German Medical Data Sciences: Bringing Data to Life
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI210063

Merging Genomics Public Datasets with Clinical Cancer Registry Data – Lessons Learned

Lorena SCHALL[a,b], Sylvia BOCHUM[b,c] and Monika POBIRUCHIN[1,d]
[a]*Dept. of Medical Informatics, Faculty of Informatics, Heilbronn University*
[b]*MOLIT Institute for Personalized Medicine gGmbH, Heilbronn*
[c]*SLK-Hospitals, Cancer Center Heilbronn-Franken*
[d]*GECKO Institute for Medicine, Informatics and Economics, Heilbronn University*

Abstract. Publicly available datasets – for example via cBioPortal for Cancer Genomics – could be a valuable source for benchmarks and comparisons with local patient records. However, such an approach is only valid if patient cohorts are comparable to each other and if the documentation is complete and sufficient. In this paper, records from exocrine pancreatic cancer patients documented in a local cancer registry are compared with two public datasets to calculate overall survival. Several data preprocessing steps were necessary to ensure comparability of the different datasets and a common database schema was created. Our assumption that the public datasets could be used to augment the data of the local cancer registry could not be validated, since the analysis on overall survival showed a significant difference. We discuss several reasons and explanations for this finding. So far, comparing different datasets with each other and drawing medical conclusions on such comparisons should be conducted with great caution.

Keywords. Databases, Factual; Genomics; Pancreatic Cancer

1. Introduction

Worldwide, exocrine pancreatic cancer (ICD-10: C25) is the ninth most common cancer in women and tenth most common cancer in men. In 2018, it accounted for 45,750 deaths [1]. Mortality in exocrine pancreatic cancer is high: 5-Year survival is estimated with 8% for women and men in Germany [2]. One reason for low survival is late diagnosis of the tumor, i.e., the tumor has already spread to a secondary site in the body (Stage IV according to Union for International Cancer Control (UICC)). Therefore, research is focusing on therapies for late-stage pancreatic cancer. Published studies showed that patients with germline mutations in BRCA genes can benefit from targeted therapies that increase progression free survival [1].

However, it is difficult for regional institutions to reproduce such analyses with their own registry data: Exocrine pancreatic cancer has a relatively low incidence and among these cases only a small subgroup shows a BRCA1 or BRCA2 mutation which further limits the sample size [3,4]. One solution for this problem might be the use of publicly available datasets, e.g., via cBioPortal for Cancer Genomics [5], to increase sample size

[1] Corresponding Author. E-Mail: monika.pobiruchin@hs-heilbronn.de

and 'augment' local datasets. However, such an approach is only valid if patient cohorts are comparable to each other and if the documentation of cancer cases is complete and sufficient. In this paper, records from Stage IV exocrine pancreatic cancer patients documented in a local cancer registry are compared with two public datasets (Memorial Sloan Kettering (MSK), The Cancer Genome Atlas (TCGA)) from cBioportal to calculate overall survival (OS). We assumed that OS for late-stage pancreatic cancer should be comparable to each other as the treatment of these patients follows internationally well-known guidelines and medical evidence. However, we encountered some pitfalls when working with such datasets which are presented and discussed in this paper.

2. Methods & Material

2.1. cBioPortal

The cBioPortal for Cancer Genomics provides data on certain types of cancer and the associated genetic data. Data can be selected from different studies and visualized directly on the web interface, or can be downloaded as `tar.gz` files. For this paper, two datasets from cBioPortal where closely considered: MSK and TCGA.

Between January 2014 and May 2016, the Memorial Sloan Kettering Cancer Center collected more than 12,000 tumors from 11,369 patients for prospective sequencing [6]. The full dataset was made publicly available via cBioPortal [7]. The TCGA data comprises molecular analyses of tumors of approximately 10,000 specimens and representing 33 types of cancer (the PanCancer Atlas) [8].

2.2. OncoKB

OncoKB - Precision Oncology Knowledge Base [9] contains information on specific gene alterations. The information is curated from various sources and institutions, e.g., FDA (U.S. Food and Drug Administration), ASCO (American Society of Clinical Oncology) or ClinicalTrials.gov.
In total, more than 5,000 alterations from 671 tumor specific genes are documented. As of June 2019, information about pathology of specific gene mutations was missing in cBioPortal's records – or at least not accessible for Website users. Therefore, we used OncoKB to decide which mutations are oncogenic or likely oncogenic.

