# Dynamic Risk Modelling: Optimising predictive modelling for CVD through landmark survival analysis to better understand risk factors in primary prevention populations

Abdulrahman Dablouk[1]

[1]Data Science and Advanced Analytics, Data Science and Artificial Intelligence, Biopharmaceuticals R&D, AstraZeneca, Cambridge, UK

## Background

- Cardiovascular disease (CVD) is the **leading cause of morbidity and mortality worldwide**[1] with research linking a variety of risk factors to CVD.
- Low density lipoprotein cholesterol (**LDL-C**) is a key risk factor. Studies imply a **strong causal relationship between CVD and LDL-C**[1], and statins are effective at lowering CVD risk through LDL-C reductions.
- A **better understanding of the risk factors associated with CVD**, amongst different populations, may **help improve outcomes** and reduce the burden of CVD on health services.

- In England, the **QRISK3**[2] model is recommended for use by clinicians to assess the risk of developing CVD and to help inform decision-making around initiating statins or other therapies.
- This prediction model calculates a person's **10-year CVD risk** using longitudinal electronic health record data on patients and models the **relationship of age on risk and risk factors**.
- Estimates of model performance may be **over-optimistic** because of the way age is modelled, which allows comparison of risk between older and younger patients.

- Here, we use the **landmark model framework** to more flexibly account for the impact of age on risk factors for CVD, in a primary prevention population.
- We used the Optum Clinformatics Data Mart dataset of **health insurance claims data of ~17 million patients** in the US collected over the course of 18 years.

## Glossary

**Survival Analysis**: Statistical methods used to analyze and predict the time until an event of interest occurs, accounting for censorship.
**Landmark Model:** Statistical method that involves predefining a set of landmark times, selecting patients who are event-free at each time point, and fitting a separate survival model for each cohort.
**Cox-Proportional Hazards Model:** Type of survival model that estimates how predictors influence the risk of an event occurring over time in survival analysis, assuming the effect of each predictor is proportional.
**LOCF**: Last observation carried forward, to impute missing values as the most recently recorded ones.
**HR per SD:** The hazard ratio per standard deviation shows how much risk changes when a variable increases by a similar distributional amount across biomarkers.
**C-Index:** Measures how well a model can correctly rank individuals by risk. 1 = perfect ranking and 0.5 = no better than chance.
**95% Confidence Interval**: Range of values that likely contain the true value, with 95% certainty based on the data.

**Cardiovascular Disease:** Group of disorders of the heart and blood vessels caused by atherosclerosis and thrombosis (Strokes, Heart Attacks etc.).
**Biomarker**: Measurable indicator of a biological state or condition.
**Primary vs. Secondary Prevention Populations:** Looking at Individuals before a CVD event vs. after.

## Methodology

**1 million patients**

**Landmark Year ≥ 2021**

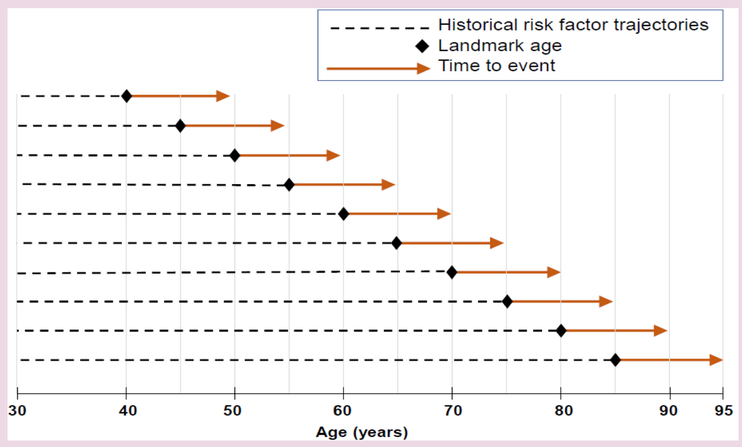
**538,243 patients**

**Event Date ≥ Landmark Date**
**✕ Prior CKD5 diagnosis**
**✕ LDL-C modifying prescriptions in the 90 days before Landmark Date**

