

QResearch primary care records linked to national hospital and cancer registry data: a validation study



Kulkarni S,¹ Perletta V,¹ Wang Z,² Tomlinson JW,³ Hippisley-Cox J,⁴ Collins GS,⁵ Clift AK,⁶ Dodwell D,² Lord SR¹

¹ Department of Oncology, University of Oxford; ² Nuffield Department of Population Health, University of Oxford; ³ Oxford Centre for Diabetes, Endocrinology & Metabolism and NIHR Oxford Biomedical Research Centre, University of Oxford; ⁴ Wolfson Institute of Population Health, Queen Mary University of London; ⁵ Department of Applied Health Sciences and NIHR Birmingham Biomedical Research Centre, University of Birmingham; ⁶ Department of Surgery & Cancer, Imperial College London.

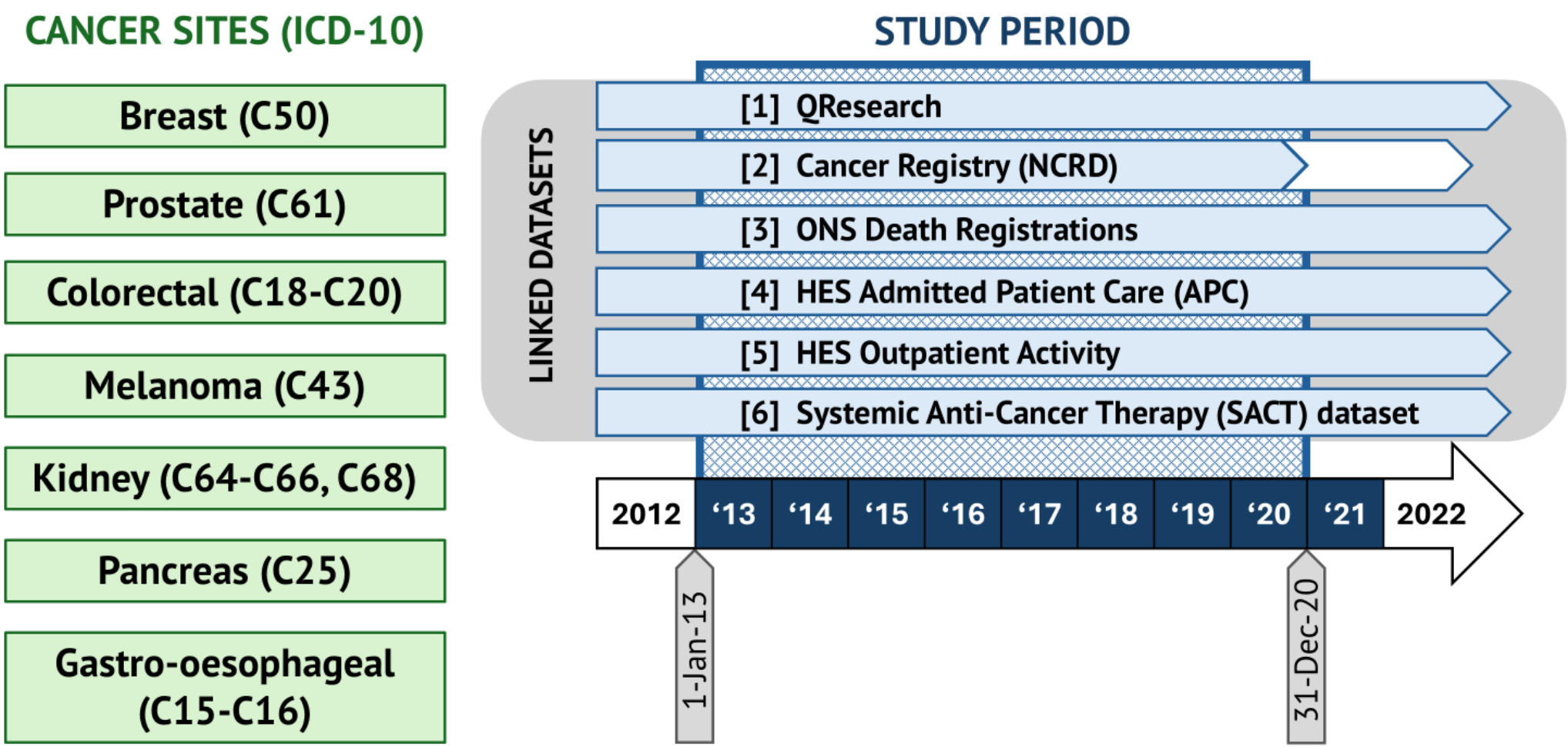
BACKGROUND

- Personalised cancer research increasingly relies on linked real-world population level data from electronic health records (EHRs)
- QResearch is a large, primary care EHR database with data from >1,500 general practices in England and >40 million patients, linked to national cancer registry data¹
- Primary care data offer detailed information on comorbidities, medications and lifestyle factors often missing from registries
- With no single national source of primary care data in England, linkage of such data limits analysis to a subset of the English population – validation is essential to ensure representativeness
- As a highly curated dataset, availability of national cancer registry data can lag behind other EHR data sources
- STUDY AIM:** to assess the demographic and treatment representativeness of the QResearch-linked cancer dataset against national cancer statistics (NDRS) and evaluate the impact of cancer registry censoring at 2020 on cohort composition

METHODS

- QResearch primary care data were linked to five national EHR data sources, including the national SACT dataset² (Fig 1.)

Fig 1. Linked data sources and included cancer sites by ICD-10 code groups



- Study population: adult patients with cancers diagnosed 2013-2020 across seven common sites (Fig 1.), identified in either QResearch, HES, the cancer registry (NCRD), or death registrations
