Use of routine healthcare data for the estimation of disease outcomes in locally advanced non-small cell lung cancer (LA NSCLC)

Swee-Ling Wong, Kate Ricketts, Gary Royle, Matthew Williams, Ruheena Mendes.

University College London Hospital

Introduction

Outcomes for patients in the UK with LA NSCLC are amongst the lowest in Europe with 5-year survival of around 10% compared to up to 20% in other countries[1,2].

Progression free survival (PFS) and overall survival (OS) are key outcome measures for lung cancer. Assessing these outcomes is important for analysing the effectiveness of current trends in practice.

Aim

This project will investigate the use of routine healthcare datasets to determine PFS and OS of patients treated with radical radiotherapy for LA NSCLC.

Method

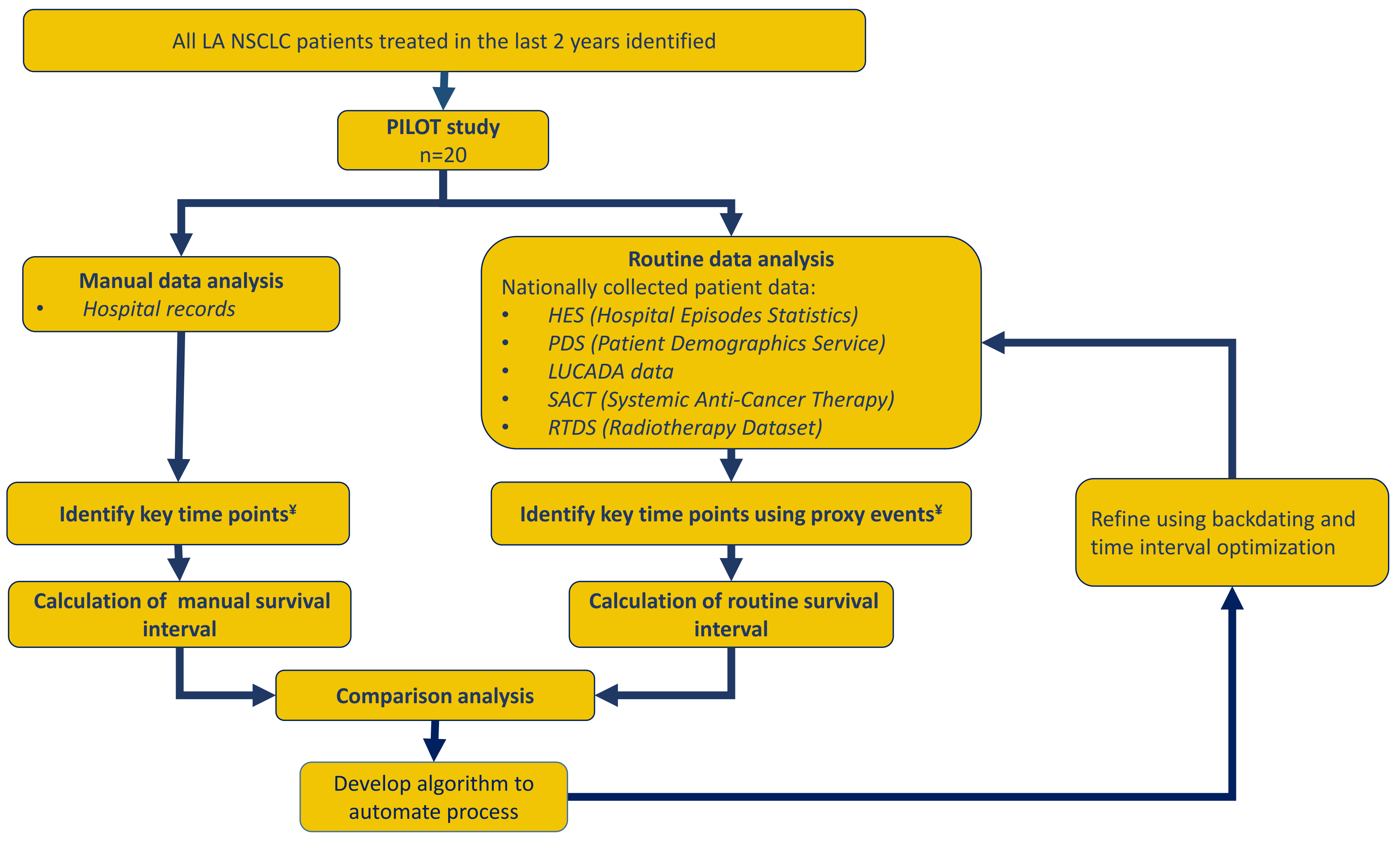

All LA NSCLC patients treated in the last 2 years identified
PILOT study
n=20
Manual data analysis
• Hospital records
Routine data analysis
Nationally collected patient data:
• HES (Hospital Episodes Statistics)
• PDS (Patient Demographics Service)
• LUCADA data
• SACT (Systemic Anti-Cancer Therapy)
• RTDS (Radiotherapy Dataset)
Identify key time points¥
Identify key time points using proxy events¥
Refine using backdating and time interval optimization
Calculation of manual survival interval
Calculation of routine survival interval
Comparison analysis
Develop algorithm to automate process

Figure 1. Flow chart showing the process of manual and routine data analysis for patients with LA NSCLC. Relevant time points are identified (¥ refer to figure 2) and used to calculate PFS and OS intervals for the data sets which are then compared to assess agreement. An algorithm is then developed to automate this process using backdating and time interval optimisation to refine the process.

