UK Validation of ENDPAC

External UK validation of the ENDPAC model to predict pancreatic cancer risk: A registered report protocol

Authors
Claire A. Price1,2 *, Hugh Claridge1,2, Simon de Lusignan3, Natalia Khalaf4, Freda Mold1, Nadia A. S. Smith5,6, Martyn Winn7, Agnieszka Lemanska1,2
1University of Surrey, School of Health Sciences, Kate Granger Building, 30 Priestley Road, Surrey Research Park, Guildford, Surrey, GU2 7YH, UK
2National Physical Laboratory, Hampton Road, Teddington, Middlesex, TW11 0LW, UK
3Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, OX2 6GG, UK
4Center for Innovations in Quality, Effectiveness, and Safety (IQuEST), Michael E. DeBakey Veterans Affairs Medical Center, Baylor College of Medicine Department of Medicine, Houston, TX, USA
5Royal Surrey NHS Foundation Trust, Scientific Computing, Egerton Road, Guildford, GU2 7XX, UK
6TÜV SÜD UK, Chadwick House, Warrington Road, Birchwood Park, Warrington, WA3 6AE, UK
7Scientific Computing Department, Science and Technology Facilities Council, Research Complex at Harwell, Didcot OX11 0FA, UK
*claire.a.price@surrey.ac.uk

Key words
Pancreatic cancer; Routine healthcare data; Early diagnosis; Primary health care; Risk algorithm

How this fits in
Pancreatic cancer is a devastating disease which is hard to diagnose. An algorithm called ENDPAC has been developed in the United States to help clinicians identify people at risk of having undiagnosed pancreatic cancer. These people can be referred for an imaging investigation to diagnose or rule out cancer. This protocol outlines a United Kingdom (UK) validation of ENDPAC so that it could be used in clinical practice in the UK.
UK Validation of ENDPAC

Abstract

Introduction
Overall cancer survival has increased over recent decades, but the very low survival rates of pancreatic cancer have hardly changed in the last 50 years. This is attributed to late diagnosis. Pancreatic cancer symptoms are non-specific which makes early diagnosis challenging. Data-driven approaches, including algorithms using combinations of symptoms to predict cancer risk, can aid clinicians. A simple but effective algorithm called Enriching New-Onset Diabetes for Pancreatic Cancer (ENDPAC) has been developed in the United States (US). ENDPAC has not yet been used in the United Kingdom (UK), our aim is to translate ENDPAC into the UK setting. The objectives are to validate ENDPAC and report its predictive utility within primary care.

Methods
A retrospective cohort study of people with new-onset diabetes using the nationally representative Oxford-Royal College of General Practitioners Clinical Informatics Digital Hub (ORCHID) database. ORCHID holds over 10 million primary care electronic healthcare records. ENDPAC scores will be calculated for eligible people along with positive predictive value, negative predictive value, sensitivity and specificity of the algorithm. We will evaluate the optimal cut-off for defining people with high-risk of having pancreatic cancer.

Discussion
Once validated within the UK, ENDPAC could be implemented in practice to improve early pancreatic cancer diagnosis by using routine data. ENDPAC is currently being tested in the US in a clinical trial to evaluate its effectiveness. ENDPAC offers an automatable and inexpensive way to improve early diagnosis as part of a sequential approach to identify individuals at high-risk of having undiagnosed pancreatic cancer.
UK Validation of ENDPAC

Introduction

Pancreatic cancer is a devastating disease due to its high mortality rate with a very low 5-year survival rate of 3-15% [1–6]. It is the 14th most common cancer globally, accounting for 3% of all new cancer cases, with nearly half a million cases worldwide [7]. The high mortality of pancreatic cancer means that it is the fourth major cause of cancer mortality in the world [8].

The need for prompt diagnosis and early detection of pancreatic cancer is widely recognised and is the most likely way in which the current dismal survival rates will improve [9]. However, there are currently no screening or biomarker tests for pancreatic cancer and diagnosis depends upon presentation with symptoms to a General Practitioner (GP). This is challenging as a full-time GP will only see one case every five years on average [10] especially as pancreatic cancer often presents with non-specific symptoms. Symptoms also vary between patients, making diagnosis challenging for clinicians [10–12]. Therefore, it is vital to provide GPs with innovative and practical ways to facilitate early detection. This includes data-driven approaches to flag people who are at high risk of having pancreatic cancer for referrals for further investigation.

