Title: The Challenges of Identifying Patients with Peripheral Artery Disease Utilising Administrative Databases

Short Title: Challenges of PAD in admin data

Authors: Pishoy Gouda MB BCh BAO, MSc\textsuperscript{1}, Douglas C. Dover PhD\textsuperscript{2}, Elinor (Mengxiao) Wang MSc\textsuperscript{2}, Victor Aboyans MD, PhD\textsuperscript{3,4}, Robert Welsh MD, MSc\textsuperscript{1,2}

Affiliations:

\textsuperscript{1} Mazankowski Alberta Heart Institute, University of Alberta, Division of Cardiology, Edmonton, Alberta, Canada

\textsuperscript{2} Canadian VIGOUR Centre, University of Alberta, Edmonton, Alberta, Canada

\textsuperscript{3} Department of Cardiology, Dupuytren-2 University Hospital, Limoges, France

\textsuperscript{4} EpiMaCT, Inserm 1094 & IRD 270, Limoges University, Limoges, France

All authors takes responsibility for all aspects of the reliability and freedom from bias of the data presented and their discussed interpretation

Word Count: 2638 words

Corresponding Author:

Robert C. Welsh, MD

Robert.Welsh@ahs.ca

Mazankowski Alberta Heart Institute

2C2 Cardiology, WMC

8440 112 St NW
Edmonton, Alberta, T6G 2B7

780-407-3613 [fax 780 407 3639]
Abstract:

Peripheral artery disease (PAD) carries a high burden of morbidity when identified in patients with coronary artery disease (CAD). However, the identification of patients with concomitant CAD and PAD remains challenging. Using linked administrative databases of 207,026 individuals with CAD between 2002 and 2019 (median follow-up 4.7 years), a model for PAD was applied to identify baseline PAD and the development of PAD during follow-up. Both baseline PAD and future PAD models demonstrated poor calibration and discrimination (c-statistic 0.618 and 0.583). In the absence of additional variables, the present models are unable to identify patients with concomitant CAD and PAD.
Introduction:

The prevalence of peripheral artery disease (PAD) is estimated to be 5.6%, affecting ~236 million individuals around the world and ~2 million Canadians. As the presence of even asymptomatic PAD is associated with a 3-fold increase in major cardiovascular events, this has led to renewed efforts to identify these patients earlier in their disease trajectory. This remains challenging due to lack of awareness of the public and the fact that up to 50% of patients with PAD are asymptomatic. While the ankle-brachial index (ABI) remains the non-invasive gold-standard to diagnose PAD, screening for asymptomatic PAD has not been recommended due to a lack of evidence to suggest that it improve outcomes or is cost-effective. Compared to the general population, patients with coronary artery disease (CAD) are twice as likely to have PAD. In addition, in the Cardiovascular Outcomes for People Using Anticoagulation Strategies (COMPASS) trial, patients with both CAD and PAD demonstrated the greatest benefit of the addition of low-dose rivaroxaban therapy to aspirin to reduce major adverse cardiovascular events (MACE). In the absence of a routine ABI screening program, alternative methods to identify patients with CAD who have concomitant asymptomatic PAD are required. One approach is to harness administrative data to provide contemporary data on the prevalence and outcomes of this high-risk population.

As administrative databases are reliant on coding of physician claims, it should not be surprising that previous studies have demonstrated that International Classification of Disease (ICD) coding has poor diagnostic characteristics and fails to identify asymptomatic PAD. Using an adapted version of a diagnostic model derived from
participants of the COMPASS trial, we aimed to ascertain its usefulness in an administrative database of patients with CAD.

**Methods:**

**Patient Population:**

Alberta with a population of approximately 4.5 million, has a single-payer, publicly funded health care system. Data for the study was obtained by linking the following administrative healthcare databases using the unique provincial personal health number: the Discharge Abstract Database of all admissions to acute care facilities containing a most responsible diagnosis and up to 24 other diagnoses and up to 20 procedure codes; the Ambulatory Care database of emergency visits and day procedures containing 10 diagnosis fields and 10 procedure code fields; Practitioner Claims of physician billings to the province containing up to 3 diagnosis codes; the Alberta Health Care Insurance Plan registry tracking Albertans eligible for health care in the province is used to determine residency and age; Alberta Vital Statistics – Capturing the date of death; Laboratory data indicating the test and test results. Using these linked administrative healthcare databases, we identified a retrospective cohort of adult patients with CAD in Alberta, Canada between 2002-2019 with lab values available for both serum creatinine and cholesterol (within 6 months of each other, data available starting 2012) defining the baseline date. The diagnoses of both CAD and PAD were determined by identifying relevant ICD codes from hospitalisation records, ambulatory care visits and practitioner claims. We required two instances to be identified in practitioner claims, at least 30-days apart within one year, or one instance from hospitalisation or ambulatory records. Codes for CAD