Extraction of Proxy Time Points

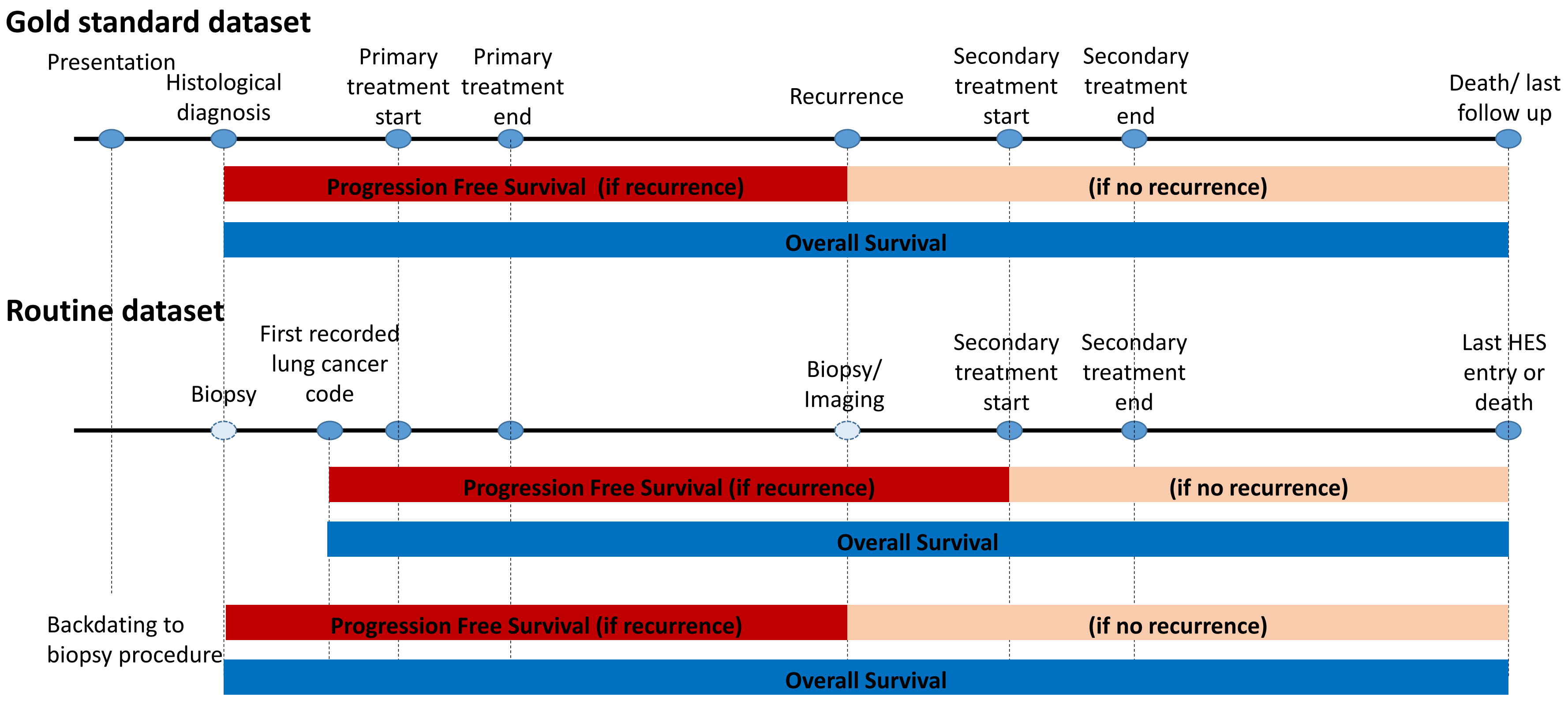


Figure 2. Diagram showing the clinical events defining PFS and OS from manual data and the identifiable proxy events used as a surrogate from routine datasets.[3]

