RESEARCH PROTOCOL:
Characterizing Patients with Metastatic Bladder Cancer using the OHDSI data network and the OMOP Oncology CDM module

TABLE OF CONTENT

List of Abbreviations

ATC	anatomical therapeutic chemical
CDM	common data model
CI	confidence interval
DTI	diagnosis to treatment interval
EHR	electronic health records
EMR	electronic medical records
ENCePP	European network of centers for pharmacoepidemiology and pharmacovigilance
GPP	guidelines for good pharmacoepidemiology practices
HemOnc	Hematology/Oncology
IQR	inter-quartile range
ISPE	international society of pharmacoepidemiology
LDH	serum lactate dehydrogenase
LoT	systemic anti-neoplastic line of treatment
NAACCR	North American Association of Central Cancer Registries
OHDSI	observational health data science and informatics collaborative
OMOP	observational medical outcomes partnership
OS	overall survival
RxNorm	
RWE	real-world evidence
SEER	surveillance, epidemiology and end results
SD	standard deviation
SMD	standardized mean difference
SNOMED	systematized nomenclature of medicine
TFI	treatment-free interval
TTNT	time to next treatment
TTD	time to treatment discontinuation
UC	urothelial cancer

Rationale and Background

Urothelial carcinoma (UC), which accounts for >90% of all bladder cancers, is one of the 10 predominant malignancies worldwide [1,2,3]. Metastatic urothelial carcinoma carries a dismal prognosis, with 5-year survival rate of 5%[1] and with limited treatment options. Despite recent improvements in clinical outcomes of patients with advanced bladder cancer following first- or second-line of anti-neoplastic treatment with immune checkpoint inhibitors, data are limited on the treatment patterns and outcomes in patients with metastatic bladder cancer in the routine clinical setting. Also, little is known about epidemiological characteristics of patients with bladder cancer who have liver metastasis at the time of diagnosis or develop liver metastasis during the course of disease.

Sufficient data on this patient population are hard to obtain from a single institution, impacting both internal and external validity of research findings. Observational Health Data Science and Informatics collaborative (OHDSI) provides Open Source resources to the community: the (observational medical outcomes partnership) OMOP Common Data Model (CDM), the OMOP Standardized Vocabularies, and a number of tools and methods to interrogate and analyze the data [4,5,6]. This allows to generate real world data (RWE) through the OHDSI Research Network, whereby each participating institution retains their data, but makes them available for querying and statistical evaluation. All data are standardized to a common data model (format) and reference data (vocabularies). That allows for queries and studies to be developed with no access to the data where they eventually are executed.

Typical observational studies base their cohort and outcome definitions on diagnostic codes, sometimes in combination with drug treatment, procedure occurrence or lab tests. Malignant neoplasms tend to have significantly more diagnostic make up: anatomical site including local penetration, morphology, metastatic spread, affected lymph nodes, mechanical consequences of tumor growth and biomarkers. All of them together define the disease and have therapeutic consequences. However, typical observational data sources do not convey the necessary level of detail, and if they do, they collect individual results of diagnostic procedures but do not summarize to the above attributes (abstraction). In addition, cancer treatment tends to also be more complex compared to other disease modalities. Cancers are treated with chemotherapy regimens, often combined with targeted therapies or immunotherapies, surgery or radiotherapy. None of these attributes follow standard definitions that can be easily applied to retrospective observational data.

The OHDSI Initiative formed a Working Group to address these issues. It standardized the topology, morphology, spreading and other tumor markers, therapies and abstractions. OMOP Cancer Module that extends the OMOP CDM and Standardized Vocabularies to support the comprehensive representation of cancer conditions, treatments, and disease abstraction required for addressing key research questions. It has the potential to allow distributed research in a network of databases all describing the patient populations in a standardized way without the need to extract the data, aggregate and interpret them at analysis time. If feasible, the OHDSI network has potentially a large number of patients available globally for systematic and standardized research.

Rationale

The foremost necessities in cancer research are to better understand unmet treatment needs, to identify special patient populations who might benefit from new treatments, and to collect baseline data that can help contextualize the impact of future novel therapies in specific patient population. Likewise, in bladder cancer research, there is immense need of characterizing the current treatment landscape of patients with and without liver metastasis in real-world setting.