Diagnostic Models:
The models utilised in the study have previously been described, developed from participants of the COMPASS trial. Derived from the COMPASS trial, symptomatic PAD was defined as known history of PAD, history of claudication, pervious arterial bypass, previous amputation, PAD identified on imaging studies. Asymptomatic PAD was defined as an ABI of <0.9 in the absence of symptoms of PAD. The model components body mass index, systolic and diastolic blood pressure, heart rate, and tobacco use were not available in the administrative health data and omitted. The following variables were included: age, sex, serum creatinine, total cholesterol, history of myocardial infarction, diabetes, hypertension, transient ischemic attack, and stroke.

Statistical Analysis:
Model 1 was a logistic regression model predicting symptomatic lower extremity PAD and carotid disease so was validated against PAD status at baseline according to administrative health records. Model 2 was a logistic regression model predicting
asymptomatic lower extremity PAD in those without a PAD diagnosis at baseline so was validated against future diagnoses of PAD in the administrative data.

In model 1, Calibration-in-the-large, the assessment of observed to model predicted outcomes was assessed visually using a calibration plot and by allowing the intercept to vary, holding the predictor effects fixed, in a logistic regression on the outcome PAD at baseline. Slope calibration, the assessment of systematic over- or under-estimation, in Model 1 was assessed using a calibration plot and testing the slope coefficient equal to 1 in a logistic regression. A fully recalibrated model was assessed using a calibration plot and a logistic regression model including the original risk score as a fixed offset and allowing coefficients for all measures to vary, capturing the change in each coefficient. Discrimination in these models was assessed using the c-statistic.

Calibration-in-the-large, Slope Calibration, and fully re-calibrated models in Model 2 were assessed as above, with the outcome of future PAD diagnosis. To fully account for time, cumulative event rates shown as Kaplan-Meier curves of PAD incidence by decile of risk score are presented and tested using the log-rank test. Discrimination was assessed using the c-statistic and the time varying and integrated AUC statistic of Uno\(^8\). Models assessing future PAD exclude patients with PAD at baseline or deaths on the index date.

Predictive validation evaluates the ability of a risk scores to predict clinically relevant events beyond the original purpose of PAD prediction. The composite outcome of myocardial infarction, stroke and death was predicted using both models. Kaplan-Meier
curves stratifying the predicted values from the slope calibrated models into deciles were tested using the log-rank test.

This study was approved by the University of Alberta Research Ethics Board (Pro00082215). The ethics panel determined that the research is a retrospective database review for which subject consent for access to personally identifiable health information would not be reasonable, feasible, or practical. All statistical analyses were carried out using SAS v9.4.
Results:

We identified 207,026 individuals with CAD between 2002 and 2019 with both lab values present (serum creatinine and total cholesterol). At baseline, 8,572 (4.1%) had a diagnosis of PAD, and 7,371 (3.6%) developed PAD over a median follow-up of 4.7 years. The mean age of patients was 67.1 +/- 12.0 years and 35.8% were female (Table 1).

Model 1 for the presence of PAD at baseline had poor calibration (rejected a calibrated slope p<.0001) and had poor discrimination (c-statistic .618; Table 2). After fully re-estimating the model coefficients (Supplementary Table S1), the model displayed good calibration and fair discrimination (c-statistic 0.743). Model 2 for the development of future PAD demonstrated poor calibration (rejected a calibrated slope p<.0001) and poor discrimination (c-statistic 0.583;Table 3). After fully re-estimating the model coefficients (Supplementary Table S2), the model displayed good calibration and poor discrimination (c-statistic 0.618).

Overall, the composite outcome occurred in 9.6%, 19.6% and 37.0% of patients at one, three and seven years of follow-up. However, deciles of the PAD risk score were poor at discriminating between MACE risk with only the two largest risk score quintiles separating from the rest (Figure 1).

Conclusion:

The models evaluated in the present study utilised commonly available variables in administrative databases. These models were not able to identify patients with PAD or
those at risk of developing PAD, limiting the current capacity utilize these databases for active patient identification. Further work is required to understand why PAD is frequently underdiagnosed and how this can be improved in the clinical setting. One potential avenue is in the introduction of simple to use ABI measuring devices that can easily be used at the bedside by front line healthcare workers. The introduction of these devices into the market, opens the possibility for the quick and reliable diagnosis of PAD. It remains to be seen whether advances in machine learning algorithms can leverage structured and unstructured data in electronic medical records to accurately identify this underdiagnosed, under-treated and high-risk population.

