org/0000-0003-2172-0536
Sabine Kesting https://orcid.org/0000-0003-3286-2249
Theres Fey https://orcid.org/0000-0002-9947-3275
Daniel Nasseh https://orcid.org/0000-0002-2167-3146

Data availability statement

Information about the selected trials can be requested by the corresponding author.

References

1. Mross K and März W. Klinische Studien: Fundament einer Evidenz-basierten Onkologie – bestandsaufnahme und Zukunft im Zeitalter des Internet. *Onkologie* 2001; 24(1):24–34. Available from: DOI: 10.1159/000055162.
2. Medical Informatics Initiative Germany. *Patientenrekrutierung für klinische Studien*. Germany: Medical Informatics Initiative Germany, 2022. Available from: https://www.medizininformatik-initiative.de/de/patientenrekrutierung-fuer-klinische-studien.
3. Wörmann B, Wulf G and Hiddemann W. Klinische studien in der onkologie. *Med Klin* 1998; 93: 181–189.
4. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemp Clin Trials Commun* 2018; 11: 156–164. DOI: 10.1016/j.conctc.2018.08.001.

5. Briel M, Elger BS, McLennan S, et al. "Exploring reasons for recruitment failure in clinical trials: a qualitative study with clinical trial stakeholders in Switzerland, Germany, and Canada". *Trials* 2021; 22(1): 844. DOI: 10.1186/s13063-021-05818-0.
6. Somkin CP, Ackerson L, Husson G, et al. Effect of medical oncologists' attitudes on accrual to clinical trials in a community setting. *J Oncol Pract* 2013; 9(6): e275–e283. DOI: 10.1200/JOP.2013.001120.
7. Schleinkofer T, Villain S, Lamla G, et al. *S3ULMU–Prototyp-Infrastruktur für die IT-Unterstützung klinischer Studien am Klinikum der Universität München*. Stuttgart: Informatik, 2012.
8. Prokosch H-U, Acker T, Bernarding J, et al. MIRACUM: medical informatics in research and care in university medicine. *Methods Inf Med* 2018; 57(S 01): e82–e91. DOI: 10.3414/ME17-02-0025.
9. Fitzer K, Haeuslschmid R, Blasini R, et al. Patient recruitment system for clinical trials: mixed methods study about requirements at ten university hospitals. *JMIR Med Inform* 2022; 10(4): e28696. DOI: 10.2196/28696.
10. Gemeinsamer Bundesausschuss Innovationsausschuss. *DNPM – deutsches Nertzwerk für Personalisierte Medizin*. Germany: Gemeinsamer Bundesausschuss Innovationsausschuss, 2022. Available from: https://innovationsfonds.g-ba.de/projekte/neue-versorgungsformen/dnpm-deutsches-netzwerk-fuer-personalisierte-medizin.419.
11. Basisdatensatz. *Einheitlicher onkologischer Basisdatensatz*. Germany: Basisdatensatz, 2021. Available from: https://basisdatensatz.de/basisdatensatz.
12. IT. *DigitNet - Digitale Vernetzung in der Onkologie*. Germany: IT, 2022. Available from: https://diginet.nngm.de/arbeitsgruppen/it-2/
13. Medical Informatics Initiative Germany. *Der Kerndatensatz der Medizininformatik Initiative*. Germany: Medical Informatics Initiative Germany, 2022. Available from: https://www.medizininformatik-initiative.de/de/der-kerndatensatz-der-medizininformatik-initiative
14. National Cancer Registration Analysis Cancer outcomes Service. *UK Cancer Registration dataset: National and Analysis Service*. https://www.ncin.org.uk/collecting_and_using_data/data_collection/cosd. Accessed October 17, 2023.
15. Surveillance, epidemiology, and end results (SEER) program SEER*stat database. USA. *Published* October 2010. https://www.seer.cancer.gov. Accessed February 1, 2012.
16. National Library of Medicine. ClinicalTrials.gov. USA: National Library of Medicine, 2022. Available from: https://clinicaltrials.gov/.
17. Doods J, Lafitte C, Ulliac-Sagnes N, et al. A European inventory of data elements for patient recruitment. *Stud Health Technol Inf* 2015; 210: 506–510. DOI: 10.3233/978-1-61499-512-8-506.
18. Bruland P, McGilchrist M, Zapletal E, et al. Common data ele ments for secondary use of electronichealth record data for clinical trial execution and serious adverse event reporting. *BMC Med Res Methodol* 2016; 16: 159. DOI: 10.1186/s12874-016-0259-3.
19. Luo Z, Yetisgen-Yildiz M and Weng C. Dynamic categorization of clinical research eligibility criteria by hierarchical clustering. *J Biomed Inf* 2011; 44: 927–935. DOI: 10.1016/j.jbi.2011.06.001.
20. R Core Team. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2021, https://www.R-project.org/.
21. Ethikkommission der Bayerischen Landesärztekammer. AMG *Ethik-Kommission der BLÄK federführend - erforderliche Unterlagen zur Bewertung klinischer Arzneimittelstudien ab Inkrafttreten des Zweiten Gesetzes zur Änderung 59 arzneimittelrechtlicher und anderer Vorschriften*. Germany: Ethikkommission der Bayerischen Landesärztekammer, 2014. Available from: https://ethikkommission.blaek.de/studien/amg-studien/antragsunterlagen-ek-federfuehrend.
22. Herbert A, Wijlaars L, Zylbersztejn A, et al. Data resource profile: hospital episode statistics admitted patient care (HES APC). *Int J Epidemiol*. 2017; 46(4): 1093–1093i. DOI: 10.1093/ije/dyx015.

23. Gulden C, Landerer I, Nassirian A, et al. Extraction and prevalence of structured data elements in free-text clinical trial eligibility criteria. *Stud Health Technol Inf* 2019; 258: 226–230.
24. Borner M, Schweizer D, Fey T, et al. *A source data verification-based data quality analysis within the Network of a German comprehensive cancer center. Transdisciplinary perspectives on public Health in Europe*. Berlin: Springer Fachmedien Wiesbaden, 2022, pp. 189–200. DOI: 10.1007/978-3-658-33740-7_11.
25. Voigt W, Steinbock R and Scheffer B. CREDOS 3.1 ein Baukasten zur Tumordokumentation für Epidemiologische-, Klinische-, Tumorspezifische- und Zentrumsregister integriert in das Kis Sap/r3 Is-h: Po344 [CREDOS 3.1 a modular system for tumor documentation for epidemiological, clinical, tumor-specifical and center-register integrated into the HIS SAP/R3 IS-H]. *Onkologie* 2010; 33: 52.
26. ForeSee Medical Inc. *Natural language processing in healthcare*. USA: ForeSee Medical, Inc, 2022. Available from: https://www.foreseemed.com/natural-language-processing-in-healthcare_:~:text=Natural_language_processing_(NLP)_is,assistants_and_language_translation_applications.
27. Schulz S, Fix S, Klügl P, et al. Comparative evaluation of automated information extraction from pathology reports in three German cancer registries. *GMS Med Inform Biom Epidemiol* 2021; 7(1): 1860–9171.
28. Obervational Health Data Sciences and Informatics. OMOP common data model. USA: Obervational Health Data Sciences and Informatics, 2022. Available from: https://www.ohdsi.org/data-standardization/the-common-data-model/.