What counts as a cancer diagnosis in real world data?
Cohort identification and oncology endpoints in a linked genomic-clinical database

Sudhir Venkatesan[*1] (Sudhir.Venkatesan@iqvia.com), Nicole Rutishauser*[*1], Fiona Grimson[1], Lorena Cirneanu[1], Svetomir Hitov[1], Valeria Lascano[1], Mounika Parimi[1], Rosalind Polya[1], Georgia Chan[2], Genomics England Research Consortium[2], Benjamin Bray[1]*

[1]EMEA Centre of Excellence for Retrospective Studies, IQVIA , [2]Genomics England, London, UK

** Joint first authors*

Background and objectives

- Participants in the Genomics England (GE) database are consenting NHS patients whose clinical samples are sequenced using Whole Genome Sequencing (WGS) technology
- All cancer health records gathered by the NHS are collected by the National Cancer Registration and Analysis Service (NCRAS) and consolidated under Cancer Outcomes and Services Dataset (COSD)
- While linking clinical health records and genomic outputs can have great potential, linking multiple sources of information can create difficulties with conflicting data or multiple methods of measurement
- Objective: To explore cohort identification approaches for Non-Small Cell Lung Cancer (NSCLC) patients in the linked Genomics England (GE) database, evaluate the level of linkage with clinical databases and its impact on oncology endpoints

Methodology

- Diagnosis information from multiple sources was combined to identify NSCLC diagnoses between 2015 and 2017
- Diagnosis will be compared between GE's Cancer Analysis (CA) and Cancer Participant Tumour (CPT) datasets, NCRAS and HES using Fleiss' kappa statistic (across all lung cancer and subsequent NSCLC diagnoses) and Cohen's kappa (CA and NCRAS) to determine the Inter-Rater Reliability (IRR) between these data sources
- Overall Survival (OS) was the primary clinical endpoint. The impact on OS from using cancer diagnosis information (diagnosis and dates) from different data sources was assessed and analyses were stratified by cancer staging (Kaplan-Meier analysis)

Results

	CA	CPT	NCRAS	HES	% Agreement	Cohen's kappa (CA vs. NCRAS)	Fleiss' kappa
Total cancer diagnoses†	*7868*	*4833*	*7868*				
Lung Cancer (2015 – 2017)	477*	373	638	663	97.00%	0.822	0.816
NSCLC (2015 – 2017)	408	373	623		97.20%	0.818	0.805
Adenocarcinoma	212	51	317		96.80%	0.695	0.484
Squamous cell	146	33	206		98.00%	0.768	0.521

*326 patients missing diagnosis date; †In patients with a single primary tumour and age ≥18 years at diagnosis

Table 1: Diagnosis concordance and IRR across four data sources

		CA	CPT	NCRAS	CA with HES proxy
Total	**Number of Patients**	**405**	**371**	**623**	**603**
	Number of Censored Patients	**337**	**317**	**515**	**503**
	Median [95% CI]	**not reached**	**not reached**	**not reached**	**not reached**
	Q1-Q3	**n.a-n.a**	**n.a-n.a**	**36.96-n.a**	**37.22-n.a**
6 Months	Number of Patients	370	336	596	560
	Probability [95% CI]	0.95 [0.93,0.97]	0.96 [0.94,0.98]	0.96 [0.94,0.97]	0.95 [0.93,0.97]
12 Months	Number of Patients	309	291	568	479
	Probability [95% CI]	0.90 [0.87,0.93]	0.92 [0.89,0.95]	0.91 [0.89,0.93]	0.90 [0.88,0.93]
18 Months	Number of Patients	196	189	349	296
	Probability [95% CI]	0.84 [0.80,0.88]	0.86 [0.82,0.90]	0.86 [0.83,0.89]	0.85 [0.81,0.88]
24 Months	Number of Patients	116	114	185	168
	Probability [95% CI]	0.80 [0.76,0.85]	0.82 [0.78,0.87]	0.81 [0.78,0.85]	0.81 [0.77,0.85]
30 Months	Number of Patients	53	52	83	83
	Probability [95% CI]	0.77 [0.72,0.83]	0.81 [0.76,0.86]	0.77 [0.72,0.82]	0.78 [0.73,0.82]
36 Months	Number of Patients	18	16	34	35
	Probability [95% CI]	0.76 [0.70,0.82]	0.79 [0.73,0.85]	0.75 [0.70,0.81]	0.77 [0.72,0.82]
42 Months	Number of Patients	<10	<10	11	11
	Probability [95% CI]	0.76 [0.70,0.82]	0.79 [0.73,0.85]	0.70 [0.62,0.79]	0.72 [0.64,0.80]

Table 2: Kaplan-Meier summary estimates comparing overall survival in NSCLC patients identified in four different data sources (diagnosis data obtained from each individual data source, death data obtained from NCRAS)

Results (cont'd)

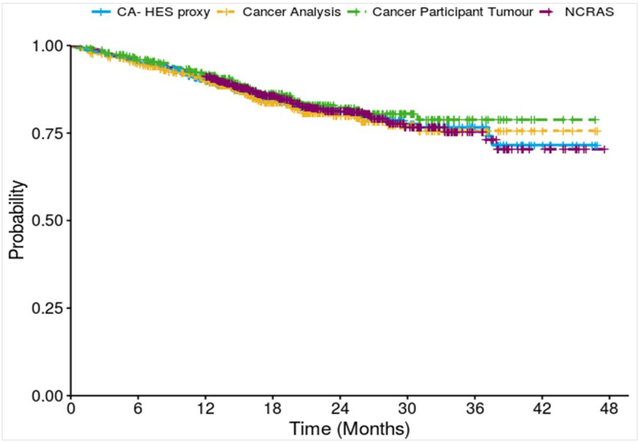

Figure 1: Kaplan-Meier plots showing overall survival in NSCLC patients identified in four different data sources

Conclusions

- Linked databases often contain multiple sources of information; it is important to consider the provenance of individual variables
- NCRAS (cancer registry) contains the most detailed and validated cancer diagnosis information
- GE data linked with the NCRAS and HES datasets provide a rich dataset well-suited for use in genomic-clinical studies

Acknowledgement

This research was made possible through access to the data and findings generated by the 100,000 Genomes Project; http://www.genomicsengland.co.uk

ISPE 36th International Conference on Pharmacoepidemiology & Therapeutic Risk Management, 16-17 September 2020