Computational Algorithm Development

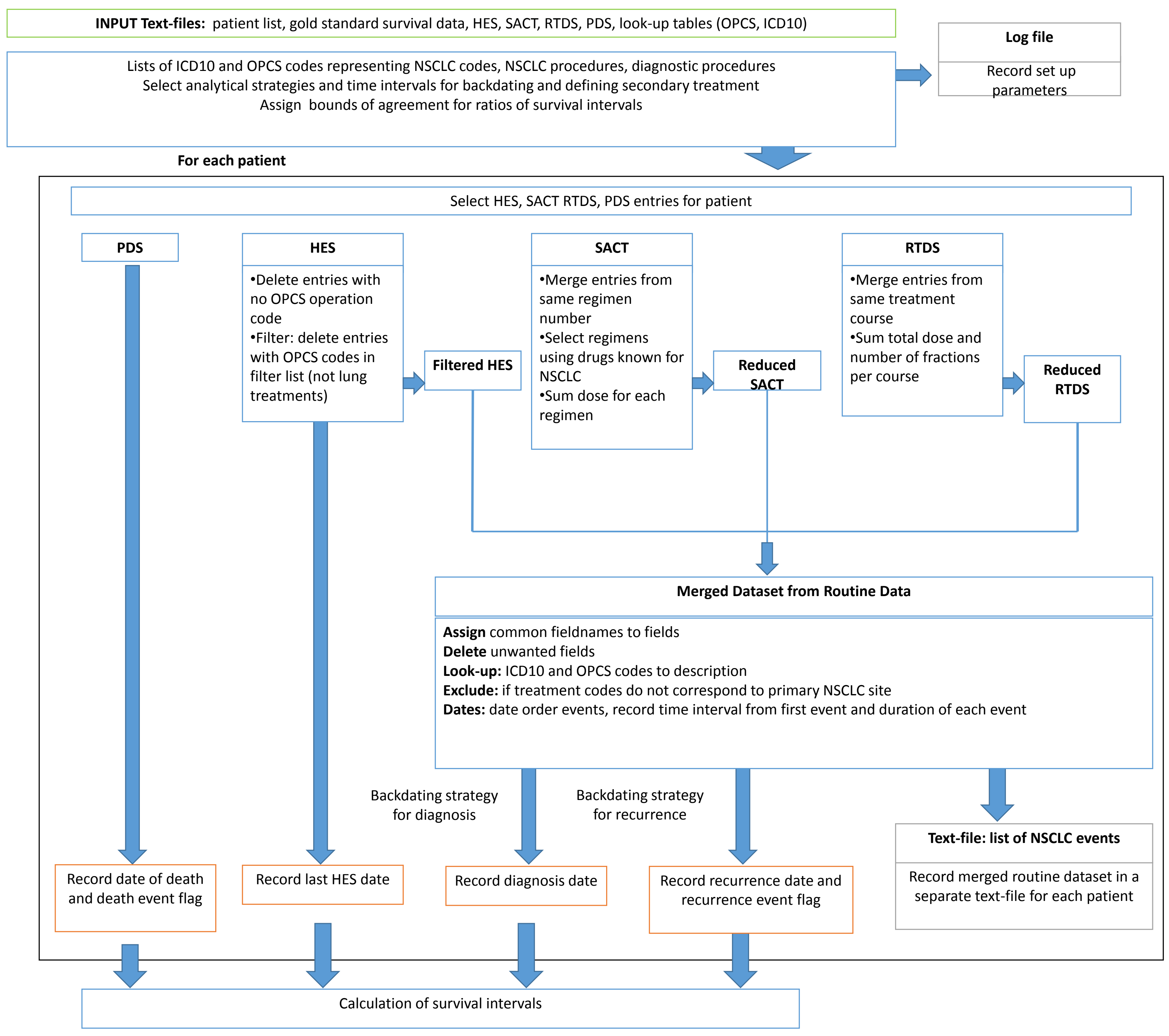

INPUT Text-files: patient list, gold standard survival data, HES, SACT, RTDS, PDS, look-up tables (OPCS, ICD10)
Lists of ICD10 and OPCS codes representing NSCLC codes, NSCLC procedures, diagnostic procedures
Select analytical strategies and time intervals for backdating and defining secondary treatment
Assign bounds of agreement for ratios of survival intervals
Log file
Record set up parameters
For each patient
Select HES, SACT RTDS, PDS entries for patient
PDS
HES
•Delete entries with no OPCS operation code
•Filter: delete entries with OPCS codes in filter list (not lung treatments)
Filtered HES
SACT
•Merge entries from same regimen number
•Select regimens using drugs known for NSCLC
•Sum dose for each regimen
Reduced SACT
RTDS
•Merge entries from same treatment course
•Sum total dose and number of fractions per course
Reduced RTDS
Merged Dataset from Routine Data
Assign common fieldnames to fields
Delete unwanted fields
Look-up: ICD10 and OPCS codes to description
