

Common data elements of breast cancer for research databases: A systematic review

Esmat Mirbagheri¹, Maryam Ahmadi², Soraya Salmanian³

¹Department of Health Information Management, School of Health Management and Information Sciences, ²Department of Health Information Management, School of Management and Medical Information Sciences, Iran University of Medical Sciences, ³Assistant Professor, Radiation Oncology, Oncopathology Research Center, Faculty of Medicine, Iran University of Medical Sciences, Tehran, Iran

ABSTRACT

Background: Common Data Elements (CDEs) are data-metadata descriptors used to collect research study data. CDEs facilitate the collection, processing, and sharing of breast cancer data. This study intended to explore the CDEs of breast cancer for research databases and primary care systems. **Methods:** This study was conducted using systematic search and review. This systematic literature review covered PubMed, Scopus, Science Direct, SID, ISC, Web of Science, and Google Scholar search engine. It included studies in English language with accessible full-text from the beginning of 2007 to September 2019. **Results:** Reviewing 25 studies revealed that 52 percent of studies were carried out in the US and most studies were conducted between 2013 and 2015. The most domains for using CDEs were: Pathology Report and Registry. The CDEs of breast cancer for research databases were categorized into three categories namely clinical, research, and non-clinical and indicate the importance of these data elements. Most of the studies focused on creating and deploying clinical CDEs as physical examination, clinical history and pathology data. **Conclusion:** The integration of biomedical and clinical data relevant to breast cancer enhances the power of research variable analysis and statistical analysis, thereby facilitating improved knowledge of effective therapeutic interventions. Also CDEs used to collect, store, and retrieve patient data in various health setting such as primary care and research databases.

Keywords: Breast cancer, common data elements (CDEs), research database

Introduction

Breast cancer, along with lung and colorectal cancer, has three types of cancer in terms of high incidence and mortality rate worldwide. Together, these three cancers account for one-third of all cancer and deaths in the world. Breast cancer as the fifth leading cause of death (627,000 deaths, 6.6%) is a relatively favorable prognosis and is at least prevalent in developed countries. Breast cancer is the most common cancer in women (24.2%, about one in every four new cases of cancer diagnosed in women worldwide, is breast cancer). Breast cancer

is also the leading cause of cancer deaths in women (15.0%).^[1] Breast cancer is the most common cancer among Iranian women, with an estimated number of cases (5 years) in 2018 for Iran, amounted to 40,825 or 32.3%.^[2]

Improving the efficacy of clinical trials in research for breast cancer will lead to the innovation and reduction of time to use new methods and drugs in the treatment of this disease. To increase efficiency, different departments involved in a research study, including sponsors, clinical researchers, and surveillance devices, and each used different systems and software to collect and analyze data that integrate these programs and systems. Integration is one of the most important factors for achieving desirable goals in the field of medical research. But, at present, the relationship between the two areas of clinical research and

Address for correspondence: Dr. Maryam Ahmadi, No. 6, Rashid Yasemi St., Vali-e-asr Ave., 1995614111 Tehran, Iran. E-mail: ahmadi.m@iums.ac.ir

Received: 22-10-2019

Revised: 11-02-2020

Accepted: 13-02-2020

Published: 26-03-2020

Access this article online

Quick Response Code:



Website:
www.jfmpc.com

DOI:
10.4103/jfmpc.jfmpc_931_19

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: reprints@medknow.com

How to cite this article: Mirbagheri E, Ahmadi M, Salmanian S. Common data elements of breast cancer for research databases: A systematic review. J Family Med Prim Care 2020;9:1296-301.

clinical care is incomplete or sometimes completely disconnected because they each use different standards and terminology systems.^[3] Also, the integration of heterogeneous datasets into clinical research is one of the complex problems that require continuous efforts to optimally utilize data and information in biomedical research.^[4]

The deficiencies and inefficiencies in follow-up care for breast cancer survivors in primary health care indicated the value of healthcare records and datasets in healthcare systems.^[5] Improving health care for patients with breast cancer need to coordinate data from health setting and research database for qualitative primary care.^[6]