2.3. Local Cancer Registry

The SLK Hospital Holding GmbH is located in Heilbronn (southern Germany) and the surrounding districts. Approximately 51,000 inpatients and 136,000 outpatients are treated annually. Between 2010 and 2017, 608 patients were diagnosed with exocrine pancreatic carcinoma at SLK, i.e., 76 pancreatic cancer patients per year. All cancer cases are recorded and stored in the local GTDS (Giessener Tumor Documentation System) database. GTDS is a common tool for information management in cancer registries in Germany [10]. It supports documentation of diagnosis, therapy and follow-up care and is able to provide anonymized data exports to ensure data privacy.

2.4. Data Preprocessing

Several data preprocessing steps were necessary to ensure comparability of the different datasets: (i) column contents with missing information ("N.A.") were removed, (ii) (re-)calculation of survival time in months, (iii) selection of relevant exocrine pancreatic cancer cases by oncotree code: "PAAD" (pancreatic adenocarcinoma), (iv) removal of cases without stage classification, (v) matching of histology notations (GTDS data) to oncotree code PAAD. For GTDS data, to differentiate between clinical and pathologic TNM stage, the column `p_y_symbol`, giving notice of neoadjuvant treatment, was used to determine if the clinical stage information (c-stadium) or the pathologic stage information (p-stadium) should be included for further analysis steps.

At the end of the selection and preprocessing a total of 979 cases (MSK=384, TCGA=173, GTDS=422) remained for further analyses, see Figure 1.

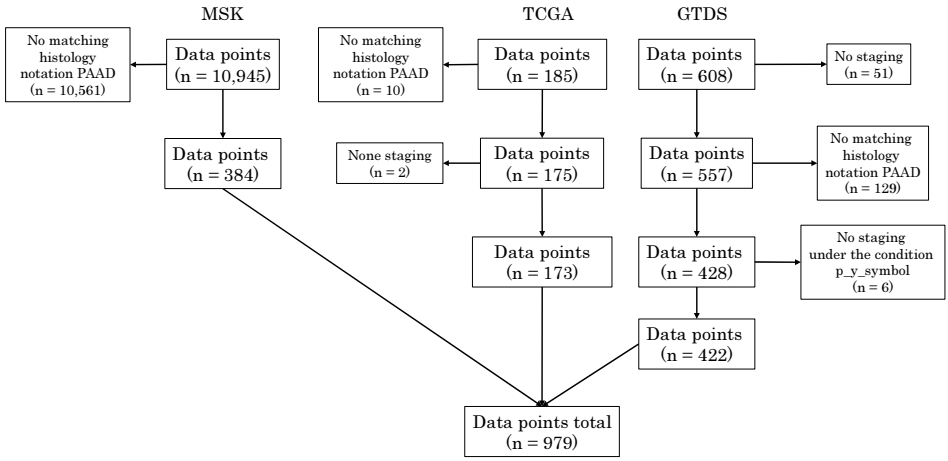

Figure 1 Flow chart of the preprocessing steps and removed tuples from the datasets.

Next, a common database schema was created for the five (MSK, TCGA, GTDS, OncoKB, Genes of MSK) datasets. This way, easy querying as well as mapping and merging of cases originating from different records was possible. PostgreSQL (version 10.3) was used as database system.

2.5. Statistical Analysis

The subsequent data analysis was performed using the programming language `R` (version 3.5.3); for drawing curves, `ggsurvplot` from `ggplot2` package (version 3.2.1) was required [11]. The survival of the cohorts was compared with Kaplan-Meier curves.

3. Results

3.1. Patient Characteristics

Sex and age distribution for pancreatic cancer patients in the three cohorts are similar to each other (see Table 1). However, the distribution of tumor stages differs: MSK cohort only contains Stage IV patients, TCGA mainly non-metastatic cancer (97.11%), whereas more than half of the GTDS patients are in Stage IV.

Table 1. Overview of the three patient cohorts. N.A. = information not available.