However, this is no easy task. Cancer patients are treated in the hospital setting, often in specialized academic cancer centres or ambulatory specialty clinics. These centres do typically not share their data, except for participation in cancer registries. These extract high-quality information about the disease, but generally omit most other aspects of patient care. In addition, cancer-specific data are not submitted for billing purposes, at least not at the detail required for research.

To address these issues, the OHDSI community and in particular the Oncology Working Group undertook the attempt to utilize the OMOP approach. In it, data are not shared, but harmonized to a rigorous standard at the institution, and RWE gets generated through federated queries and analytics.

This study aimed at testing the feasibility of the OHDSI approach for RWE generation on bladder cancer in a network of databases and utilizing this information to inform on key aspects of treatment patterns in patients with metastatic bladder cancer (with and without liver cancer). In particular, it established the feasibility of the OHDSI Research Network, and that of participating cancer centres, to generate detailed epidemiological insights into cancer patients on the exam.

Objective(s)

The aim of this study was to assess the feasibility of OHDSI approach in generating RWE on the example of bladder cancer by identifying bladder cancer patients, those with metastases and those with liver metastases, and describing the characteristics (including age, gender, and comorbidities), the treatment patterns, and clinical outcomes of patients with metastatic bladder cancer and those with metastasis to liver in the real-world.

1. To transform anonymized, pre-existing data from real-world datasets of cancer patients to the OMOP CDM Oncology Module, and then use this data to identify cohorts of two patient populations using a standardized query developed centrally and distributed to all participating sites:
    - Patients with metastatic bladder cancer
    - Patients with bladder cancer with metastasis to liver
2. To describe demographic and clinical characteristics of patients at the time of diagnosis with bladder cancer in each cohort, in particular: age, gender and where available race distribution of patients, in addition to comorbid conditions
3. To characterize detailed treatment patterns among patients with advanced bladder cancer and bladder cancer patient with liver metastasis. Specifically,
    - Distribution of treatment regimens, their dose, cycle, and scheduling per systemic anti-neoplastic line of treatment LoT (up to three LoTs)
    - Treatment flow across 1st, 2nd and 3rd LoT (treatment pathways)
    - Time to treatment discontinuation
    - Time to next treatment
4. To estimate overall survival

Methods

Study Design

This is a retrospective cohort study of patients diagnosed with metastatic bladder cancer and bladder cancer with metastasis to liver.

Patients diagnosed with metastatic bladder cancer and between 1 January 2000 and 1 December 2019 are included in the study. Patients are followed longitudinally until death or end of data capture set at 30 July 2020, unless lost to follow-up (censored).

Setting

This study is conducted in a network of data sources. That means data are not extracted and pooled to conduct the study. Instead, each participating institution develops and maintains its own data asset.

Table 1 data sources description of participating institutions

Institution	Data source description
Northwestern University **Memorial Sloan Kettering Cancer** **Tufts University** **Columbia University**	All data assets are based on Cancer Center electronic health record (EHR) in combination with Surveillance, Epidemiology and End Results (SEER) cancer registry reporting, Records from both were linked at the patient level and a combined OMOP CDM instance was formed. The ETL for the SEER piece was developed centrally and
IQVIA	1) OncoEMR: a data asset derived from a variety of oncology specific electronic medical record (EMR) vendors in the out- patient office setting. Data are geographically representative of the US population and cover all cancer types. 2) OpenClaims: A United States database of open, pre-adjudicated claims from January 2013 to May 2020. Data are reported at anonymized patient level collected from office-based physicians and specialists via office management software and clearinghouse switch sources for the purpose of reimbursement. A subset of medical claims data has adjudicated claims.
Ajou University	EHR based database of Ajou University School of Medicine (AUSOM) standardized into the OMOP-CDM

Study period

The study period starts from the date of database inception and end at latest available date for all data sources.

Study identification period

The identification period is from the date of database inception to six months prior to the latest available data in each database. This allows for a potential of 6 months of follow-up data for the last person included in the study.

Study population: inclusion criteria

Metastatic Bladder Cancer

Adult patients (≥18 years at) diagnosed with metastatic bladder with no history of other malignancies three years prior to index date

Bladder Cancer with Metastasis to Liver

Adult patients (≥18 years at) diagnosed with bladder cancer with liver metastasis with no history of other malignancies three years prior to index date

Participants are identified using pre-specified concept sets developed and reviewed by a core team of epidemiologist, clinicians, vocabulary experts and data scientist with extensive expertise in the use of OMOP CDM and the OHDSI tools.