Integration of clinical data into electronic health records and clinical trials will increase the likelihood of intervention for disease prevention and treatment.^[7] However, data integrity in the studies are always complex and difficult.^[8] Conventional data collection methods are slow and costly.^[9] The use of CDEs is one of the data integration methods that has increased ability to analyze stored data and combine different findings from studies, thereby reducing the cost of clinical research.^[10] It also requires the definition of a specific set of features to identify CDEs. Among other elements of the data are data synchronization and integration in clinical settings can facilitate synchronization of data and spatial data in a specific field.^[11]

At present, data sharing in the clinical setting and full semantic interoperability between heterogeneous systems have not yet been realized. However, significant progress has been made in this area. In the International Classification of Diseases, a set of standards for controlled clinical terms has been widely used.^[12] Integrating or combining data from different sources and providing it to users with the same vision, Researchers help to coordinate data elements in a particular area or a particular subject.^[13] Data heterogeneity is a major source of challenge in integrating data and inability to interact with health information systems to deliver accurate and effective health care. Knowledge generation is based on clinical data in the context of clinical research.^[14,15] Therefore, solving the semantic heterogeneity problem is the key to achieving interoperability between health care systems and integrating different datasets related to different domains.^[16] Although common data elements were introduced in 2015 by Mesh, since the early 20th century and before, common data elements were used to exchange the same data in different environments of computer systems. The purpose of this study was to retrieve common data elements used in breast cancer databases in order to integrate data elements into heterogeneous clinical research systems.

Methods

Search strategy

Databases including PubMed, Scopus, Science Direct, SID, ISC, Web of Science, and Google Scholar search engine were searched from 2007 to 2019. The Mesh term “CDEs” and all its entry terms^[17] were

searched with “OR” operator. In addition, we searched different terms for breast cancer such as breast neoplasm, breast tumor, mammary cancer, cancer of breast, malignant neoplasm of breast, and breast malignant tumor using “OR” operator. Additionally, the term “research databases” was also searched. All of the three search strategies were conducted in titles, abstracts, and keywords. Finally, the results of the above searches were combined together using “AND” operator. The research team also checked the references of the retrieved articles to find any related articles missed through the searching process. We included English papers dealing with CDEs of breast cancer from 2007 to 2019 (the last 12 years). Papers for which full-text was not accessible and not available in English were excluded. After importing the selected articles into the EndNote, the duplicate items were excluded.

Study selection

The produced list was then independently checked by two raters of the research team in terms of title, abstract, and the content given the inclusion and exclusion criteria. In this way, 25 articles were finally included in the study. The cases of inter-rater disagreement were resolved by holding a mutual meeting by two research team. Furthermore, the inter-rater agreement estimated by using kappa coefficient (κ) was found to be 0.85 (statistically significant at $P < 0.001$). Study selection steps were as per PRISMA flow diagram.

Results

A total of 396 relevant studies were retrieved by the database search. After removing 184 duplicates, 212 studies remained. We excluded 127 studies based on title, another 52 studies based on abstract or full text, and 8 studies because the full-text article was not available. The remaining 25 studies were included for this review. Figure 1 depicts the details of the selection of the studies based on PRISMA flow diagram.

Trend:

Figure 2 shows the trend of studies from 2007 to 2019. Most of studies were carried out in 2013 and 2015. Ninety-five percent of the studies were published in scientific journals and only one study was published in the Australian report that used CDEs to design the national registration system.

Country

Most the articles were published in US. The extracted studies showed that 48 percent were published in US and 32 percent in Europe. Furthermore, England had the highest studies with 20 percent in Europe [Table 1]. Asia and Australia each accounted for 8% of studies, indicating that these countries were less in proportion to this domain.

Domain

The most use of CDEs was in the domain of pathology reporting and registration system with 16%. The domain of integration