**303,902 patients**

**Prepared model cohort**

### Setup landmark model




**Split cohort dataset into training & testing sets: 2/3 Train | 1/3 Test**

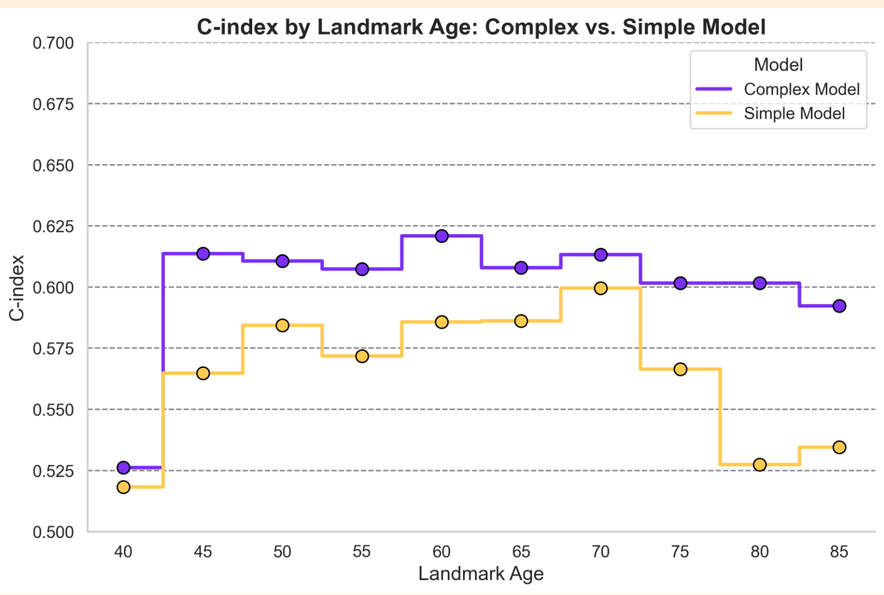
**Fit a Cox-Proportional Hazards Model for each landmark age**

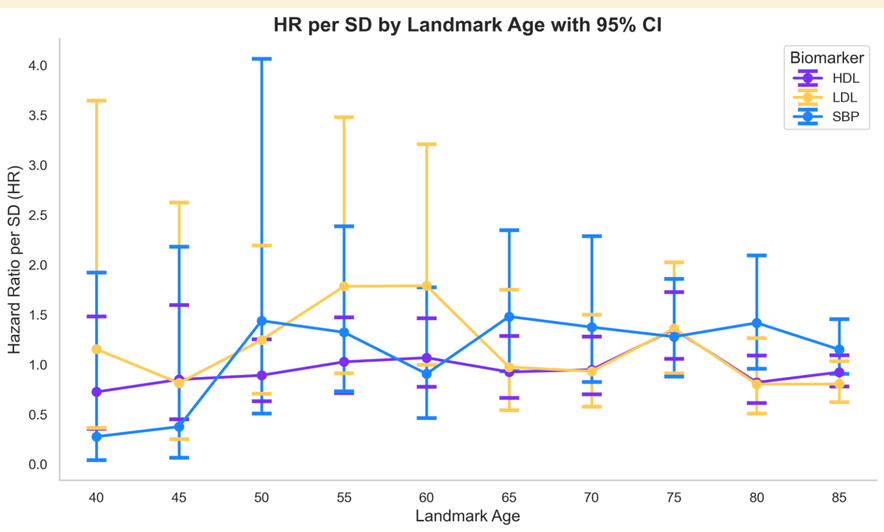
**Flagged missing data (missing data indicators) and employed a Last Observation Carried Forward (LOCF) framework to deal with missingness**