- Treatments were identified using SACT records, HES OPCS codes and NCRD treatment flags, using NDRS-defined treatment attribution windows specific to tumour site and modality
- Demographic representativeness was assessed by comparing distributions reported as *n* (%) for sex, age, year, region, stage, ethnicity and deprivation (IMD) with NDRS statistics
- Treatment representativeness assessed by comparing absolute differences in proportions receiving systemic anti-cancer therapy (SACT) or radiotherapy with NDRS treatment statistics
- NCRD censoring was evaluated by characterising tumours identified only in the registry in comparison to other cases

RESULTS

- 196,565 patients with cancers across seven sites were captured in the QResearch-linked data, from 1,348,243 eligible national registrations in England (2013-2020), equating to 14.6% of cases
- Capture increased over time from 11.9% (2013) to 17.2% (2020), consistent with expanding practice participation

REFERENCES

- QResearch. QResearch: Generating new knowledge to improve patient care, <<http://www.qresearch.org/>> (2025).
- Bright, C. J. et al. Data Resource Profile: The Systemic Anti-Cancer Therapy (SACT) dataset. Int J Epidemiol 49, 15-151 (2020). <https://doi.org/10.1093/ije/dyz137>

ACKNOWLEDGEMENTS

We acknowledge the contribution of EMIS practices who contribute to QResearch® and NHS England, Public Health England and the Office for National Statistics for providing the HES, Cancer Registry, SACT and death registration datasets used in this analysis. Funding for grant IIG_FULL_2023_007 was obtained from World Cancer Research Fund (WCRF), as part of the World Cancer Research Fund International grant programme.

DEMOGRAPHIC REPRESENTATIVENESS

- Age, sex, stage, ethnicity and deprivation distributions closely matched national cancer statistics across all sites (Table 1)
- Regional variation reflected QResearch practice distribution but did not materially affect demographic or clinical characteristics of the cohort (Fig 2.)