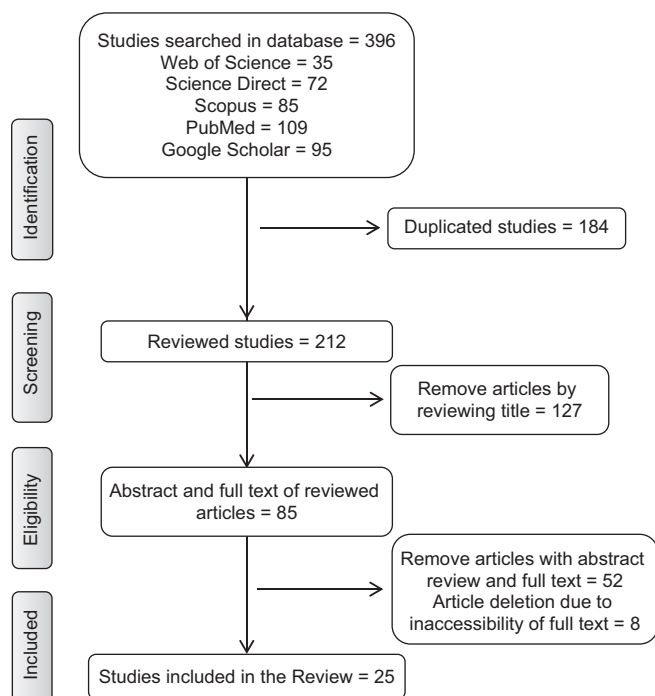


Figure 1: Details of selection of the studies based on PRISMA flow diagram

and diagnosis and screening were the next priorities of the studies. (8%) In general, the use of CDEs of breast cancer in research centers in different domains indicates the ease of creation and use of these data elements [Table 2].

CDEs

CDEs of breast cancer for research databases were categorized into three categories. These CDEs were clinical, research, and non-clinical indicate the importance of these data elements and frequency of use [Table 3].

Discussion

The purpose of this study was to provide an overview of breast cancer data elements in research databases. The results are generally divided into four categories: 1) the study trend over the time, 2) the study site, 3) the domain of studies, and 4) the CDEs of breast cancer derived from these articles.

Studies were conducted before 2007, but due to limitations in the study, most studies were conducted in the period 2007–2019 between 2013 and 2015. Most studies were conducted in the United Kingdom with 8% in 2013 and the United States and the Netherlands with 4% each. Most studies were conducted in 2015 in the US with 8% and Iran and Thailand each with 4%.

Most studies on the creation and use of CDEs were from the US and European countries. 52 percent of studies were in the US and 32 percent in the EU. In the European Union, the UK accounts for 20 percent of the most common data on breast cancer data with the other European countries, the Netherlands and

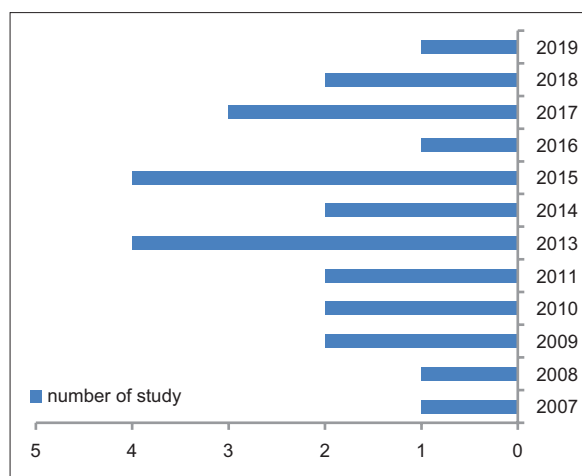


Figure 2: Trend of studies from 2007 to 2019

Table 1: The frequency and percentage of published papers by country

Country	Number	Percentage
USA	13	52%
England	5	20%
The Netherlands	2	8%
Iran	1	4%
Germany	1	4%
New Zealand	1	4%
Thailand	1	4%
Australia	1	4%
Total	25	100%

Table 2: The frequency of domain CDEs in breast cancer

Domain of Studies	Number	Percentage
Pathology Report	4	16%
Registry	4	16%
Integration	2	8%
Screening and Diagnosis	2	8%
Mammography	1	4%
Treatment	1	4%
Data Harmonization	1	4%
Biomarkers	1	4%
Immunology	1	4%
Big Data	1	4%
Interoperability	1	4%
Imaging	1	4%
Documentation and Medical Forms	1	4%
Minimum Data Set	1	4%
Clinical Trial	1	4%
Mesothelioma Virtual Tissue	1	4%
Virtual Biorepository	1	4%
Total	25	100%

Germany following with 8% and 4%. Countries such as France, Italy, and Norway lacked studies on CDEs for breast cancer.