For this study, two study populations are defined through condition and tumor attribute concepts.

Metastatic Bladder Cancer:

- Have a diagnosis of metastatic bladder cancer, defined as having either condition SNOMED 432851 "Secondary malignant neoplastic disease" or any of its descendants, OR tumor attribute North American Association of Central Cancer Registries (NAACCR) 35918319 "TNM Path M" or 35918383 "TNM Clin M" with one of the value concepts for "cM1", "cM1a", "cM1b", "cM1c", "pM1", "pM1a", "pM1b", "pM1c" OR Tumor attribute NAACCR Variables 35918527 “Mets at DX-Other”, 35918559 “Mets at DX-Lung”, 35918290 “Mets at DX-Liver”, 35918491 “Mets at DX-Distant”, 35918692 “Mets at DX-Brain”, 35918581 “Mets at DX-Bone” with concept values of “Yes; distant metastases in known site(s) other than bone, brain, liver, lung or distant lymph nodes”, “Yes; distant lymph node metastases”, “Yes; distant lung metastases”, “Yes; distant liver metastases”, “Yes; distant brain metastases”, “Yes; distant bone metastases”
- Have a prior diagnosis of bladder cancer; condition of SNOMED 196360 "Primary malignant neoplasm of bladder" or any of its descendants
- No prior diagnosis of other malignancies excluding non-melanoma skin cancers three years before the index date
- ≥18 years at index date

Bladder Cancer with Liver Metastasis:

- Have a diagnosis of bladder cancer with metastasis to liver, defined as having either condition SNOMED 432851 " Secondary malignant neoplasm of liver and intrahepatic bile duct" or any of its descendants, OR tumor attribute NAACCR 35918319 "TNM Path M" or 35918383 "TNM Clin M" with one of the value concepts for "cM1", "cM1a", "cM1b", "cM1c", "pM1", "pM1a", "pM1b", "pM1c" AND Tumor attribute NAACCR Variable 35918290 "Mets at DX-Liver" in combination with NAACCR Value 35941492 "Distant liver metastases" with a concept value of “Yes; distant liver metastases”
- Prior condition of SNOMED 196360 "Primary malignant neoplasm of bladder" or any of its descendants
- No prior diagnosis of other malignancies excluding non-melanoma skin cancers three years before the index date
- ≥18 years at index date

Follow up

Index date is defined as the **date of initiation of first LoT**. Patients are followed until death, last known activity or end of the study period at 31 July 2020.

Variables

Exposures

Treatment-related information are standardized through RxNorm, defining their active ingredient (agent), dose, and form. An algorithm is developed to parse regimens (defined through Hematology/Oncology (HemOnc))[7] from the individual drug exposure records by utilizing the ingredients, their timing and their dosing.

HemOnc treatment regimens are used to derive LoT. Date of the date of the first drug episode (e.g., first administration or non-cancelled order) for anti-neoplastic agents (ATC 21601387) after diagnosis of metastatic bladder cancer or bladder cancer with metastasis is considered as the start of the first LoT. For Claims databases, we require a six months washout period with no drug exposure to anti-neoplastic agents (ATC 21601387) prior to the initiation of the treatment regimen. The date patient discontinues first LoT treatment is considered the end date of the first LoT. Discontinuation is defined as having a subsequent systemic anti-neoplastic regimen after the first line of treatment; having a gap of more than 120 days with no systemic anti-neoplastic therapy following the last administration; or having a date of death while on the regimen. Patients are censored at their last known usage within the database or end of follow-up. Receipt of a new anti-neoplastic regimen after the end of the first LoT initiates a subsequent LoT; and the start and end of subsequent LoTs are identified using the logic described above. Same logic is used to derive the third LoT.

Outcomes

The following outcomes are assessed in the study:

Diagnosis to Treatment Interval (DT): the number of days from date of diagnosis to the initiation of the first LoT.

Treatment-free Interval (TFI): Time from discontinuation of one LoT to initiation of the subsequent LoT, or date of death if death occurs prior to start of the subsequent LoT. Patients are censored at their last activity within the database or end of follow-up.

