A study protocol for the validation of a primary care-based data-driven algorithm to predict pancreatic cancer in the UK setting: challenges of open research using routine healthcare data

Claire A. Price[1], Debbie Cooke[1], Martyn Winn[2], Nadia A. Smith[3], Agnieszka Lemanska[1]

[1] School of Health Sciences, University of Surrey; [2] Computational Biology, Science and Technologies Facility Council; [3] Data Science, National Physical Laboratory

Aim

To validate and enrich a data-driven algorithm for the early detection of pancreatic cancer in the UK setting

Method

A retrospective case-control study using nationally representative routine primary care data

Open Research

Practices: **Open access code & publication**
Protocol pre-registration
Challenges: **Data cannot be shared**

Background
Pancreatic Cancer

Overall cancer survival has increased over the recent decades, but the dismal 3-15%[1–3] survival rates of pancreatic cancer have not changed in the last 40 years.[4] This high mortality is largely due to most cases (80%) being diagnosed at a late stage when it has spread to other organs and is no longer curable.[5,6]

Detecting and diagnosing pancreatic cancer earlier is challenging as it often presents with non-specific symptoms. However, these symptoms are often recorded in routine primary care data.

Background
Routine Primary Care Data

Routine primary care data are a massive resource that is increasingly being employed to answer research questions. However, such data are currently underutilised due to restricted access as well as privacy and ethical implications. This project will serve as a case study to support the safe and trustworthy use of patient data in research. Validating the use of primary record care, which is not being collected for research purposes, is vital for providing evidence of the real-world utility of the approach.

Methods
Data Driven Algorithm

Data-driven approaches, including predictive algorithms that use a combination of symptoms have been developed to aid earlier detection and diagnosis.

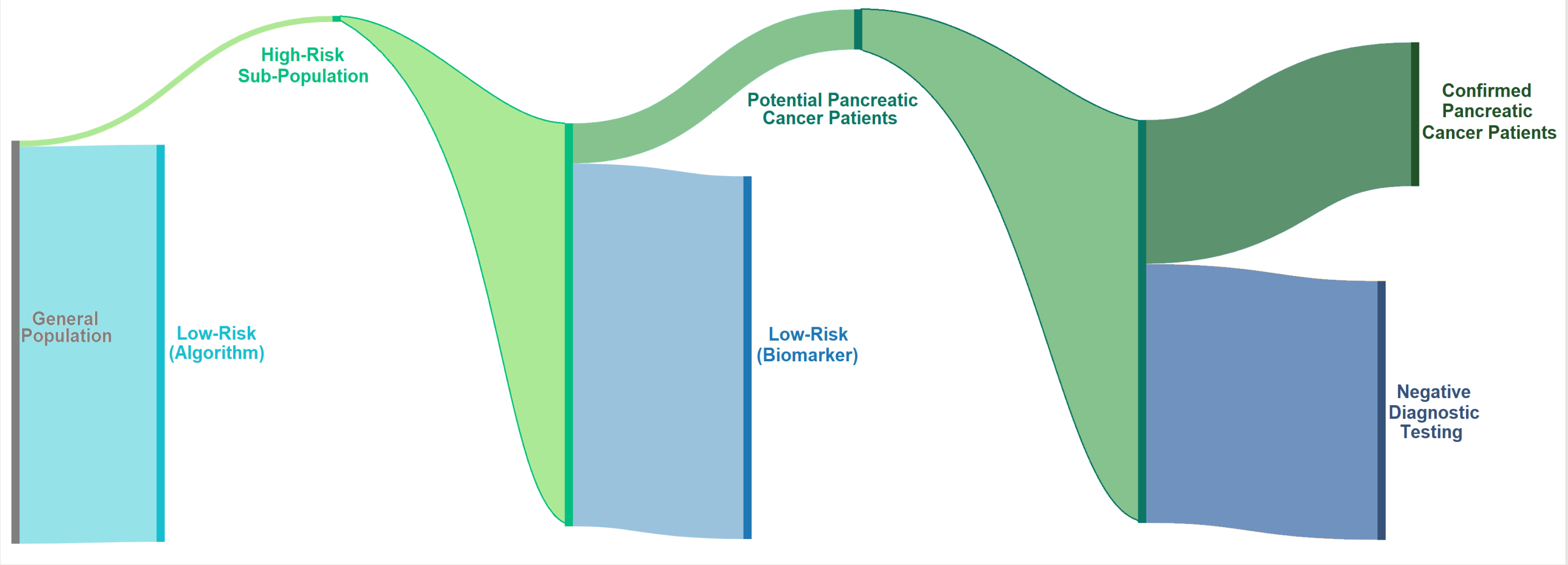
One such algorithm which flags patients at high risk of pancreatic cancer is ENDPAC (Enriching New-Onset Diabetes for Pancreatic Cancer)[7] which was developed in the US primary care setting. The aim of this project is to validate ENDPAC for the UK, using real world primary care data. This algorithm uses the age of diabetes onset, weight loss and changes in blood glucose to predict pancreatic cancer risk. These factors have been linked with an increased risk of pancreatic cancer.[8,9]