A simple but promising data-driven algorithm for the earlier detection of pancreatic cancer has been developed in the United States (US). The Enriching New-onset Diabetes for Pancreatic Cancer (ENDPAC) algorithm [13] uses changes in weight and blood glucose to calculate someone’s risk of having pancreatic cancer. ENDPAC is applied to people aged 50 years or older with new-onset diabetes, as this is the cohort with the highest risk of developing non-familial or sporadic pancreatic cancer [14,15]. This clear clinical rationale means ENDPAC is easily explainable to both GPs and patients. Its relative simplicity and use of measures routinely collected in primary care make ENDPAC well suited for implementation in the UK primary care setting.

One significant advantage of ENDPAC is that it targets early detection as it is based on glucose and weight changes. In contrast, other data-driven risk assessment algorithms [11,16–23] incorporate symptoms like back pain and jaundice, which typically manifest in the later stages of pancreatic cancer. Given that pancreatic cancer is often diagnosed too late for effective treatment, leading to high mortality rates, the ability of ENDPAC to identify cases earlier in the disease progression could potentially save lives. Furthermore, ENDPAC stands out for its use of a relatively small number of routinely collected measurements, addressing concerns about data availability that arise with other algorithms.

Data-driven algorithms, such as Risk Assessment Tools (RATs) like QCancer [24–26] are showing promise in aiding clinical decisions and expediting referrals. However, before ENDPAC can be added to these tools in clinical practice, it requires validation in the UK. ENDPAC was developed in the US and validation in the UK is necessary due to significant differences between the healthcare systems of the two countries [27–30]. Furthermore, previous validations of ENDPAC [21,31,32] have also included secondary care data. Additionally, adjustments to the ENDPAC algorithm are necessary for its application in UK primary care due to how blood glucose is recorded (ENDPAC was developed using fasting blood glucose but in UK primary care, HbA1c (haemoglobin A1c) is recorded and used to diagnose diabetes). Therefore, validation specific to primary care settings in the UK is essential to ensure the algorithm’s suitability for such data.

The aim is to validate the ENDPAC algorithm using UK primary care data. We will extract the cohort of people with new-onset diabetes and their information on pancreatic cancer diagnosis. We will calculate the ENDPAC score at the time of diabetes onset (defined by one HbA1c measure ≥ 48 mmol/mol) reporting predictive value, sensitivity, and specificity of ENDPAC.
UK Validation of ENDPAC

Methods

Study design and data source
This is a retrospective cohort study of adults with new-onset diabetes between 2006 and 2017. They were followed up for 3 years after their first elevated HbA1c reading (defined as diabetes diagnosis) to determine their pancreatic cancer diagnosis status. We used the 2006 cut-off because coding of the metabolic markers such as weight and HbA1c in primary care improved significantly from 2004 with the introduction of the Quality and Outcomes Framework and the pay-for-performance incentive scheme [33,34]. Therefore, the 2006 cut-off was set to improve data completeness for at least 2 years before the index date (defined below).

Data from the Oxford-Royal College of General Practitioners Clinical Informatics Digital Hub (ORCHID) database will be used [35]. This is a nationally representative database [36] which downloads electronic healthcare records from GP practices which belong to the Royal College of General Practitioners, Research and Surveillance Centre. In May 2021, using the systematized nomenclature of medicine clinical terminology (SNOMED CT) system [37], we extracted a dataset that included demographic information, ages at diagnoses of diabetes and pancreatic cancer, pancreatic cancer-related symptoms (change in bowel habit, constipation, diarrhoea, abdominal pain, back pain, nausea, vomiting, abdomen scan, operation on pancreas, jaundice, weight loss, indigestion, digestive disorders, suspected pancreatic cancer or cancer referral) as well as all measurements of weight and HbA1c recorded in 2001 or after.

