Together we are beating cancer

Using multiple imputation to account for missing stage and to generate a population-based late-stage incidence rate

Shane Johnson[1,2,5], Kirstin Roberts[1,2], Sophia Lowes[1], Fahmina Fardus-Reid[1], Carolynn Gildea[2], Sean McPhail[2], Jason Poole[3], David Foord[3], Lorna Wills[1], Samantha Hinks[3], Mark Rutherford[4], Matthew Barclay[5], Jon Shelton[1]

[1]Cancer Intelligence, Cancer Research UK, [2]National Disease Registration Service, NHS England, [3]Cancer Analysis and Insights Team, NHS England, [4]Division of Public Health & Epidemiology, University of Leicester, [5]Epidemiology of Cancer Healthcare & Outcomes, University College London

Background

- Reliable stage at diagnosis data are important for monitoring ambitions to diagnose cancers at an earlier stage.
- Completeness of staging data is variable over time, making accurate population-based surveillance of stage at diagnosis difficult.
- As an alternative to stage proportion metrics[1], a rate-based metric of stage at diagnosis has been proposed to aid interpretation of trends, by taking into account population and age structure, and changes in cancer incidence over time.
- We aimed to develop a robust multiple imputation model for missing stage data, enabling calculation of incidence rates by stage at diagnosis

Methods

Data were extracted for cancer diagnoses in England from 2014 to 2022 from the National Cancer Registration Dataset (NCRD). Multiple imputation (MI) models were run to impute missing stage at diagnosis values.

Key variables included in the imputation models:

- tumour characteristics (diagnosis year, stage at diagnosis, cancer site, route to diagnosis [screening, emergency presentation or other])
- patient characteristics (age, gender and deprivation quintile)
- treatment (fact of first systemic anti-cancer therapy, radiotherapy and surgery[2])
- vital status (at 12 months post-diagnosis for cumulative hazard[3])

Stage distribution of the original and imputed datasets were compared to visualise the impact of stage imputation.

We conducted a sensitivity analysis where we compared imputation results from models with and without treatment information. Using complete case data, we made 20% of known stage values missing-completely-at-random. Stage was imputed using each model and we assessed concordance between complete case and each set of simulated stage values using:

- Stage distribution
- Cohen's kappa – measure of agreement between the known and imputed stage values[4]

National age- and gender- standardised late-stage (stage 3 and 4) incidence rates were calculated for each imputed dataset and pooled using Rubin's rules to provide an average rate[5].

Conclusions

- Multiple imputation can be a useful tool to address challenges of missing data in cancer registries.
- Using national registry data to include both survival and treatment information has enhanced the strength of the model.
- For the 19% of records missing stage, the imputed values indicated that stage 4 diagnoses were predicted more often than other stages.
- Our sensitivity results showed that including treatment details in the model improved accuracy in imputation at a tumour-level.
- Use of MI brings more stability to rates of stage of disease and can be used for creating indicators to support the evaluation and impact of national cancer strategies.

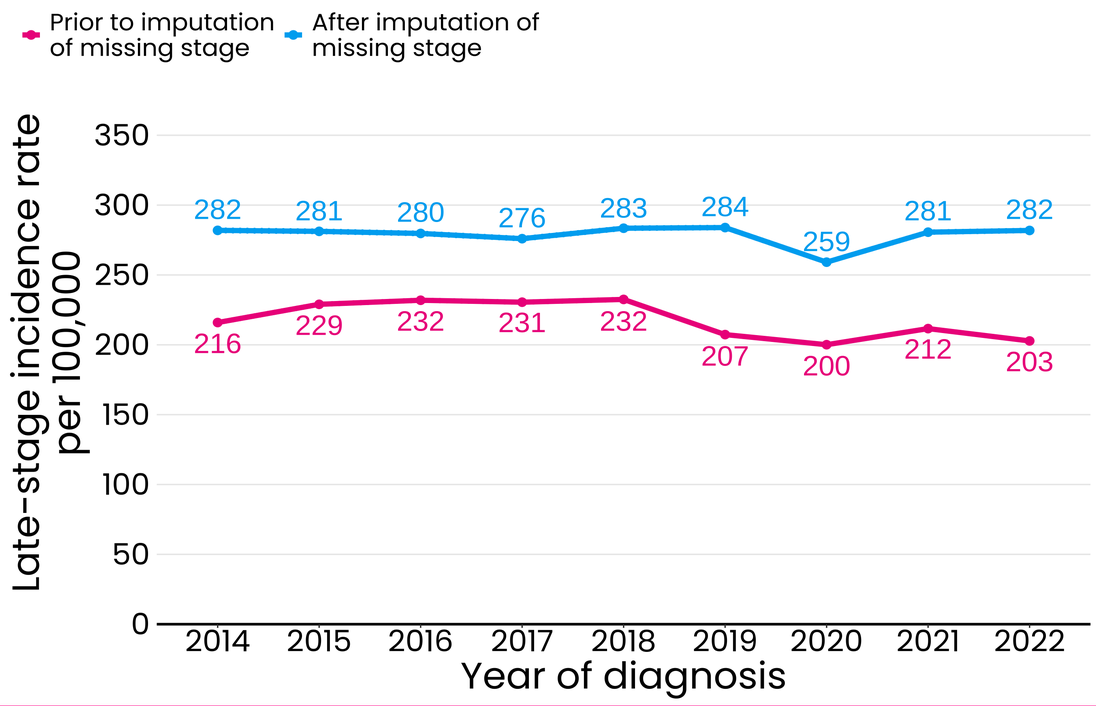
This work uses data that has been provided by patients and collected by the NHS as part of their care and support. The data are collated, maintained and quality assured by the National Disease Registration Service, which is part of NHS England.