In the domain of study, most of the studies were in the field of pathology reporting and registration system with 16% each.

Table 3: CDEs of breast cancer

Category (Clinical CDEs)	Subcategory
Personal history ^[18-22]	Life style, physical activity and diet habits, Quality of Life, Sleep Habits, Comorbidity Diseases, Clinical History, Menopausal Status
Physical Examination & Clinical History ^[11,18,19,23-34]	Vital Sing, Size of Tumor, Side of Tumor (left, right or bilateral), Main Illness, Status of Body Systems, Morbidity, Referral Data, Body Mass Index, Main Complaint
Family History ^[22,33]	Morbidity
Diagnosis ^[18-23,25,30,33,35,36]	Mammography, sonography, MRI
Cancer Data ^[31,32,37-39]	Type of Cancer (In Situ, Ductal Carcinoma in Situ & Metastasis), Metastasis Status, Date of First Diagnosis, Date of First Metastasis
Surgical Data ^[18,20,26,27,33,34,38]	Date of Surgery, Type of Surgery, Important Findings, Intraoperative Data
Pathology Data ^[11,18,20,21,23-27,29,30,32-34,37,39]	Surgical Pathology, Grade of Tumor (I, II, III, IV), Margin of Mass, Macroscopic & Microscopic Data, Morphology Data, block level annotation, DCIS
Core Needle Biopsy ^[19,33,40]	Size and type of core needle biopsies
Specific Biomarkers & Hormone therapy ^[11,21,28,30-34,37,38,41]	ER, PR, HER2
Epidemiologic Data ^[25,29]	Reproductive data, hereditary status
Laboratory Test ^[19,23]	CA 15.3, CEA (carcinoembryonic antigen), CA125
Genetic & Genomic Data ^[21,28,33,35,38,41]	Gene Sequence Number, Cancer Phenotype
Lymph Node Status ^[18,20,24,27,30-32]	Lymph Node Involvement, Examination of Lymph Nodes, Metastasis to Lymph node, Number of Lymph Nodes Tested
Histology Data ^[21,24,31-33,39]	Grade of Histology, Immunohistology
Type of Treatment & Results ^[11,19,21-23,26,30,31,33,35,36,38,39]	Pharmaceutical, Surgical, Chemotherapy, Radiotherapy, Palliative care
Fallow Up Patient ^[18,23,25,29,30,33,35,38,39]	Fallow up of Treatment, Recurrence, Managing and Controlling Patient Pain
Radiology, Mammography, Ultra Sonography, MRI Data ^[19,35,36,42]	Location of Mass (left, right or bilateral), BIRAD Imaging & Result, Calcification, Cystic Breast
Category (Non Clinical CDEs)	Subcategory
Demographic ^[11,21-23,25-29,34,36,37,39]	Name, Family Name, Patient ID, Birth Date, Marital Status, Address of Home and Work
Identification Information ^[23,34,36]	Social Security Number, Patient Identification Number, Medical Record Number
Contact Information ^[11,33,36]	Information of Contact with other Treatment providers, Address
Financial information/refund ^[11,19,36]	Government or institutions
Managerial & Legal information ^[21,22,27,31]	Patient Consent, Date of Death (If Happen), Cause of death
Category (Research CDEs)	Subcategory
Type of Study ^[11,18,22,28,30]	Cohort, Clinical Trial, ID of study
Location of Study ^[11,18,22]	Hospital, Palliative Center, Day Clinic
Date of Start (study) ^[11,18,22,28]	Date
Date of Last Contact Patient ^[18,33]	Date

Integration with 8% and screening and diagnosis with 8% were also studied. Other areas of importance for the creation and use of CDEs include: mammography, cancer treatment, data coordination, biomarker immunology, Big Data, interoperability, imaging, documentation and medical forms, minimal data sets, clinical trials, mesothelioma, virtual tissue and virtual biorepository each made up 4%.

In the field of CDEs used for the Cancer Research Database, CDEs were divided into three categories, clinical, research, and nonclinical, by the research team. Common clinical data elements had the most duplication in the study literature, which were subdivided into categories.