Dataset	MSK (n=384)	TCGA (n=173)	GTDS (n=422)
Female	179 (46.61%)	77 (44.51%)	189 (44.79%)
Mean Age	N.A.	65	71
Stage			
IA	0 (0.00%)	2 (2.60%)	3 (1.59%)
IB	0 (0.00%)	5 (6.49%)	0 (0.00%)
IIA	0 (0.00%)	10 (12.99%)	14 (7.41%)
IIB	0 (0.00%)	54 (70.13%)	45 (23.81%)
III	0 (0.00%)	3 (3.90%)	32 (16.93%)
IV	179 (100.00%)	3 (3.90%)	95 (50.26%)
Male	205 (53.39%)	96 (55.49%)	233 (55.21%)
Mean Age	N.A.	65	69
Stage			
IA	0 (0.00%)	5 (2.89%)	0 (0.00%)
IB	0 (0.00%)	10 (5.78%)	2 (0.86%)
IIA	0 (0.00%)	30 (17.34%)	11 (4.72%)
IIB	0 (0.00%)	119 (68.79%)	52 (22.32%)
III	0 (0.00%)	4 (2.31%)	36 (15.45%)
IV	205 (100.00%)	5 (2.89%)	132 (56.65%)

3.2. Survival Analysis

A survival analysis for Stage IV cases was conducted. As the TCGA dataset only comprised n=8 Stage IV cases, MSK and TCGA cases were combined. Comparing the MSK/TCGA and the GTDS cohort revealed a significant difference in overall survival (see Figure 2). The median survival for GTDS was 10.1 months, for MSK/TCGA 17.9 months.

4. Discussion

Publicly available datasets are a huge step forward for collaborative research. However, we encountered pitfalls and learned some major lessons that other researchers should keep in mind when working with similar data.

Our assumption that the MSK/TCGA datasets could be used to ‘augment’ the data of the (anonymized) local cancer registry could not be validated, since the analysis on OS showed a significant difference. In the local cancer registry, a median OS of approximately 8 months in Stage IV exocrine pancreatic cancer cases and 11 months in

patients treated with at least one line of chemotherapy cancer is reported, which is in accordance with data published in several national and international guidelines [12,13].

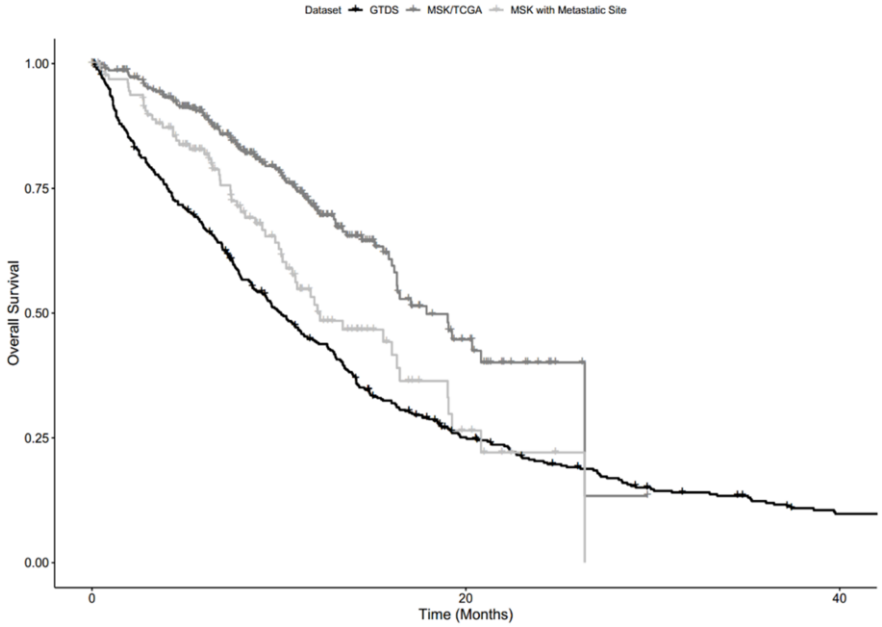

Figure 2. Kaplan-Maier plot of overall survival of Stage IV exocrine pancreatic cancer patients comparing the cohorts a) GTDS, b) MSK/TCGA all sites including 'NA', c) MSK with specified site of metastasis

The MSK/TCGA dataset displays an unprecedented superior median OS of approximately 20 months in Stage IV. Based on the documented data in the MSK/TCGA data we cannot explain such a gap. However, if we calculate survival only for those cases where the site of metastasis is specified (see Figure 2), then median OS is 10 months, giving strong evidence that documentation of the clinical stage in the MSK/TCGA datasets might be not faultless with probably several patients with initially localized pancreatic cancer (Stage I - III) wrongly assigned to Stage IV. Nevertheless, it is also known that germline mutations BRCA genes and different therapy strategies can have an influence on survival in pancreatic cancer. However, treatment information and information, if reported genetic mutations are present in germline, is usually missing in the available public datasets.