Contact: cancerintelligence@cancer.org.uk

Results

After imputing missing stage, late-stage incidence rates were more consistent and higher over time compared to incidence rates prior to imputation (**Figure 1**). Late-stage incidence rates in 2022 were 282 late-stage diagnoses per 100,000 persons where stage was imputed compared to 203 prior to imputation.

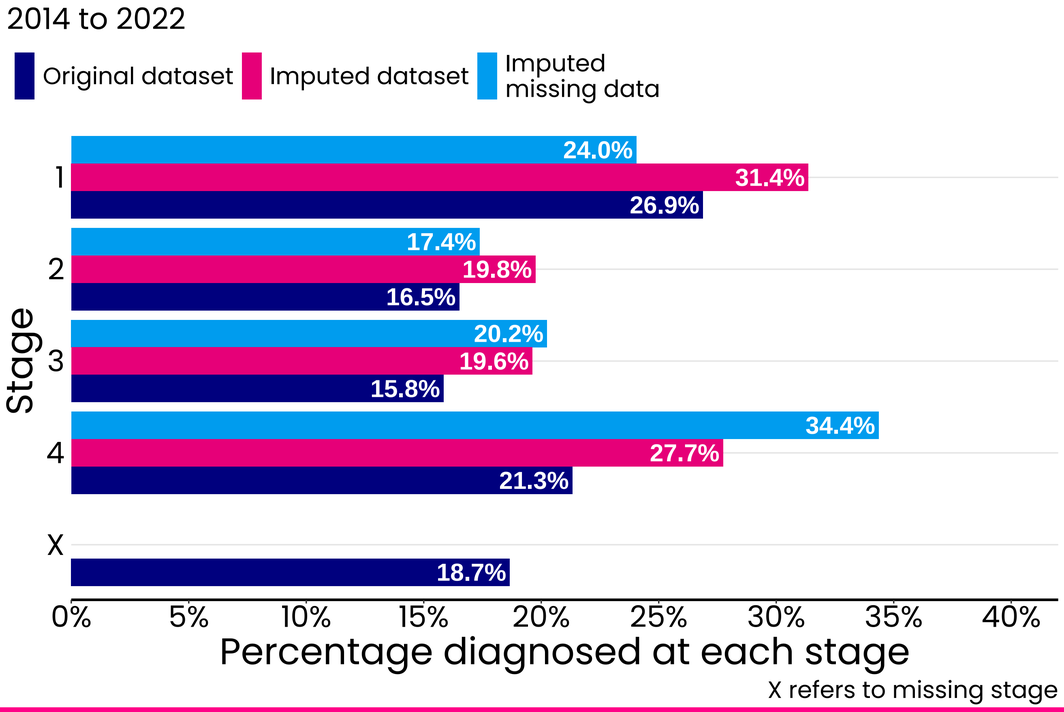
Figure 1: Late-stage (3 and 4) incidence rate over time, with and without imputation of missing stage (NCRD)



Around 19% of the original dataset had missing stage values (**Figure 2**). After stage imputation, stage 4 increased by 6.4 percentage points and stage 1 increased by 4.5 percentage points. This is because stage 4 was more common (34%) than stage 1 disease (24%) in the imputed missing stage patients.

Sensitivity analysis showed models with and without treatment imputed a similar stage distribution to the complete case known stage dataset. Average weighted Cohen's kappa values showed there was higher agreement when treatment was included in the model (0.42), compared to the model without treatment (0.30).

Figure 2: Stage distribution of original dataset, imputed dataset, and the subset of missing data post-imputation



References

1. National Disease Registration Service, NHS England. Staging data in England. 2025. Available: https://nhsd-ndrs.shinyapps.io/staging_data_in_england/
2. National Disease Registration Service, NHS England. Linking treatment data for cancer diagnoses CAS SOP 4v10. 2025. Available: https://digital.nhs.uk/ndrs/data/data-outputs/cancer-data-hub/cancer-treatments
3. Keogh R, Morris T. Multiple imputation in Cox regression when there are time-varying effects of covariates. Statistics in Medicine. 2018;37(15):3661-3678.
4. Cohen J. Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960;20:37-46.
5. Rubin D. Multiple imputation for nonresponse in surveys. New York; Chichester: Wiley. 1987.