Exclude: if treatment codes do not correspond to primary NSCLC site
Dates: date order events, record time interval from first event and duration of each event
Backdating strategy for diagnosis
Backdating strategy for recurrence
Text-file: list of NSCLC events
Record merged routine dataset in a separate text-file for each patient
Record date of death and death event flag
Record last HES date
Record diagnosis date
Record recurrence date and recurrence event flag
Calculation of survival intervals

Figure 3. Flow chart showing the algorithm process of merging the various routine datasets to identify dates of diagnosis, recurrence and death or last follow-up appointment and then calculation of PFS and OS.

Motivations

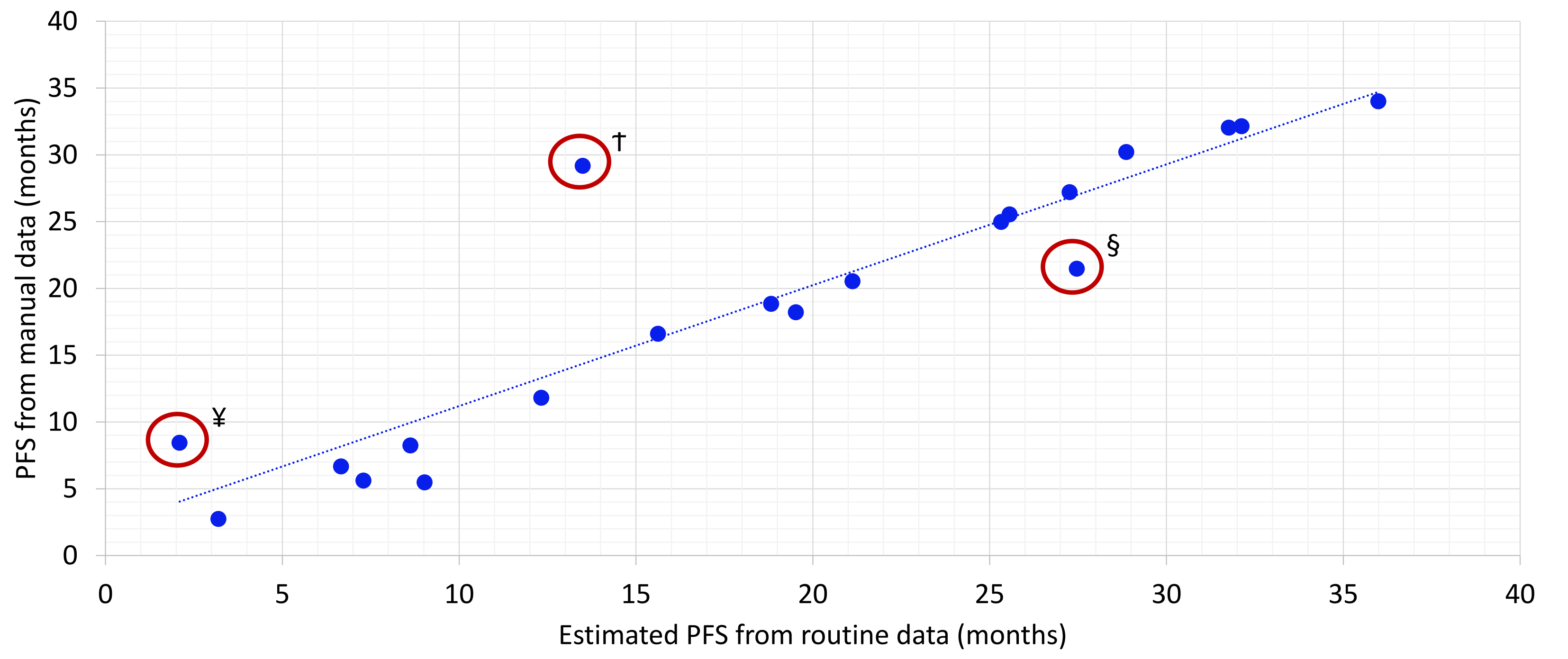
- "A Vision for Radiotherapy 2014- 2024"- national strategy to evaluate radiotherapy services
- Radiotherapy plays an important role in the treatment of patients with LA NSCLC so developing an algorithm that rapidly analyses outcomes of these patients is valuable for research and strategic planning of service provision.

