PiCture

Chronic Pain IdentifiCation
Through Using electRonic rEcords

Data Analysis Plan (DAP)

Draft 7th Feb 2024

STUDY FULL TITLE	Chronic Pain Identification Through Using Electronic Records (C-PICTURE). Development and validation of an algorithm to identify people with chronic pain through primary care-based records
DAP VERSION	0.2
DAP VERSION DATE	07/02/2024
STATISTICIAN	?
STUDY CHIEF INVESTIGATOR	Professor Lesley Colvin
DAP AUTHOR	Dr Nouf Abutheraa, Professor Peter Donnan

Research questions:

- Primary Research Question: What is the optimal combination of routinely coded data (such as diagnostic clinical Read codes, prescribing data, secondary care referral) needed to construct an algorithm that could be used efficiently to identify people with chronic pain in the community using primary care health records?
- Secondary Research Questions: What are the factors that determine accurate identification of cases using this algorithm? What issues need to be addressed to maximise the algorithm's ability to produce an accurate, meaningful record of people living with chronic pain in the community, including associated healthcare needs? What are the sensitivity, specificity, and positive and negative predictive values of the refined algorithm compared to the reference datasets (medical notes review and survey)?

Comparison of Medical Notes Review (A) and Survey questionnaire (B)

Phase A (the medical notes review) will collect data from individual patient records including read codes for medical conditions and prescribing data. In addition, some free text will be collected from referral letters, GP notes, scans and lab results. These data will be used to classify patients as having chronic pain (yes) or not having chronic pain (no). The sample size is 200 from 6 practices giving 1,200 subjects.

This proportion (i.e. 200) of each practice population undergoing the medical notes review will also be invited to participate in Phase B (survey data collection tool). The survey will ask them if they have chronic pain = yes or don't have chronic pain = No. We will compare these 2 outcomes using Chi-Square tests (χ^2). An additional 1000 are collected in the survey and these can be compared to A using unpaired Chi-Square test (χ^2)

> The *null hypothesis* is that there is no difference between the number of patients identified with chronic pain using medical review notes and the survey data collection tool.

	Chronic Pain *(Yes)*	*No Chronic Pain* *(No)*
Medical notes review	***n***	***n***
Survey data collection	***n***	***n***

For the 200 in both A and B these will be analysed as paired data using McNemar's test. We can also compare the total in the survey which is larger than the medical notes review using the unpaired Chi-squared test.

Comparison of Medical Notes Review (A) with results from the SPIRE algorithm (D)

Using data from Phase A as described above with its binary outcome of yes= chronic pain and no = no chronic pain. This data will be compared with Phase D (the SPIRE report) which used an algorithm to identify patients with chronic pain using the same read codes used in Phase A. The Medical Notes Review (A) will be considered the gold standard and so dependent variable in the binary logistic regression. Read codes that were used in both methods will be added as independent variables and the outcome will be chronic pain = yes or don't have chronic pain = no according to A.

The list of patients included in this analysis will be patients who identified with chronic pain from Phase A (as a gold standard) as well as those identified as no chronic pain (n = 1200) and the output of yes/no will be coming from Phase D (the SPIRE report) to assess how far away we are from the gold standard.

The output will include sensitivity, specificity, Positive Predictive Values (PPV), Negative Predictive Values (NPV), and the Area Under the Receiver Operating Characteristic (AUROC) curve measuring discrimination along with calibration summaries and plots.

> The *null hypothesis* states that all coefficients in the model are equal to zero.

The coefficients (i.e. the B column) will be multiplied by the variables to provide a prediction of chronic pain after adding the constant value of the regression. Having a coefficient of zero means that the variable does not affect the prediction of the outcome. The equation will be similar to:

> ***Predictor of chronic pain = log (p / 1 – p)*** = the constant value + coefficient1*variable1 + coefficient2*variable2 …… + coefficientz*variablez
>
> Where p = probability of chronic pain

Missing data

Assessment of missing data will be carried out and if missing at random is a reasonable assumption multiple imputation may be utilised as a sensitivity analysis.

For an AUROC derived from logistic regression with an assumed prevalence of 40% and a sample size of 1200, the precision for a 95% CI will be +/- 2.6 (PASS 2021 software).

Reporting Conventions

P-values ≥0.001 will be reported to 3 decimal places; p-values less than 0.001 will be reported as “<0.001”. The mean, standard deviation, and any other statistics other than quantiles, will be reported to one decimal place greater than the original data. Quantiles, such as median, or minimum and maximum will use the same number of decimal places as the original data. Estimated parameters, not on the same scale as raw observations (e.g. regression coefficients) will be reported to 3 significant figures.

Qualitative data analysis plan

Data collected from Phase B (the survey) will use mainly a descriptive analysis of numbers and percentages. While interview and focus group data collection will be analysed using a thematic analysis approach to explore broad themes in the data, while also drilling down through the data for a deeper analysis. The analysis will adopt a six-phase approach, described below:

Phase 1: Data familiarization
Phase 2: Initial code generation
Phase 3: Generating (initial) themes
Phase 4: Theme review
Phase 5: Theme defining and naming
Phase 6: Report production