Paper RE09

Real-world outcomes and biomarker testing in cancer patients: exploration of a novel genetic database from routine clinical practice in England

Fiona C Ingleby, IQVIA, London, UK
Saskia P Hagenaars, IQVIA, London, UK
Alexandrina Lambova, IQVIA, Sofia, Bulgaria
Mounika Parimi, IQVIA, London, UK
Stephen Benson, IQVIA, London, UK
Sophie Jose, HDI, London, UK
Lora Frayling, HDI, London, UK
Valeria Lascano, IQVIA, London, UK

ABSTRACT

Genetic biomarker testing offers valuable insight into cancer diagnosis, prognosis, and treatment. This study explores potential for novel real-world biomarker data linked to the English cancer registry (National Cancer Registration Dataset, NCRD) to enable biomarker-stratified analyses of routine clinical care data. Using biomarker-tested cohorts of cancer patients diagnosed between 2012 and 2019, the content, structure, and capabilities of the data were explored. Patient cohort and diagnosis identification methods were examined, and descriptive analyses of timing of biomarker tests and biomarker test results were used to explore patient distributions and characteristics. Potential for biomarker-stratified analysis of outcomes was also explored via stratified analysis of overall survival by key biomarkers of interest per each cancer type. The results highlight the need for careful methodology and data checks to ensure reliable analyses, and consideration of multiple biomarker tests per patient for appropriate stratification of analyses. Overall, the study demonstrates the crucial novel genetic real-world data provided by NCRD for identifying genetic biomarkers and aiding research to improve cancer patient outcomes in a targeted therapy setting.

INTRODUCTION

The field of clinical oncology has progressed in recent years from making treatment decisions largely based on the type of tumour or anatomical location, to managing patient pathways based on genetic biomarker testing and known genetic alterations. As such, an increasing number of available therapies are tumour agnostic and instead target specific genes and alterations. In fact, approximately 90% of the oncology drugs approved by the FDA in 2020 were targeted therapies. In line with this new paradigm, cancer research is increasingly interested in real-world data that provides information in genetic biomarker testing in cancer patients. While genetic testing technologies are far from novel, and genetic data in clinical trials is relatively commonplace, there is a lag in terms of integrating genetic testing consistently into routine clinical care, and, subsequently, having data available from this part of the patient pathway in real-world clinical databases. This type of data is a key component of cancer research, by offering a real-world and population representative perspective on clinical care, as opposed to more targeted studies that rely on clinical trials or other voluntary sampled patient cohort studies.

In England, the National Cancer Registration Dataset (NCRD) holds information on all diagnosed malignancies amongst those resident in England. This data, alongside additional linked National Health Service (NHS) datasets, comprise the Cancer Analysis System (CAS) database. The CAS includes data on patient and tumour characteristics, death registrations, systemic treatments, radiotherapies, surgeries, and secondary care records in a hospital setting. This is a large population-representative database, representing approximately 99% of the English cancer patient population, and is accessible for research studies demonstrating clear patient benefit. The CAS data is not new itself, but the Molecular Diagnostics dataset within the CAS, CAS-MDx, is novel and presents an exciting opportunity to access and analyse genetic biomarker test data within a cancer registry setting. CAS-MDx currently contains genetic biomarker test records dated from start of 2016 until end of 2020, with updates expected annually. The test records cover a variety of testing methods, from molecular genetics lab testing and pathology lab testing, and therefore capture a wide range of types of genetic abnormalities. At time of writing, CAS-MDx contains >572,000 test records from >188,000 tumour samples and >300 labs across England.

In this paper, we summarise the methods and outcomes of an exploratory pilot study analysis of the CAS-MDx data, which aimed to explore data availability and potential capability to address research questions of key interest to studies of targeted therapies for cancer.

METHODS

The pilot study focused on cancer types with high incidence in the English population (breast cancer, BrCa; colorectal cancer, CRC; and non-small cell lung cancer, NSCLC) and a set of genetic biomarkers of interest within each cancer type depending on biomarkers of interest for targeted therapies in each setting (PIK3CA, HER2, and PD-L1 for BrCa; EGFR, ALK, ROS1, KRAS, BRAF, and PD-L1 for NSCLC; BRAF, MLH1, PMS2, MSH2, MSH6, KRAS, and NRAS for CRC). Study dates were determined largely by data availability at time of analysis (noting that the CAS data is regularly updated). Patients were included if they had a relevant incident diagnosis of BrCa, CRC, or NSCLC from 01-January-2012 until 31-December-2019 and they had at least one biomarker test record in CAS-MDx from 01-January-2016 until 31-December-2019 for one of the biomarkers of interest per each cancer type. Only adult patients aged 18 years or more at time of diagnosis were included. In line with requirements for use of the CAS data, patients who received treatments included in the Cancer Drugs Fund (CDF) at time of analysis were excluded from analyses. Patients were indexed at date of incident cancer diagnosis and followed up until date of death or censoring, whichever was earliest. Censor date was defined as the earliest date of either loss to follow-up or end of the available survival and treatments data, which at time of analysis was 31-August-2021.