The most CDEs were in the clinical category, which is referred to in most articles. CDEs, physical examination and clinical history,^[11,18,19,23-34] diagnosis,^[18-23,25,30,33,35,36] pathologic data,^[11,18,20,21,23-27,29,30,32-34,37,39] type and outcome of treatment,^[11,19,21-23,26,30,31,33,35,36,38,39] and specific biomarkers of breast cancer and hormone therapy^[11,21,28,30-34,37,38,41] are of

greater importance and have been suggested to be used in most studies. Other CDEs include patient follow up, surgical data,^[18,23,25,29,30,33,35,38,39] lymph nodes status,^[18,20,24,27,30-32] genetic and genomic data, histological data,^[21,28,33,35,38,41] personal history,^[18-22] cancer data,^[31,32,37-39] radiology, mammography, ultrasonography and MRI data,^[19,35,36,42] Core Needles Biopsy,^[19,33,40] epidemiologic data,^[25,29] and laboratory test.^[19,23]

The review revealed that CDEs of physical examination and clinical history, diagnosis and pathologic data are very important in collecting and organizing CDEs in research databases and need to be used in the design and creation of registers and databases. In general, the most important of clinical CDEs of breast cancer were shared pathology and physical examination data items (64%).

In nonclinical category, demography data^[11,21-23,25-29,34,36,37,39] is important for designing CDEs and other subcategories included in identification information,^[23,34,36] managerial and legal information,^[21,22,27,31] and contact information and financial

information.^[21,22,27,31] For the CDEs of the nonclinical, the demographic with 52% is most important.

In the category of research CDEs, the type of study^[11,18,22,28,30] was more important, date of start (study),^[11,18,22,28] location of studies,^[11,18,22] and the date of last visit or contact^[18,33] with the patient were identified in the relevant articles. In research CDEs, type of study with 20 percent is important.

Storing and retrieving data through original data definitions using CDEs is a way of integrating clinical data from different databases.^[21,28-30] The use of CDEs to underlie clinical research such as tissue banks demonstrates the efficacy and standardization of data model that can be applied in other domains of biomarkers and bioinformatics associated with breast cancer.^[22,28,29,39] CDEs are also suitable for short-term studies on large datasets, so that these data elements act as a “mediator” of a unified model for mapping biomedical ontologies.^[26,39] Furthermore, CDEs facilitate the creation and use of EHR data, as an interface for connecting local data to EHR integrated or national registry.^[27,34] This study is like the Sluijter study, the importance of pathological data elements has been emphasized pathological data elements for description of ‘resection margins’, ‘DCIS size’, ‘location’ and ‘presence of calcifications’.^[32] There were limitations to the present study, such as, missing some studies with other language or studies that full text not available.

Conclusion

Medical research into various diseases, including cancer, requires the collection, processing, and exchange of data with different centers. Data exchange leads to data efficiency, preventing rework, saving time and cost, and ultimately enhancing the quality of medical research. One of the standard methods in this field is the use of data standards including data collection in the form of CDEs. Integrated datasets lead to integrated terminology to facilitate data management across the mass of patient data collected. Accordingly, CDEs can be the basis for achieving higher standard levels and data quality and facilitating the application of health information technology in breast cancer research centers. However, the identification and using of CDEs needs to be coordinated by different healthcare systems to use standard data for breast cancer patients to improve primary care.

Financial support and sponsorship

This study is a part of a PhD dissertation granted by Iran University of Medical Sciences (Grant No: IUMS/SHMIS_2017/9321563004).

Conflicts of interest

There are no conflicts of interest.