Acknowledgement:

This study is based in part on data provided by Alberta Health and Alberta Health Services. We thank the Customer Relationship Management and Data Access Unit at Alberta Health for creating the linked database. The interpretation and conclusions are those of the researchers and do not represent the views of the Government of Alberta. Neither the Government of Alberta nor Alberta Health express any opinion in relation to this study.

Funding and Disclosures: All authors declare no conflicts of interests. This work was partially funded by the CCS-Bayer Cardiovascular Research Award which was awarded to PG.
<table>
<thead>
<tr>
<th>Table 1 – Patient Demographics Stratified by Presence of Baseline PAD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No PAD at baseline</strong></td>
</tr>
<tr>
<td>Total N</td>
</tr>
<tr>
<td>Age (years)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Age Group</td>
</tr>
<tr>
<td>&lt;60</td>
</tr>
<tr>
<td>60-69</td>
</tr>
<tr>
<td>70-79</td>
</tr>
<tr>
<td>80+</td>
</tr>
<tr>
<td>Sex - female</td>
</tr>
<tr>
<td>Serum Creatinine (mg/dl)</td>
</tr>
<tr>
<td>Cholesterol (umol/l)</td>
</tr>
<tr>
<td>MI</td>
</tr>
<tr>
<td>Diabetes</td>
</tr>
<tr>
<td>Hypertension</td>
</tr>
<tr>
<td>TIA</td>
</tr>
<tr>
<td>Stroke</td>
</tr>
<tr>
<td>Develop PAD</td>
</tr>
</tbody>
</table>

**Abbreviations:** MI – myocardial infarction; TIA – transient ischemic attack; PAD – peripheral arterial disease
Table 2. Calibration and Discrimination for Reduced Model 1 Predicting Symptomatic PAD

<table>
<thead>
<tr>
<th></th>
<th>Ideal (well calibrated)</th>
<th>External Validation Sample</th>
<th>p-value</th>
<th>c-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibration-in-the-large&lt;sup&gt;a&lt;/sup&gt;</td>
<td>--</td>
<td>-5.22 (intercept)</td>
<td>--</td>
<td>0.618</td>
</tr>
<tr>
<td>Calibration slope&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1</td>
<td>0.39 (slope)</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>Fully re-estimated</td>
<td>n/a</td>
<td>n/a</td>
<td>&lt;.0001</td>
<td>0.743</td>
</tr>
</tbody>
</table>

<sup>a</sup> intercept not reported, as such no formal test is possible
<sup>b</sup> testing slope equal to 1
Table 3. Calibration and Discrimination for Reduced Model 2 Predicting Future PAD

<table>
<thead>
<tr>
<th>Time to Event Models</th>
<th>Binary Outcome Models</th>
<th>Calibration-in-the-large&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Calibration slope&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Fully re-estimated</th>
<th>Integrated AUC&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ideal (well calibrated)</td>
<td>External Validation Sample</td>
<td>p-value</td>
<td>c-statistic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>--</td>
<td>-10.4 (intercept)</td>
<td>--</td>
<td>0.583</td>
<td>--</td>
</tr>
<tr>
<td>Calibration slope</td>
<td>1</td>
<td>0.057 (slope)</td>
<td>&lt;.0001</td>
<td>0.592</td>
<td>0.592</td>
</tr>
<tr>
<td>Fully re-estimated</td>
<td>n/a</td>
<td>n/a</td>
<td>&lt;.0001</td>
<td>0.618</td>
<td>0.428</td>
</tr>
</tbody>
</table>

<sup>a</sup> Intercepts not reported as such no formal test is possible
<sup>b</sup> Testing slope equal to 1
<sup>c</sup> For both models, the AUC was virtually constant over time. The time-dependant AUC in the calibration slope model ranged from 0.587 to .603 and in the fully re-estimated model ranged from 0.422 to .432.
Figure 1 Legend - Kaplan Meier Curve of Major Adverse Cardiovascular Events

Stratified by Deciles of the Slope Calibrated Model

Abbreviations: MACE – major adverse cardiovascular events (Composite of death, myocardial infarction and stroke)
References:


Using Slope Calibrated Reduced Model 1 from Gouda et al 2021