A range of exploratory analyses were carried out, reported here under three broad scenarios aimed to explore data availability and capabilities:

(1) Identifying patient cohorts and diagnoses of interest – these steps involved cohort identification and overall descriptive analyses, with an overall aim to investigate feasibility and potential pitfalls in analysis of novel data;
(2) Describing genetic biomarker test data – descriptive summary statistics were used to report various patient demographics, clinical characteristics, and descriptive statistics of genetic biomarker test records and results, with an overall aim to explore available data;
(3) Analysis of overall survival stratified by biomarker status – with an overall aim to explore feasibility of biomarker-stratified research questions and stratification methods using the novel data.

Descriptive summary statistics were used to describe the data. Categorical variables were described using patient totals (N) and percentage (%) per category. Continuous variables were described using mean, standard deviation (SD), median, interquartile range (IQR), and 5th and 95th percentile values. Analyses were carried out using available data without any imputation of missing values, except for date variables, where missing day of the month was imputed as the 15th of the month. Kaplan-Meier methods were used to estimate overall survival, using time-to-event data per patient defined as time from index date to either date of an event (i.e., date of death from any cause), or date of censoring, whichever date is earliest per patient. Kaplan-Meier median survival (with IQR) was reported, scaled to months. Note that analysis results are not reported in full; instead, samples of representative results tables are shown to visualise example outputs from the novel data.

Analyses were carried out using R statistical software with the support of Health Data Insight CIC (HDI). Analysis programs were first prepared using the Simulacrum; a synthetic dataset designed to facilitate cancer research without the need for researchers to directly access any patient-level data. The programs were then run on CAS data via a virtual analysis session with HDI analysts. Outputs of analyses (i.e., aggregated tables of completed analyses, showing descriptive statistical results of the patient cohorts) were available following formal review to safeguard against release of identifiable information. Outputs were produced in a draft format for the purposes of the pilot study, which means all estimates are intended as illustrative only, with patient numbers rounded to the nearest 10; percentages rounded to the nearest 5; and masking applied to patient numbers <10.

RESULTS

Following the patient selection criteria outlined above, we included 4,320 BrCa patients, 52,690 NSCLC patients, and 32,020 CRC patients in our exploratory analyses. Note that these cohorts are not intended to fully capture all the patient data available in CAS-MDx, but instead to focus on specific patient cohorts with commonly-occurring cancer diagnoses and specific genetic biomarker test results of interest, as defined in the Methods.

(1) Identifying patient cohorts

For real-world clinical epidemiological studies, a key step is the identification of a cohort of eligible patients who are uniquely identified in the database. The CAS database uses patient ID numbers to achieve this reliably, but note that study documents will need to precisely define which records to use per patient, and account for potential of multiple tumour diagnosis records per patient. In addition, records in CAS-MDx are uniquely identified per each combination of tumour ID and name of gene. As such, CAS-MDx is likely to contain multiple rows of data per a patient ID, via a combination of multiple potential tumours per patient and multiple potential genes tested per tumour. This means that studies using the novel CAS-MDx data will need to additionally define precisely which biomarker records to use per patient. Note that the unique records in CAS-MDx per tumour and gene combination are achieved via data aggregation by the data holders, and in cases of multiple tests for a given tumour and given gene, some granularity of test-level information is not accessible in CAS-MDx. Based on the results of the sense checks carried out in this pilot study, we recommend the following steps: (i) exclude patients with prior malignancies from studies to focus on incident cases; (ii) consider only the earliest eligible incident tumour diagnosis record per patient; (iii) exclude biomarker test records that are

dated significantly earlier than diagnosis date; and (iv) take particular care with studies concerned with multiple cancer types and/or multiple biomarkers of interest, to ensure appropriate cohort definitions.

(2) Describing genetic biomarker test data

Using linkage between CAS-MDx and other datasets within CAS that contain patient demographics and clinical characteristics, we were able to describe the patient cohorts included in our analyses. A sample of this analysis is shown in Table 1 (the variables indicated are not exhaustive but intended to be illustrative). Age at diagnosis, sex, ethnicity, and stage at diagnosis were each largely representative of what would be expected for these cancer cohorts overall, suggesting good coverage of the overall cohorts within CAS-MDx. Distribution across socio-economic status showed some variation from what would be expected in the overall cancer patient cohort, indicating that the biomarker-tested cohorts may not be fully representative of a general population. Although a limitation of the data, note that the CAS database and CAS-MDx are routinely updated, and so coverage may be expected to increase over time.