Methods
Validating ENDPAC

A retrospective case-control study using the nationally representative Oxford-Royal College of General Practitioners Clinical Informatics Digital Hub (ORCHID) database[10] will be undertaken. ORCHID holds over 10 million primary care electronic healthcare records including nearly 11,000 people diagnosed with pancreatic cancer. We will use this data to

- Determine positive and negative predictive values, sensitivity and specificity of ENDPAC
- Establish the cut-off for a high-risk pancreatic cancer cases using the Youden index
- Measure ENDPAC's discrimination with ROC and AUC

Screening Stages


High-Risk Sub-Population
General Population
Low-Risk (Algorithm)
Potential Pancreatic Cancer Patients
Low-Risk (Biomarker)
Confirmed Pancreatic Cancer Patients
Negative Diagnostic Testing

We propose using the enriched ENDPAC algorithm as part of a multi-stage screening process. This starts with identifying a high-risk sub-population of people using the enriched ENDPAC algorithm, followed by a second "sieve" of further biomarker testing. These potential pancreatic cancer patients would then undergo CT for diagnostic testing. Our aim is to improve early detection of pancreatic cancer without increasing unnecessary diagnostic CT screening.

Methods
Enriching ENDPAC

The ENDPAC algorithm categorises a patients' risk of pancreatic cancer based on a designated cut-off. Patients with a score above that cut-off require follow-up testing to confirm the pancreatic cancer diagnosis. However, due to the rarity of pancreatic cancer even a test with 99% sensitivity and specificity only has a positive predictive value of 3.6%, resulting in too many false positives to justify invasive screening tests.[6,11]

Therefore, we will enrich this algorithm, by finding additional differences between matched controls and pancreatic cancer patients. In addition, we hope to widen the algorithm to detect pancreatic cancer patients who do not develop diabetes.

Open Research
Challenges

Transparency and reproducibility are challenging when using primary care data as it is personal and highly sensitive. Datasets that consist of healthcare records cannot be made open access. Only authorised and trained researchers can access this type of data.

Open Research Practices

Pre-registration of the study protocol, enabling peer-review of methods

Results will be published open access in peer reviewed journals

The software developed in this project will be deposited in repositories such as GitHub to enable scrutiny and reuse

claire.a.price@surrey.ac.uk
@UniClairePrice

Funding: This project is jointly funded by the University of Surrey and National Physical Laboratory as part of the EPSRC iCASE 2021-22 programme

References

1. Rahib, L. et al. (2014) 'Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States'. *Cancer Res.,* 74(11), pp. 2913–2921.
2. Arnold, M. et al. (2019) 'Progress in cancer survival, mortality, and incidence in seven high-income countries 1995-2014 (ICBP SURVMARK-2): a population-based study'. *Lancet Oncol.,* 20(11), pp. 1493–1505.
3. Rachet, B. et al. (2009) 'Population-based cancer survival trends in England and Wales up to 2007: an assessment of the NHS cancer plan for England'. Lancet Oncol., 10(4), pp. 351–369.
4. Anon (2015) 'Pancreatic cancer survival statistics'. Cancer Research UK. *https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/pancreatic-cancer/survival* (Accessed 21 February 2022)
5. Pereira, S.P. et al. (2020) 'Early detection of pancreatic cancer'. *Lancet Gastroenterol. Hepatol.,* 5(7), pp. 698–710.
6. Pannala, R. et al. (2009) 'New-onset diabetes: a potential clue to the early diagnosis of pancreatic cancer'. *Lancet Oncol.*, 10(1), pp. 88–95.
7. Sharma, A. et al. (2018) 'Model to Determine Risk of Pancreatic Cancer in Patients With New-Onset Diabetes'. *Gastroenterol.*, 155(3), pp. 730–739.e3.
8. Chari, S.T. (2007) 'Detecting early pancreatic cancer: problems and prospects'. *Semin. Oncol.,* 34(4), pp. 284–294.
9. Damiano, J. et al. (2004) 'Should pancreas imaging be recommended in patients over 50 years when diabetes is discovered because of acute symptoms?' *Diabetes Metab.,* 30(2), pp. 203–207.
10. de Lusignan, S. et al. (2020) 'The Oxford Royal College of General Practitioners Clinical Informatics Digital Hub: Protocol to Develop Extended COVID-19 Surveillance and Trial Platforms'. *JMIR Public Health Surveill.,* 6(3), p. e19773.
11. Khan, S. et al (2021) 'Validation of the ENDPAC model: Identifying new-onset diabetics at risk of pancreatic cancer'. *Pancreatology,* 21(3), pp. 550–555.