Dynamic Incidence Prediction using Categorical EHR Data with Timestamps (DIPCAT)

Ben Jacob, Patrick Redmond
RCSI University of Medicine and Health Sciences, Dublin, Ireland

Background: Routine data in the electronic health record (EHR) is known to contain signals for as-yet-undiagnosed cancer, e.g., investigation results, disease/symptom codes, or prescribing patterns. While certain biomarkers (e.g. Haemoglobin or HbA1c) are represented as a continuous variable, it is anticipated that complicated biomarker tests which are interpreted by a machine learning classifier (Yes/No) will also be included in the EHR, after they are approved for cancer screening in a primary care population.

Problem: If every positive test is expected to lead to thorough investigation (e.g. a CT scan), the test must have very high specificity. The PATHFINDER study of the Galleri cfDNA MCED test matches this description, however the SYMPLIFY study of the same biomarker test illustrated its poor ability to function as a rule-out test (due to its low sensitivity for early-stage cancer). Since the primary care population is characterised by a high prevalence of cancer symptoms but a low prevalence of undiagnosed cancer, high-sensitivity tests for ruling out cancer are needed and will soon be used repeatedly in the same patient (e.g., the need the rule out ovarian cancer in a patient with frequent attacks of bloating), thus begging the question of how to utilise repeated results from high sensitivity tests (known to produce transient false positives).

Methods: Here we outline:

1. the design of the “DIPCAT algorithm” which:
   • predicts cancer incidence using five types of input data: (1) Date of birth (timestamp), (2) Sex (binary, baseline = female), (3) Comorbidities (multiple binary variables, timestamped), (4) Signal data (categorical, timestamped), (5) Timepoint, t , for which the model will make the prediction
   • utilises six types of functions (with their associated parameters)—a baseline function, three signal functions, a comorbidity modifier function, a probability mapping function— to output a probability of cancer
2. an approach to cleaning the data to produce a data set of all temporally-contiguous registered data in “no-cancer-controls”, the registered data >2 years prior to cancer diagnosis in “other-cancer-controls”, and the pre-cancer records of cancer cases
3. an approach to using non-linear regression (NLR) to find the optimal values for the model parameters (outlined below), by minimising the following cost function: $AUC_{outside} - AUC_{inside} \cdot k$
   • the curve is instantaneous risk of cancer (quantified as real number monotonically related to the probability)
   • the AUC outside the pre-diagnostic window (i.e. not occurring ~12months before a cancer diagnosis) is considered a false positive signal
   • the AUC inside the pre-diagnostic window is considered a true signal
   • k is simply a pre-defined scale factor to prevent $AUC_{outside} \gg AUC_{inside}$, given the rarity of cancer, which would reduce NLR algorithm efficiency
4. an approach to training the probability mapping function instantaneous risk to probability using Cox Regression
5. a suggested approach to validation given a variety of categorical variables
6. preliminary results from an Irish dataset

Implications: Our ultimate aim is to find a way to determine the need for further cancer investigations given series of high sensitivity categorical MCED blood test results, which are known to commonly include transient false positives. However, the approach can immediately generalise beyond lab results to any categorical EHR signal (e.g. consultation dates, symptom codes, symptom phrases, or suspicious prescribing patterns). We believe that this work moves forward the discussion on how best to utilise EHR records for early cancer detection.