Mapping of different datasets to establish a common data source for combined analyses is no trivial task and requires in-depth knowledge of the oncology domain and tumor documentation. As already mentioned in many other articles: Interdisciplinary teams are needed. If a common mapping cannot be established, it could be difficult and/or impossible to assess if the documented values really mean the same. The concept of semantic interoperability should be paid attention to – especially for scientific use files. As of today, GTDS and cBioPortal do not support clinical standards, e.g., HL7 FHIR.

The quality of the public data for an in-depth comparison is insufficient: Necessary information such as the age at initial diagnosis of the carcinoma, general condition of the

patients, or forms of therapy (surgery, neo-/adjuvant chemotherapy, immunotherapy, etc) and medication are missing. It is therefore not feasible to make conclusions regarding differences in overall survival. Furthermore, because of missing and not-accessible values in the datasets e.g. pathogenicity of mutations, more than one external database (for example OncoKB) is needed.

For the evaluations with the programming language `R`, no simplified application programming interface (API) for the respective databases were available (cBioPortal and OncoKB). When using the `CGDS-R` package, the numbering and changes of the study order had to be taken into account when generating and using new datasets. Otherwise, incorrect study data were evaluated (as of June 2019).

In the future, publicly available datasets could be potential sources for benchmarks and comparisons for other researchers if these records are systematically documented. So far, comparing different datasets with each other and drawing medical conclusions on such comparisons should be conducted with great caution.

References

[1] T. Golan, P. Hammel, M. Reni, et al. Maintenance Olaparib for Germline BRCA-Mutated Metastatic Pancreatic Cancer. *N Engl J Med* **381** (2019), 317-327. DOI: 10.1056/NEJMoa1903387

[2] Zentrum für Krebsregisterdaten und der Gesellschaft der Epidemiologischen Krebsregister in Deutschland e.V. Krebs in Deutschland 2015/2016. RKI, Berlin 2019. DOI: 10.25646/5977

[3] A. Sehdev, O. Gbolahan, B.A. Hancock et al., Germline and Somatic DNA Damage Repair Gene Mutations and Overall Survival in Metastatic Pancreatic Adenocarcinoma Patients Treated with FOLFIRINOX. *Clin Cancer Res* **24** (2018). DOI: 10.1158/1078-0432.CCR-18-1472

[4] M.J. Pishvaian, R.J. Bender, D. Halverson, et al., Molecular Profiling of Patients with Pancreatic Cancer: Initial Results from the Know Your Tumor Initiative. *Clin Cancer Res* **24** (2018), 5018-5027. DOI: 10.1158/1078-0432.CCR-18-0531

[5] E. Cerami, J. Gao, U. Dogrusoz, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov* **2** (2012), 401–404. DOI: 10.1158/2159-8290.CD-12-0095

[6] A. Zehir, R. Benayed, R.H. Shah, et al. Mutational Landscape of Metastatic Cancer Revealed from Prospective Clinical Sequencing of 10,000 Patients. *Nat Med* **23** (2017), 703–713. DOI: 10.1038/nm.4333

[7] J. Gao, B. Aksoy, U. Dogrusoz, et al., Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal, *Sci Signal* **6** (269). DOI: 10.1126/scisignal.2004088

[8] K.A. Hoadley, C. Yau, T. Hinoue, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173** *(*2018), 291–304. DOI: 10.1016/j.cell.2018.03.022

[9] D. Chakravarty, J. Gao, S.M. Phillips et al., OncoKB: A Precision Oncology Knowledge Base, *JCO Precision Oncology*, 1, 1-16. DOI: 10.1200/PO.17.00011

[10] U. Altmann, J. Dudeck J. The Giessen Tumor Documentation System (GTDS) – Review and Perspectives. *Methods Inf. Med.* **45** (2006), 108–115. DOI: 10.1055/s-0038-1634046

[11] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2016

[12] Leitlinienprogramm Onkologie (Deutsche Krebsgesellschaft, Deutsche Krebshilfe, AWMF). S3-Leitlinie Exokrines Pankreaskarzinom, Langversion 1.0, 2013, AWMF Registernummer: 032-010OL. Berlin; 2013. Available from: http://leitlinienprogramm-onkologie.de/Leitlinien.7.0.html

[13] M. Ducreux, C. Caramella, A. Hollebecque, et al., (2015). Cancer of the pancreas: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology* **26** (2015), v56-v68.