**Fit Simpler Model**

| Biomarker | Simple Model | Complex Model |
|---|---|---|
| Sex | ✓ | ✓ |
| Biomarkers typically associated with CVD (SBP, DBP, LDL-C, HDL-C levels) | ✓ | ✓ |
| 9 Additional lab tests (Hb1Ac, eGFR etc.) | ✕ | ✓ |
| 28 Previous events associated with increased CVD risk (Af, CKD3 etc.) | ✕ | ✓ |
| History of 4 Procedures that impact CVD risk (VR, AA etc.) | ✕ | ✓ |
| Recent Prescriptions of 13 Medications (ACE, BB etc.) | ✕ | ✓ |

## Results




- **Complex Model outperforms Simple Model**
  - Improvement however is incremental.
- 0.52 – 0.62 C-indices range for both models implies **modest discrimination**
  - C-indices > 0.7 indicate good discrimination.
- **C-index lowest at age 40**, **Highest in 45-70 age range** and drops off beyond that
  - Age 40: at younger ages, harder to predict events due to **fewer cases** and **more randomness.**
  - Ages 75+: possibly result of **competing risks** in older ages & the **challenge of identifying old age risk factors.**

- **HDL-C:** HR per SD ranges from ~0.72 (age 40) to ~1.07 (age 60), generally close to 1, some <1 (suggesting **higher High density Lipoprotein Cholesterol (HDL-C) may be associated with lower risk).**
- **LDL-C:** For Ages 40-75, HR per SD is generally >1 (e.g. 1.78 at 55 & 1.36 at 75) or very close to 1 (0.97 at 65) indicating **increased CVD risk with higher LDL-C**. However, **beyond that it is <1 (e.g. 0.8 at 80)**, which can suggest lowered CVD risk with increasing LDL-C but could also be the **result of competing risks**.
- **SBP:** HR per SD is **<<1 at ages 40 & 45** but this is likely the **result of higher degrees of missingness**. For ages 50-85, the HR per SD >1 indicating **increased CVD risk with higher SBP levels**.
- It is however important to note that for all biomarkers, **confidence intervals are wide** and generally include 1 at all ages.

## Wider Impact & Clinical Applications

**1. Identifying high risk populations for clinical trials**
- Patients at high risk for CVD events may be included in clinical trials, particularly if their risk remains high on current standard of care, such as statins.
- Trials that include high risk patients are likely to be shorter given the expected increased event rate.

**2. More tailored preventative care**
- This same model can be applied to assessing risks in other diseases to better individualized preventive care.

**3. Resource allocation & screening**
- To help identify higher-risk individuals for targeted follow up.

**4. Patient communication & engagement**
- Providing individualized risk estimates can enhance shared decision-making, improve patient motivation, and aid understanding of the benefits (or lack thereof) of preventive therapies.

**5. Help identify most pertinent risk factors in old age**
- The predictive value of traditional risk factors such as LDL-C & SBP are inconsistent in old age[3].

**6. Confirm or challenge clinical assumptions around various risk factors**

## Building on this Research

- **Fitting the model to the entire 17 million patient claims dataset**
  - Will help tackle issues to do with biomarker missingness and smaller event rates within different groups.
  - This will likely result in a more robust identification of risk factors amongst primary prevention populations. Additionally, higher and a more consistent model performance.
- **Rebuilding the model for secondary prevention populations**
  - Will help test the model's efficacy in individuals who have already had a CVD event and enable a better understanding of how risk factors change.
  - It will also allow us to test the efficacy of medications prescribed post event.
- **Validate model against external dataset**

**References**
1. Brian A. Ference et al., Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel, *European Heart Journal*, Volume 38, Issue 32, 21 August 2017, Pages 2459–2472, https://doi.org/10.1093/eurheartj/ehx144
2. https://qrisk.org/
3. E.F. van Bussel et al., Predictive value of traditional risk factors for cardiovascular disease in older people: A systematic review, Preventive Medicine, Volume 132, 2020,105986, ISSN 0091-7435, https://doi.org/10.1016/j.ypmed.2020.105986

**Acknowledgements**
I would like to extend my gratitude to Matthew Arnold for his guidance and support and the Healthcare Analytics team at AstraZeneca and HDRUK for providing me with this amazing opportunity.

*Supported by* AstraZeneca