Challenges to assessing outcomes

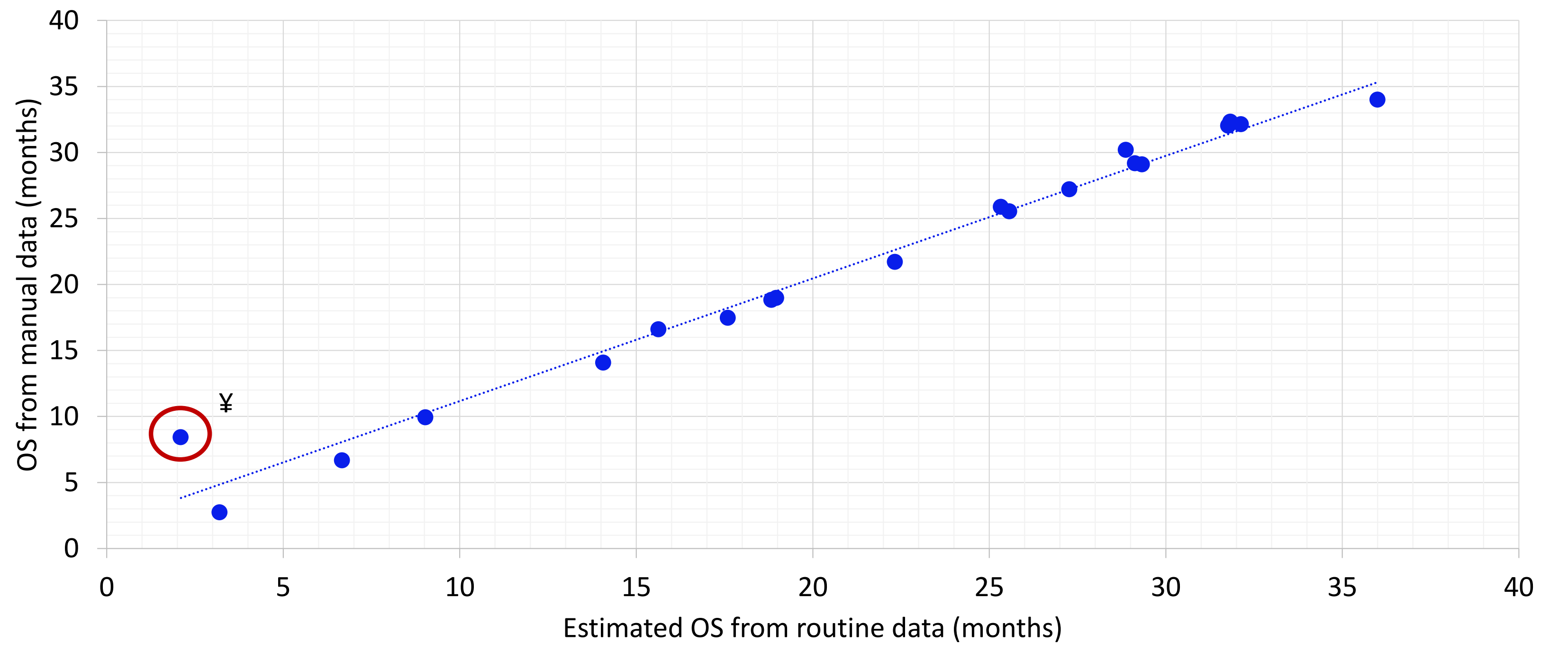
- Currently reliant on the quality, completeness and consistency of data from hospital records (manual data)
- **Manual data**
- Advantages: most accurately identifies clinically significant events; considered the Gold Standard
- Disadvantages: data quality can be inconsistent; collecting and analysing data is labour-intensive
- **Routine data**
- Advantages: Nationally collected data is an alternative source for data analyses
- Disadvantages: Does not include a diagnosis or recurrence date so proxy time points are used to extract this data from the routine dataset