References

1. Latest Global Cancer Data: Cancer Burden Rises to 18.1 Million New Cases and 9.6 Million Cancer Deaths in 2018. Geneva, Switzerland: International Agency for Research on Cancer (IARC); 2018.
2. Zendehelel K. Cancer statistics in I. R. Iran in 2018. *Basic Clin Cancer Res* 2019;11:1-4.
3. Laleci GB, Yuksel M, Dogac A. Providing semantic interoperability between clinical care and clinical research domains. *IEEE J Biomed Health Inform* 2013;17:356-69.
4. Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, *et al.* From big data to precision medicine. *Front Med* 2019;6:34.
5. McBride ML, Groome PA, Decker K, Kendell C, Jiang L, Whitehead M, *et al.* Adherence to quality breast cancer survivorship care in four canadian provinces: A canimpact retrospective cohort study. *BMC Cancer* 2019;19:659.
6. Jansana A, Posso M, Guerrero I, Prados-Torres A, Del Cura MI, Castells X, *et al.* Health care services use among long-term breast cancer survivors: A systematic review. *J Cancer Survivorship* 2019;13:477-93.
7. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, *et al.* Electronic health records to facilitate clinical research. *Clin Res Cardiol* 2017;106:1-9.
8. Boyd D. Six provocations for big data [internet]. Rochester: Social Science Research Network, 2011. [Cited 2014 Sep 14]. Report No: 1926431.
9. Kaplan RM, Riley WT, Mabry PL. News from the nih: Leveraging big data in the behavioral sciences. *Transl Behav Med* 2014;4:229-31.
10. Riley WT, Glasgow RE, Etheredge L, Abernethy AP. Rapid, responsive, relevant (r3) research: A call for a rapid learning health research enterprise. *Clin Transl Med* 2013;2:10.
11. Luo Z, Miotto R, Weng C. A human-computer collaborative approach to identifying common data elements in clinical trial eligibility criteria. *J Biomed Inform* 2013;46:33-9.
12. Coustasse A, Paul DP IIIrd. Adoption of the icd-10 standard in the united states: The time is now. *Health Care Manag (Frederick)* 2013;32:260-7.
13. Ichihara K. Statistical considerations for harmonization of the global multicenter study on reference values. *Clin Chim Acta* 2014;432:108-18.
14. Ohno-Machado L. Nih’s big data to knowledge initiative and the advancement of biomedical informatics. *J Am Med Inform Assoc* 2014;21:193.
15. Schuurman N, Leszczynski A. A method to map heterogeneity between near but non-equivalent semantic attributes in multiple health data registries. *Health Informatics J* 2008;14:39-57.
16. Weng C, Gennari JH, Fridsma DB. User-centered semantic harmonization: A case study. *J Biomed Inform* 2007;40:353-64.
17. NCBI. Available from: <https://www.ncbi.nlm.nih.gov/mesh/?term=common+data+elements>. [Cited 2019 Sep 27].
18. Breast Cancer Specific Data Items for Clinical Cancer Registration. In: Centre NBaOC, editor. Surry Hills, NSW: National Breast and Ovarian Cancer Centre; 2009.
19. Hu H, Correll M, Kvecher L, Osmond M, Clark J, Bekhash A, *et al.* Dw4tr: A data warehouse for translational research. *J Biomed Inform* 2011;44:1004-19.
20. Kussaibi H, Macary F, Kennedy M, Booker D, Brodsky V, Schrader T, *et al.* H17 cda implementation guide for structured anatomic pathology reports methodology and tools. *Stud Health Technol Inform* 2010;160:289-93.