Diabetes Status

ENDPAC requires two paired elevated glycaemic values, within 90 days, to diagnose diabetes. However, this requirement is challenging in real-world settings in which people will not necessarily have a second test within 90 days. Therefore, in this study, diabetes diagnosis is defined by one HbA1c measure of ≥48mmol/mol and encompasses new-onset biochemically-defined hyperglycaemia and diabetes (glycaemic onset). This is the index date.

In the UK HbA1c is measured in mmol/mol (IFCC) [38]. Therefore, HbA1c will be extracted in mmol/mol.

ENDPAC

The ENDPAC model [13] was developed for people who are over 50 and have developed new-onset diabetes. It combines their age at diabetes diagnosis with changes in their blood glucose and weight over the year before their diabetes diagnosis to produce a score for their risk of having pancreatic cancer. This score consists of three sub-scores: A (Δ blood glucose category) + B (Δ weight category) + C (age category at diabetes diagnosis), see Table 1. The final ENDPAC score produces one of three outcomes: low (<0), intermediate or high (≥3) risk of having pancreatic cancer.

<table>
<thead>
<tr>
<th>Fasting Blood Glucose (FBG) categories</th>
<th>HbA1c Blood Glucose (HbA1c BG) categories</th>
<th>Δ BG Category score (NOD – 1y): A</th>
<th>Δ Weight category</th>
<th>Δ Weight score: B</th>
</tr>
</thead>
<tbody>
<tr>
<td>BG range (mg/dl)</td>
<td>BG range (HbA1c mmol/mol)</td>
<td>Score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BG category at ~ 1 years</td>
<td></td>
<td>Score range</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 100</td>
<td>&lt; 42</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 100 to &lt; 110</td>
<td>≥ 42 to &lt; 44</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 110 to &lt; 126</td>
<td>≥ 44 to &lt; 48</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BG category at glycaemically-defined new-onset diabetes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 126 to &lt; 160</td>
<td>≥ 48 to &lt; 60</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥ 160</td>
<td>≥ 60</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
UK Validation of ENDPAC

<table>
<thead>
<tr>
<th>( \Delta ) Weight (kg)</th>
<th>Score</th>
<th>Score range</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \leq -6.0 )</td>
<td>+6</td>
<td></td>
</tr>
<tr>
<td>( &gt; -6.0 ) to ( -4.0 )</td>
<td>+4</td>
<td></td>
</tr>
<tr>
<td>( &gt; -4.0 ) to ( -2.0 )</td>
<td>+2</td>
<td>-6 to +6</td>
</tr>
<tr>
<td>( &gt; -2.0 ) to ( +2.0 )</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>( \geq +2.0 ) to ( +4.0 )</td>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>( \geq +4.0 ) to ( +6.0 )</td>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>( \geq +6.0 )</td>
<td>-6</td>
<td></td>
</tr>
</tbody>
</table>

<p>| Age at glycaemically-defined new-onset diabetes | Age score: C |</p>
<table>
<thead>
<tr>
<th>Age Range</th>
<th>Score</th>
<th>Score range</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 60</td>
<td>-1</td>
<td>-1 to +1</td>
</tr>
<tr>
<td>( \geq 60 ) to ( &lt; 70 )</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>( \geq 70 )</td>
<td>+1</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: ENDPAC score parameters with blood glucose (HbA1c) categories equivalent to fasting blood glucose categories presented, based on American Diabetes Association blood glucose levels for healthy, prediabetes, diabetes and uncontrolled diabetes. NOD: New-Onset Diabetes, \( \Delta \) represents the change in the measurement. Table adapted from Sharma et al. [13]

Patient Population

Study participants will be aged 50 years or older. They will have a diagnosis of new-onset diabetes between 2006 and 2017. The date of diagnosis was defined as the date of the first elevated blood glucose reading (HbA1c \( \geq 48 \text{ mmol/mol} \)) and defined as the index date. Eligible patients were required to have:

1. An HbA1c \( \geq 48 \text{ mmol/mol} \) reading preceded by at least one HbA1c <48 mmol/mol reading in the past 3-24 months
2. No HbA1c \( \geq 48 \text{ mmol/mol} \) reading before the index date
3. No history of pancreatic cancer before the index date
4. Weight on the index date (can be \( \pm 30 \) days)
5. Weight 1 year prior to the index date (between 24 months prior and 6 months prior)

This definition of new-onset diabetes was adapted from the study by Sharma et al. [13]. If multiple readings for weight or HbA1c are available the flowchart by Sharma et al. [36] will be followed to select which readings are used. Sensitivity analysis will then be conducted to investigate the impact of these decisions and whether the predictions change when using highest, lowest or mean readings (when more than one reading is available). Sensitivity analysis will also be conducted on the effect of requiring people to have a second HbA1c reading within 90 days of the first reading as a definition of new-onset diabetes.

Statistical analysis

Descriptive Statistics

Baseline demographic statistics (gender, ethnicity, region and index of multiple deprivation (IMD) quintile) will be extracted.

Predictors

The study will include all three risk factors that are part of the ENDPAC model. These are: age at index date, change in body weight (kg), and change in blood glucose (HbA1c).

Outcome identification

The outcome is the diagnosis of pancreatic cancer within three years after the index date which is the period of greatest risk of developing pancreatic cancer [14,15,39–41]. Pancreatic cancer
UK Validation of ENDPAC

diagnosis is based on SNOMED CT system codes, using code 363418001 Malignant tumor of pancreas (disorder) and all sub-codes (excluding 94164003, 94212002, 94325008, 94354002, 94459006, 9460001, 944618007, 285614004) [37]. Excluded codes relate to cancers that have metastasized to the pancreas.

Missing data
This study is a validation of the ENDPAC model so ENDPAC scores will only be calculated for people who have the required information (complete cases). Therefore, there will be no missing data.

Model validation
The positive predictive value, negative predictive value, sensitivity, specificity as well as the percentage of the cohort to be tested further will be calculated for different cut-off values within the model.

Calibration (predicted probability is in line with observed outcomes) will be assessed using calibration plots of the slope of the observed proportion of events against the predicted risk. The model's discrimination (ability to distinguish between those with and without the pancreatic cancer outcome) will be assessed by area under the receiver operating characteristic curve.

The optimal cut-off value to identify people at high-risk of pancreatic cancer will then be found, using the Youden index, which maximises both the sensitivity and specificity of the model [42].

Subgroup analyses
Risk score stratification will be utilised to explore the differences in cancer rates between the risk score categories (low-, intermediate- and high-risk). The proportion of people who would be referred for further testing will also be calculated (overall and when data is stratified). Model sensitivity will also be investigated for different lead times (3-6 months, 6-9 months, 9-15 months, 15-18 months, 18-21 months and 21-24 months) of the measurements utilised. The effect of using a single elevated HbA1c reading as the definition of diabetes instead of paired values will also be investigated.

Software and reproducibility
The database is managed in Structured Query Language (SQL) Server Management Studio version v18.9.1. Data analyses will be conducted using R Studio version 2023.03.0+386 "Cherry Blossom". Software will be published open access via GitHub for reuse and review.

Ethical Implications
The study was approved by the University of Surrey Ethics Committee (reference number: FHMS 21-22 269 EGA). Access to ORCHID data has been approved by Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC) under data request RSC_0420.

Data Availability
Data will remain under the control of the Oxford-Royal College of General Practitioners Clinical Informatics Digital Hub (ORCHID, orchid.phc.ox.ac.uk) and can be accessed following all necessary approvals.

Dissemination
Dissemination is planned through peer-reviewed publications and conference presentations.
UK Validation of ENDPAC

Acknowledgements
We thank patients and practices who are members of the ORCHID network. This project was funded as part of an EPSRC iCase studentship undertaken by CP. The work of NPL co-authors was funded by the UK Government’s Department for Science, Innovation & Technology through the UK’s National Measurement System programmes.

References
UK Validation of ENDPAC


UK Validation of ENDPAC


[38] HbA1c reporting methods set to change. Diabetes UK 2011.