Fig 2. Comparison of region of residence

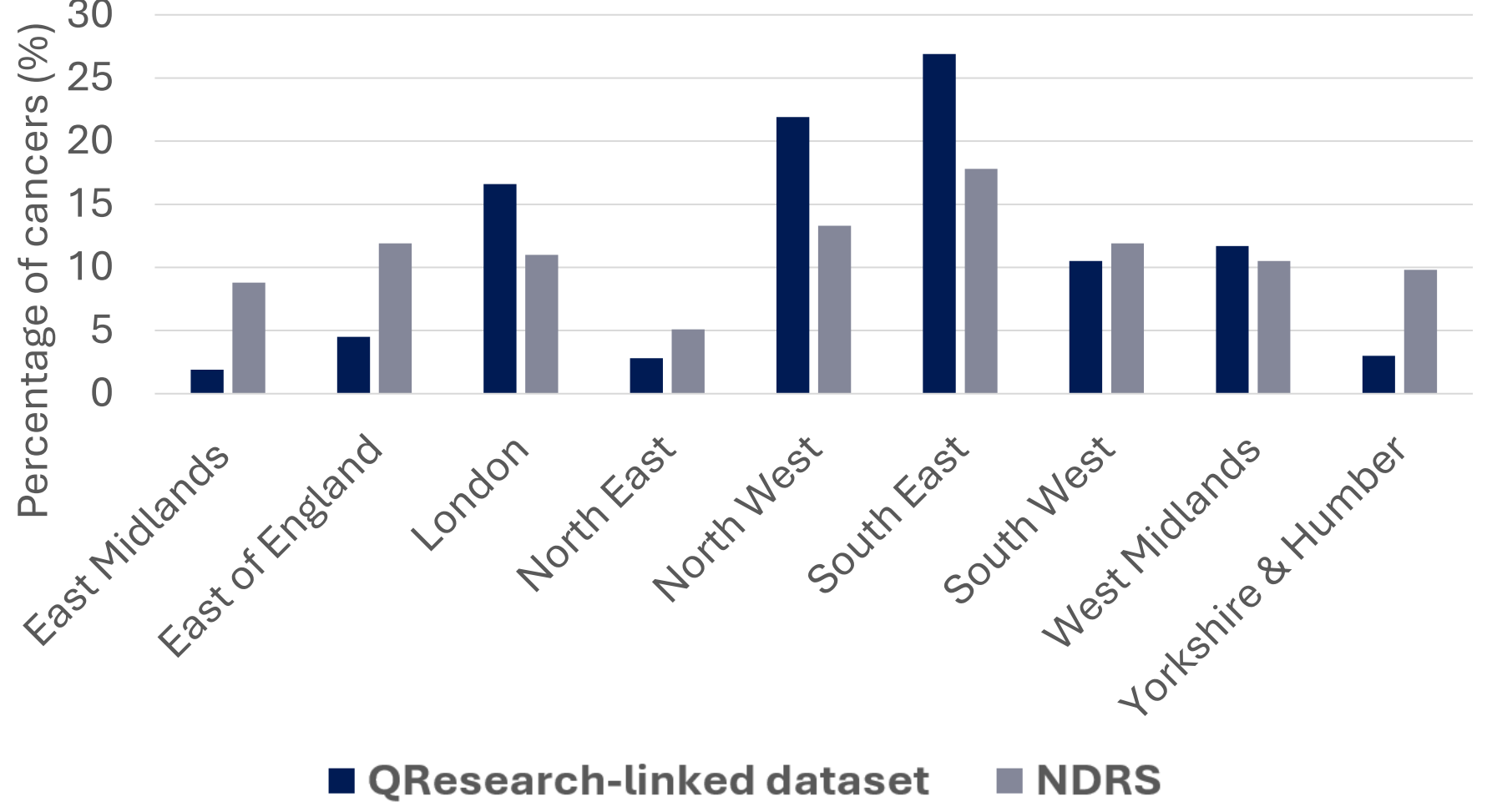


Table 1. Comparison of demographic characteristics

Characteristic	QResearch N (%)	NDRS N (%)
Total cancers	196,565	1,348,243
Sex		
Male	104,253 (53.0)	707,447 (52.5)
Female	92,313 (47.0)	640,796 (47.5)
Age at diagnosis (years)		
<70	95,596 (48.6)	673,375 (49.9)
≥70	100,969 (51.4)	674,868 (50.1)
Ethnicity†		
White	48,626 (91.6)	282,208 (92.6)
Asian	1,727 (3.3)	8,715 (2.9)
Black	1,311 (2.5)	7,295 (2.4)
Other	1,412 (2.7)	6,379 (2.1)
Stage†		
Stage I–II	24,982 (58.9)	159,505 (59.5)
Stage III–IV	17,405 (41.0)	108,032 (40.3)
IMD quintile†		
1 (most deprived)	7,967 (15.0)	51,136 (15.1)
2–3	20,699 (38.9)	131,937 (38.9)
4–5 (least deprived)	24,559 (46.1)	155,607 (45.9)