Table 1. Descriptive summary statistics of a sample of patient demographic and clinical characteristics for each cohort.

	Patient cohort	BrCa		NSCLC		CRC	
		N	%	N	%	N	%
	Cohort size	4320	5%	52690	60%	32020	35%
Age in years at diagnosis	18-30	30	0%	50	0%	220	0%
	31-40	260	5%	300	0%	1340	5%
	41-50	720	15%	1980	5%	2530	10%
	51-60	1070	25%	7330	15%	6030	20%
	61-70	1100	25%	17220	35%	8870	30%
	71-80	750	15%	19400	35%	9400	30%
	81+	380	10%	6420	10%	3620	10%
Sex	Male	20	0%	27680	55%	18260	55%
	Female	4290	100%	25010	45%	13750	45%
Ethnicity	White	3940	90%	48480	90%	28370	90%
	Asian	110	5%	940	0%	820	5%
	Black	50	0%	630	0%	590	0%
	Chinese/other	50	0%	730	0%	510	0%
	Mixed	20	0%	200	0%	160	0%
	Missing	140	5%	1720	5%	1560	5%
Socio-economic status	1 - most deprived	890	20%	13210	25%	5420	15%
	2	740	15%	10830	20%	5850	20%
	3	910	20%	10390	20%	6580	20%
	4	950	20%	9810	20%	7100	20%
	5 - least deprived	830	20%	8460	15%	7060	20%
Stage at diagnosis	I	1570	35%	6940	15%	3150	10%
	II	1610	35%	4360	10%	6690	20%
	III	420	10%	12180	25%	9770	30%
	IV	200	5%	27900	55%	10060	30%
	Missing	520	10%	1300	0%	2350	5%

In addition to patient characteristics, we were also able to explore availability of data pertaining to the biomarker tests and results. Full information of available data can be found in NCRAS resources online, but a sample of variables is illustrated in Table 2, using an example of CRC patients with MSH2 biomarker test records in CAS-MDx. As shown, a range of details on test results, types of genetic abnormality, and test dates can be accessed. This data may be useful for reporting prevalence of abnormities, types of abnormality and testing patterns. For example, prevalence of MSH2 abnormalities in the CRC patient cohort was ~1%. It may also be useful for further analyses stratified by biomarker status (see subheader [3] below).

We were able to explore more fine-grained genetic analysis and identification of specific genetic sequence variants of interest in the pilot study cohorts (results not shown). This is made possible via structured data fields within CAS-MDx which contain information on DNA sequence changes identified per tumour sample, using standard HGVS notation. This data can be parsed with careful data cleaning and use of online databases of mutations.

Table 2. Descriptive summary statistics of a sample of MSH2 biomarker test result data fields for CRC patients.

		N	%
	CRC with MSH2 test cohort size	19590	100%
Year of test	2016	560	5%
	2017	1310	5%
	2018	3290	15%
	2019	14430	75%
Overall biomarker status	Normal	18980	95%
	Abnormal	340	0%
	Borderline	50	0%
	Failed	60	0%
	Missing	170	0%
Number of abnormalities per patient	0	19250	100%
	1	320	0%
	2	20	0%
Type of abnormality	DNA sequence	40	0%
	Methylation	0	0%
	Under-expression	280	0%
	Over-expression	0	0%
	Copy number loss	0	0%
	Copy number gain	0	0%
	Copy number neutral	0	0%
	Fusion/translocation	0	0%
	Multiple	20	0%
	No abnormality	19080	95%
	Missing	170	0%
Time between first and last biomarker test date	Total	19590	100%
	Mean (SD)	14.93 (80.87)	
	Median (Q1-Q3)	0.00 (0.00-0.00)	
	Percentile 5th - 95th	0.00 - 43.00	
	N	19590	100%
	Missing	0	0%

(3) Analysis of overall survival stratified by biomarker status

Analysis of overall survival by biomarker status is feasible using CAS-MDx, and a range of different approaches can be considered. Choice of method will depend on the research question of interest, and also depend on sample size considerations to determine how fine-grained stratifications can be. Another key consideration will be whether mutually exclusive strata are needed for statistical comparisons, as opposed to overlapping sub-cohorts of patients which may be sufficient for descriptive analysis only.