Results

- 20 patients identified for the pilot study
- **Manual data:** median PFS=19.68m; median OS= 23.61m
- **Routine data:** median PFS= 19.18m; median OS=23.83m
- 15/20 and 11/20 routine diagnosis dates are within 4 weeks and 2 weeks of manual diagnosis dates, respectively; 5/20 diagnosis dates match exactly.
- 2 patients have $\geq$60day difference in routine and manual diagnosis dates: 1 patient had repeated negative biopsies prior to a positive diagnosis; 1 patient was referred from a peripheral hospital where diagnosis and chemotherapy had been initiated.
- 8 recurrences detected by algorithm. 2/8 recurrences not detected by algorithm as those patients did not receive secondary treatment. 1 incorrect recurrence event detected by algorithm due to unconventional use of systemic treatment.



Graph 1. Correlation between manual and routine derived PFS intervals determined by algorithm. Pearson correlation coefficient of 0.916. ¥ Patient diagnosed and chemotherapy initiated in peripheral hospital resulting in routine diagnosis date being later with a subsequently shorter PFS interval. † Unconventional use of systemic treatment resulting in algorithm incorrectly identifying a recurrence event. § Algorithm identifies a later date as the recurrence date from routine data resulting in longer PFS interval.



Graph 2. Correlation between manual and routine derived OS intervals determined by algorithm. Pearson correlation coefficient of 0.990. ¥ Patient diagnosed and chemotherapy initiated in a peripheral hospital resulting in routine diagnosis date being later with a subsequently shorter OS interval.

Limitations

- Quality of routine data eg. missing ICD10 codes means that the earliest lung cancer ICD10 identified by algorithm is not actually the first known instance; missing OPCS codes means algorithm cannot backdate to a diagnostic event.
- Routine data is incomplete for patients referred from other hospitals.
- Recurrence is not reliably detected if no secondary treatment is delivered.

Conclusions

- This is a novel approach to use routine datasets to determine outcome indicators in patients with LA NSCLC that will be a surrogate to analysing manual data.
- An algorithm has been developed to enable automated interpretation of routine datasets for patients with LA NSCLC and is being refined to improve data correlation.
- This method can be adjusted to auto-analyse outcomes for other stages of NSCLC.
- The ability to enable efficient and large scale analysis of current lung cancer strategies has a huge potential impact on the healthcare system.

References

1. Verdecchia A, Francisci S, Brenner H, et al. Recent cancer survival in Europe: a 2000-02 period analysis of EUROCARE-4 data. *Lancet Oncol.* 2007. 8:784e96

2. Coleman MP, Forman D, Bryant H, Butler J, Rachet B, Maringe C, Nur U, Tracey E, Coory M, Hatcher J, McGahan CE, Turner D, Marrett L, Gjerstorff ML, Johannesen TB, Adolfsson J, Lambe M, Lawrence G, Meechan D, Morris EJ, Middleton R, Steward J, Richards MA, ICBP Module 1 Working Group (2011). 'Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995–2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data'. *The Lancet*. 2011. 377: (9760) 127–38.

3. Ricketts, K., Williams, M., Liud, ZW., and Gibson, A. Automated estimation of disease recurrence in head and neck cancer using routine healthcare data. *Computer methods and programs in biomedicine*. 2014. 117: 412- 424