Time to Next Treatment (TTNT): Time from the index date to the date the patient received their next systemic anti-neoplastic treatment regimen or to their date of death if death occurs prior to having another systemic anti-neoplastic treatment regimen. Patients are censored at their last activity within the database or end of follow-up.

Time to Treatment Discontinuation (TTD): Length of time from the initiation of each LoT to the date the patient discontinues the treatment (i.e., the last administration or noncancelled order of a drug contained in the same regimen). TTD is described for the first three LoTs. *Discontinuation* is defined as having a subsequent systemic anti-neoplastic therapy regimen after the first LoT; having a gap of more than 120 days with no systemic anti-neoplastic therapy following the last administration; or having a date of death while on the regimen. Patients are censored at their last known usage within the database or end of follow-up.

Overall Survival (OS): Length of time from the index date to the date of death, or loss to follow up or end of the study. Patients are censored at last recorded clinical activity within the database or end of follow up.

Covariates

Baseline covariates are defined by observations prior to or at the index date.

The following characteristics are identified **prior to or at index date**.

Age:
- age is categorized into 10-year groups. To ensure patient's privacy, patients above 85 are grouped together.
- age is also described as a continuous variable

Biologic sex
Race
Charlson comorbidity index
Diabetes
Hypertension
Cardiovascular disease
Cerebrovascular Disease
Chronic Obstructive Pulmonary Disease
Moderate/Severe Liver Disease
Mild liver disease
Renal disease
Dementia
Autoimmune disease
Serum lactate dehydrogenase (LDH) levels, Hemoglobulin and Creatinine levels
Site of metastasis **at index date**: categorized into lung, liver, bone, brain and others
Smoking

Study size

This study is undertaken using routinely collected data, all patients meeting the eligibility criteria above are included. No formal sample size and power calculation is performed.

Data Management

Source data at the participating institutions is converted to the OMOP CDM through a process known as ETL (extraction, transformation, loading)[8]. This process defines the rules for how source codes are mapped to standardized concepts in OMOP while maintaining quality of information. The data is refreshed periodically responding to updates by the data source. The OMOP process maps source codes to a standardized set of concepts depending on their domain. There are five key standardized domains: condition, drug, measurement, procedure and observation. Data sources participating in the study network does not provide ETL information. These data sources are responsible for their own validation to ensure that the OMOP dataset matched the quality of information provided in the original source.

Analysis

Demographic and clinical characteristics are summarized for both cohorts; metastatic bladder cancer and bladder cancer with metastasis to liver. For categorical variables, frequencies and percentages are presented for each level. Continuous variables are summarized and mean (standard deviation (SD)), minimum, maximum, median (within each database), and interquartile range (within each database). The proportions of missing data are captured for each variable. Standardized mean differences (SMD) are calculated when comparing characteristics of study cohorts.

Distribution of treatment regimens by LoT (up to three LoTs) is summarized for each cohort. Pending data availability, distribution of treatment regimens by year is summarized in each LoT and for each cohort.

For each cohort, DTI is reported as median (inter quartile range (IQR)). Time to evet outcomes including TFI, TTD, TTNT and OS are evaluated descriptively by Kaplan-Meier curves along with two-sided 95% (confidence interval (CI)) for the median time to event estimates.

The extracted treatment regimens are used to describe treatment flow across 1L, 2L and 3L treatments (treatment pathways) using Sankey diagrams.

Treatment regimens are categorized into the following groups:

- Cisplatin & Gemcitabine
- Methotrexate, vinblastine, doxorubicin, and cisplatin (MVAC)
- Other Cisplatin containing regimens
- Carboplatin & Gemcitabine
- Other Carboplatin containing regimens
- PD(L)1 monotherapy
- Gemcitabine monotherapy
- Paclitaxel monotherapy
- Others

All the proposed analyses are conducted for the top five treatment groups in each LoT, separately. Analyses are conducted in each database separately using R[9] and Redshift SQL.