For this exploratory analysis, we took an approach that considered all available biomarker test results for the full set of biomarkers selected of interest per each cancer patient cohort. We allowed overlap between sub-cohorts defined by abnormalities in each biomarker. A sample of the survival analysis results is shown in Table 3, which contains the survival estimates for breast cancer patients split into the following overlapping sub-cohorts: (i) patients with no known abnormalities for any of the biomarkers of interest; (ii) patients with a PIK3CA abnormality; (iii) patients with a HER2 abnormality; and (iv) patients with a PD-L1 abnormality. The results allow reliable description of clinical outcomes across patient groups defined by biomarker status, and, in the case of the breast cancer patient cohort example shown in Table 3, longer survival was observed in the patients with no known abnormalities for the biomarkers of interest. While this approach has the advantage of creating clear and biologically-meaningful groups for analysis, a potential drawback is that patients can fall into more than one 'abnormality' sub-cohort, if they have multiple co-occurring mutations, meaning that statistical comparisons between groups would not be not valid. In addition, the sub-cohort without any evidence of abnormalities in the biomarkers of interest should be interpreted with caution, given that this group is a mix of 'normal' test results and 'unknown' biomarker status, and patients may have abnormalities for other biomarkers or have unknown mutations.

Table 3. Descriptive summary statistics from a Kaplan-Meier analysis of overall survival timed from date of diagnosis for BrCa patients grouped in mutational sub-cohorts for biomarkers of interest.

	Normal/unknown result for PIK3CA, HER2, and PD-L1	Abnormal result for PIK3CA	Abnormal result for HER2	Abnormal result for PD-L1
N patients	3580	30	610	100
N events	740	30	110	60
N censored	2840	<10	500	40
Median OS (months)	80	59.88	76.48	40.08
IQR Q1-Q3 (months)	58.09 - 100.34	40.87 - 97.02	61.14 - NR	20.24 - 83.94

Notes: Patient sub-cohorts may overlap, see text for detail; IQR=inter-quartile range; NR = not reached; 'HER2' refers to protein coded for by ERBB2 gene.

CONCLUSION

Overall, the exploratory analyses were conducted successfully and demonstrate as a general proof-of-concept that CAS-MDx can be applied to cancer research precision medicine studies to answer questions relating to patient demographics and clinical characteristics, biomarker testing patterns, clinical outcomes, and genetic biomarker-stratified analyses. In addition, although not explicitly reported as part of the pilot study, this proof-of-concept equally applies to other research questions that were explored with other component parts of the CAS data, for example, making use of treatments data or secondary care hospitalization records to demonstrate HCRU. The scope of possible research questions with such a rich database is extremely broad. That said, there are some considerations of the novel CAS-MDx data to note when planning further research:

(1) The CAS data overall includes approximately 99% of the English cancer patient population, however, CAS-MDx is not yet fully representative of the overall cohort. Biomarker test data from pathology labs represents the full population for diagnoses from 2019 onwards. Data from the molecular labs is not fully representative as yet, but is expected to improve with subsequent data refreshes. In spite of this limitation, CAS-MDx covers large numbers of patients and tests, and may be more generalizable as compared to other available databases based on much smaller, sampled cohorts.

(2) Linkage between a variety of datasets within the CAS offers data richness and flexibility, but note that reliable use of the data will require detailed knowledge of the database. It is recommended to make use of sufficiently-detailed data checks as part of a study.

(3) Patients can have multiple different biomarker test records for any given tumour, and equally there can be multiple tests carried out for the same biomarker for a given tumour. This is the nature of biomarker testing procedures within routine clinical care, and is reflected in the data. It may even be an aspect of interest in the data, depending on the focus of a study. CAS-MDx records are aggregated to a level where records are unique per tumour and biomarker combination. While this aggregation simplifies the overall database structure, there is potential loss of granularity of information in cases where there are multiple tests of interest within a unique tumour and biomarker combination.

(4) CAS-MDx data represents structured, processed genetic test data, as opposed to raw genomics data. This simplifies the use and interpretability of the data, however, it is important to note that analysis methods should still be informed by genetic knowledge.

(5) Patients may have multiple genetic abnormalities identified in the data. Any planned stratified analyses will need to account for this on a case-by-case basis. Depending on the research question of interest, analysis plans should consider appropriate statistical comparisons of either mutually exclusive strata or overlapping sub-cohorts, as needed.

The CAS data therefore represents an accessible, rich, population-representative repository of cancer patient data based on the English cancer registry, for use in research studies with clear patient benefit. Data on patient characteristics, treatments, secondary care, clinical outcomes, and genetic biomarker testing are now all available within the CAS data. The novel addition of genetic biomarker data in CAS-MDx provides data for large parts of the overall CAS population and introduces opportunity for future research studies to include analyses of genetic biomarker testing patterns, patient pathways, and genetically-stratified outcomes.

ACKNOWLEDGMENTS

This work uses data that has been provided by patients and collected by the NHS as part of their care and support. The data are collated, maintained and quality-assured by the National Disease Registration Service, which is part of NHS England. Access to this data was facilitated by the Simulacrum produced by Health Data Insight CIC with generous support from AstraZeneca and IQVIA.