21. Patel AA, Gilbertson JR, Showe LC, London JW, Ross E, Ochs MF, *et al.* A novel cross-disciplinary multi-institute approach to translational cancer research: Lessons learned from pennsylvania cancer alliance bioinformatics consortium (pcabc). *Cancer Inform* 2007;3:255-74.
22. Sherman S, Shats O, Fleissner E, Bascom G, Yiee K, Copur M, *et al.* Multicenter breast cancer collaborative registry. *Cancer Inform* 2011;10:217-26.
23. Ghaneie M, Rezaie A, Ghorbani NR, Heidari R, Arjomandi M, Zare M. Designing a minimum data set for breast cancer: A starting point for breast cancer registration in iran. *Iran J Public Health* 2013;42(Supple1):66-73.
24. Hassell LA, Parwani AV, Weiss L, Jones MA, Ye J. Challenges and opportunities in the adoption of college of american pathologists checklists in electronic format: Perspectives and experience of reporting pathology protocols project (rpp2) participant laboratories. *Arch Pathol Lab Med* 2010;134:1152-9.
25. Jazayeri SB, Saadat S, Ramezani R, Kaviani A. Incidence of primary breast cancer in iran: Ten-year national cancer registry data report. *Cancer Epidemiol* 2015;39:519-27.
26. Keshtkaran A, Sharifian R, Barzegari S, Talei A, Liu S, Tahmasebi H. Agreement of iranian breast cancer data and relationships with measuring quality of care in a 5-year period (2006-2011). *Asian Pac J Cancer Prev* 2013;14:2107-11.
27. Kilburn LS, Aresu M, Banerji J, Barrett-Lee P, Ellis P, Bliss JM. Can routine data be used to support cancer clinical trials? A historical baseline on which to build: Retrospective linkage of data from the tact (cruc 01/001) breast cancer trial and the national cancer data repository. *Trials* 2017;18:561.
28. Krumm R, Semjonow A, Tio J, Duhme H, Burkle T, Haier J, *et al.* The need for harmonized structured documentation and chances of secondary use-results of a systematic analysis with automated form comparison for prostate and breast cancer. *J Biomed Inform* 2014;51:86-99.
29. Mohanty SK, Mistry AT, Amin W, Parwani AV, Pople AK, Schmandt L, *et al.* The development and deployment of common data elements for tissue banks for translational research in cancer - An emerging standard based approach for the mesothelioma virtual tissue bank. *BMC Cancer* 2008;8:91.
30. Papatheodorou I, Crichton C, Morris L, Maccallum P, Davies J, Brenton JD, *et al.* A metadata approach for clinical data management in translational genomics studies in breast cancer. *BMC Med Genomics* 2009;2:66.
31. Roelands J, Decock J, Boughorbel S, Rinchai D, Maccalli C, Ceccarelli M, *et al.* A collection of annotated and harmonized human breast cancer transcriptome datasets, including immunologic classification. *F1000Research* 2017;6:296.
32. Sluijter CE, van Lonkhuijzen LRCW, van Slooten HJ, Nagtegaal ID, Overbeek LIH. The effects of implementing synoptic pathology reporting in cancer diagnosis: A systematic review. *Virchows Arch* 2016;468:639-49.
33. Warner JL, Maddux SE, Hughes KS, Krauss JC, Yu PP, Shulman LN, *et al.* Development, implementation, and initial evaluation of a foundational open interoperability standard for oncology treatment planning and summarization. *J Am Med Inform Assoc* 2015;22:577-86.
34. Zuley ML, Nishikawa RM, Lee CS, Burnside E, Rosenberg R, Sickles EA, *et al.* Linkage of the acr national mammography database to the network of state cancer registries: Proof of concept evaluation by the acr national mammography database committee. *J Am Coll Radiol* 2019;16:8-14.
35. Lee JS, Kibbe WA, Grossman RL. Data harmonization for a molecularly driven health system. *Cell* 2018;174:1045-8.
36. Yancy B, Royalty JE, Marroulis S, Mattingly C, Benard VB, DeGross A. Using data to effectively manage a national screening program. *Cancer* 2014;120(Suppl 16):2575-83.
37. Kern DM, Barron JJ, Wu B, Ganetsky A, Willey VJ, Quimbo RA, *et al.* A validation of clinical data captured from a novel cancer care quality program directly integrated with administrative claims data. *Pragmat Obs Res* 2017;8:149-55.
38. Nelson HD, Weerasinghe R. Actualizing personalized healthcare for women through connected data systems: Breast cancer screening and diagnosis. *Glob Adv Health Med* 2013;2:30-6.
39. Noor AM, Holmberg L, Gillett C, Grigoriadis A. Big data: The challenge for small research groups in the era of cancer genomics. *Br J Cancer* 2015;113:1405-12.
40. Alonso-Calvo R, Perez-Rey D, Paraiso-Medina S, Claeherhout B, Hennebert P, Bucur A. Enabling semantic interoperability in multi-centric clinical trials on breast cancer. *Computer Methods Programs Biomed* 2015;118:322-9.
41. Breitenstein MK, Liu H, Maxwell KN, Pathak J, Zhang R. Electronic health record phenotypes for precision medicine: Perspectives and caveats from treatment of breast cancer at a single institution. *Clin Transl Sci* 2018;11:85-92.
42. Iacson R, Harris K, Brawarsky P, Tosteson TD, Onega T, Tosteson AN, *et al.* Evaluation of an automated information extraction tool for imaging data elements to populate a breast cancer screening registry. *J Digit Imaging* 2015;28:567-75.