Footnote: Unknown categories not presented; percentages calculated excluding unknown values. † Available for diagnoses in 2019–2020 only.

TREATMENT REPRESENTATIVENESS

- Proportions receiving SACT and radiotherapy closely aligned with national statistics across sites and years (Fig 3. and Fig 4.), with absolute differences generally within 2-3%

Fig 3. Comparison of treated proportions

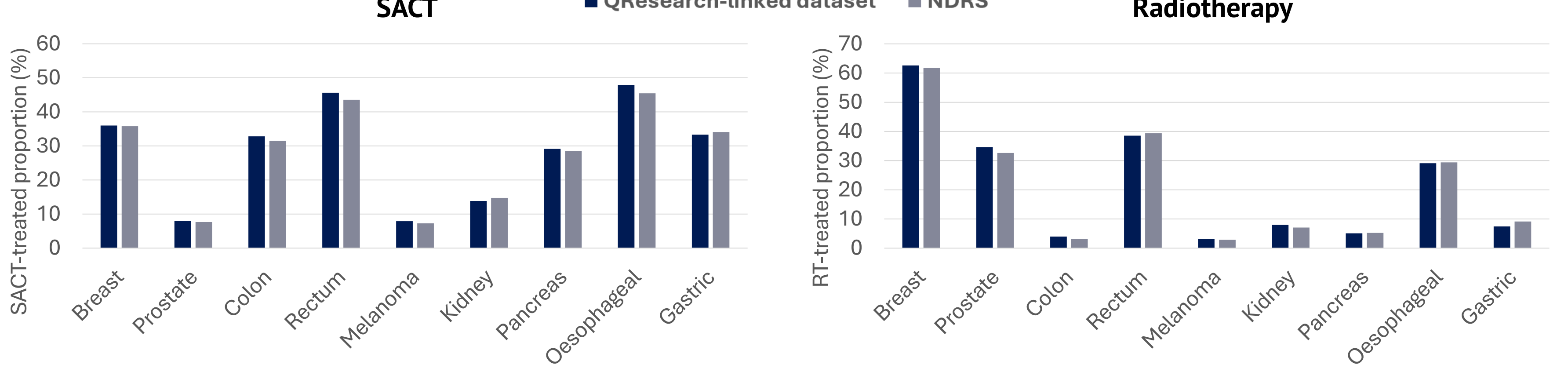


Fig 4. Absolute percentage difference in SACT-treated proportions (QResearch-NDRS)