Limitations

This study is carried out using data recorded in a collection of EHR systems. Misclassification may occur in identification of patients, outcomes and covariates. As with any healthcare database used for secondary data analysis, the patient records are expected to be incomplete in many respects and may have had erroneous entries. Data regarding diagnosis of metastatic disease, chemotherapy regimens used or baseline covariates prior to enrollment within the database may not be available. Treatment provided in hospitals or any other setting outside each participating institution is not included. Finally, comorbidities that could influence treatment given in the oncology clinic but that were themselves treated outside of the oncology setting tend to be somewhat under-reported in the databases. The primary purpose of the study is to describe a real-world cohort of patients with metastatic bladder cancer in the institutions participating in the study and, therefore, the results should not be interpreted as generalizable to the overall population. In the contrary, the institutions are expected to have preferences for patients with certain condition settings and differ in their selection of therapies.

Protection of Human Subjects

This study is conducted in accordance with the International Society of Pharmacoepidemiology (ISPE) Guidelines for Good Pharmacoepidemiology Practices (GPP)[10] and EMA, European Network of Centers for Pharmacoepidemiology and Pharmacovigilance (ENCePP) Guide on Methodological Standards in Pharmacoepidemiology[11]. The use of the OMOP common data model and OHDSI tools enable the federated analysis of these different databases without accessing patient-level data outside the participating institutions. The study is conducted in compliance with all applicable data protection, security and privacy laws, rules and regulations with respect to the collection, production, use, processing, storage, transfer, modification, deletion, and/or disclosure of any information related to this study under the Study Agreement. Each data partner is required to provide statement about IRB approval or exemption to participate.

Management and reporting of adverse events/adverse reactions

According to the new guidelines for good pharmacovigilance practice (EMA/873138/2011)[12] and ISPE[10], there is no requirement for expedited reporting of adverse drug reactions from studies with secondary use of data (such as electronic health care databases).

Plans for disseminating and communicating study results

Results of this study will be published following guidelines, including those for authorship, established by the International Committee of Medical Journal Editors[13]. When reporting results of this study, the appropriate Strengthening the Reporting of Observational Studies in Epidemiology checklist will be followed[13].

References

1. Powles T. Immunotherapy: The development of immunotherapy in urothelial bladder cancer. Nat Rev Clin Oncol. 2015;12(4):193-4.
2. Fleshner NE, Herr HW, Stewart AK, Murphy GP, Mettlin C, Menck HR. The National Cancer Data Base report on bladder carcinoma. The American College of Surgeons Commission on Cancer and the American Cancer Society. Cancer. 1996;78(7):1505-13.
3. Powles T. Immunotherapy: The development of immunotherapy in urothelial bladder cancer. Nat Rev Clin Oncol. 2015;12(4):193-4.
4. Hripcsak G, Ryan PB, Duke JD, et al. Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci USA. 2016;113(27):7329-36.
5. Banda JM. Fully connecting the Observational Health Data Science and Informatics (OHDSI) initiative with the world of linked open data. Genomics Inform. 2019;17(2):e13.
6. Wang Q, Reps JM, Kostka KF, et al. Development and validation of a prognostic model predicting symptomatic hemorrhagic transformation in acute ischemic stroke at scale in the OHDSI network. PLoS ONE. 2020;15(1):e0226718.
7. Warner JL, Dymshyts D, Reich CG, et al. HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. J Biomed Inform. 2019;96: 103239
8. OHDSI. The Book of OHDSI: OHDSI, 2019:458.
9. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/
10. International Society for Pharmacoepidemiology. Guidelines for good pharmacoepidemiology practices (GPP), Revision 3; June 2015. Available at: https://www.pharmacoepi.org/resources/policies/guidelines-08027/. Accessed 14 August 2020.
11. The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). Guide on Methodological Standards in Pharmacoepidemiology (Revision 5). EMA/95098/2010. Available at http://www.encepp.eu/standards_and_guidances. /. Accessed 14 August 2020.
12. European Medicines Agency. Guidelines on good pharmacovigilance practices (GVP): Module VIII – Post-authorisation safety studies. * EMA/813938/2011 Rev 3*. 9 October 2017. Available at: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-good-pharmacovigilance-practices-gvp-module-viii-post authorisation-safety-studies-rev-3_en.pdf. Accessed 6 Aug 2020
13. ICMJE. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. International Committee of Medical Journal Editors; December 2019. Available at: http://www.icmje.org/urm_main.html. Accessed 11 August 2020.
14. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008 Apr;61(4):344-9.