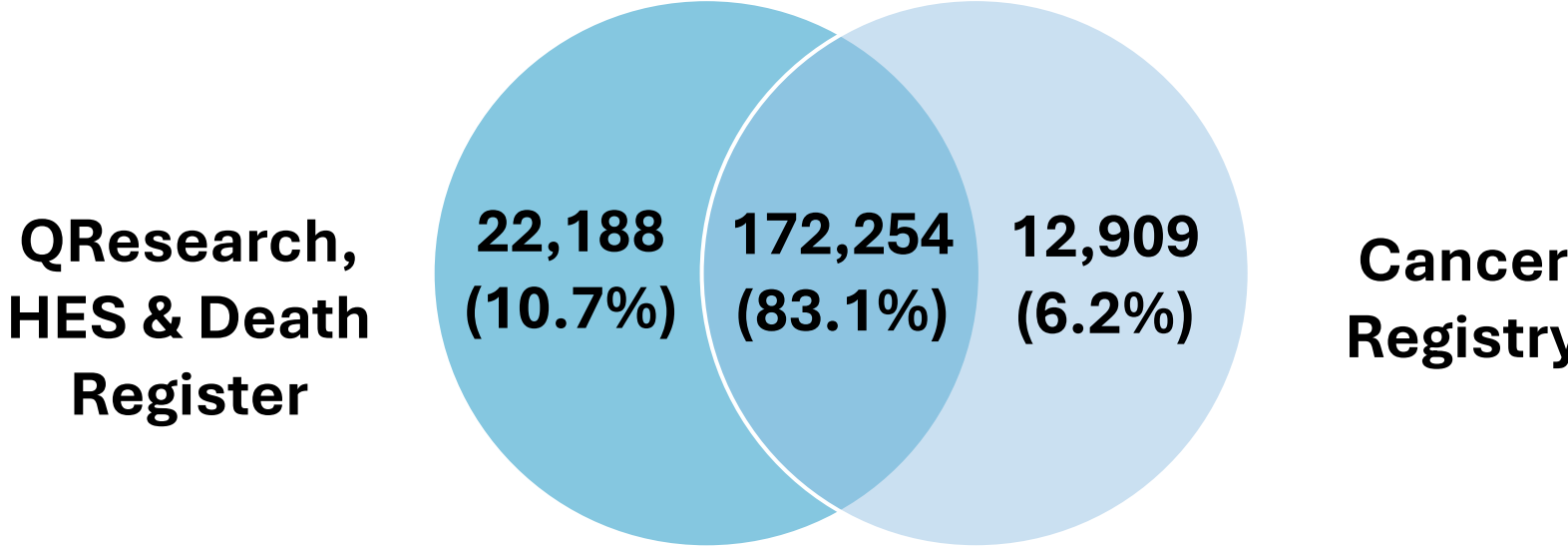
	Breast	Prostate	Colon	Rectum	Melanoma	Kidney	Pancreas	Oesophageal	Gastric	Legend
2013	1.4	0.1	3.2	4.1	-0.7	0.3	0.0	5.6	2.3	Over-capture 10
2014	1.0	0.0	1.1	1.8	0.3	-1.3	4.5	1.8	-2.1	5
2015	-0.4	0.2	1.6	1.8	0.4	-1.1	-2.3	-0.4	-4.0	2.5
2016	0.6	-0.2	0.3	-0.3	0.6	-1.5	-0.8	4.7	-3.1	0
2017	0.4	-0.3	2.0	3.2	0.4	-0.4	1.3	2.2	0.3	-3
2018	-0.8	0.0	0.8	2.0	1.1	-1.7	1.2	2.5	1.0	-5
2019	-0.2	-0.2	0.8	1.2	-0.5	-1.6	0.2	1.3	-1.3	-10
2020	-0.3	0.8	1.6	3.0	0.1	-0.2	0.9	2.7	3.1	
Overall	0.2	0.3	1.3	2.1	0.6	-0.9	0.6	2.5	-0.8	

Bold value indicates a statistically significant difference based on non-overlapping 95% confidence intervals

IMPACT OF REGISTRY CENSORING

- 93.8% of patients with cancer (*n*=194,442) remain identifiable without registry linkage
- Registry-only cases were older (≥80y: 27.7% vs 21.5%), earlier stage (stage I-II: 76.3% vs 58.8%), and less often treated (SACT: 12.3% vs 26.3%; RT: 14.4% vs 31.7%)
- Excluding them has minimal impact on cohort structure or treatment patterns

N = 207,351 eligible cancer records (2013-2020)



SUMMARY

- The QResearch-linked cancer dataset provides a validated, nationally representative cancer cohort with close alignment to national statistics in terms of demographics and treatment
- Surgical resections were not considered due to data availability
- Linked rich primary care data offers the opportunity for detailed evaluation of treatment selection, toxicities and outcomes, including the role of multimorbidity, polypharmacy and lifestyle factors
- In summary, the linked dataset offers a robust platform for real-world population oncology research



UNIVERSITY OF
BIRMINGHAM

IMPERIAL



Queen Mary
University of London