Forecasting disease burden in Ulcerative Colitis

SCAN-2030 Work Package 2 [WP2] Analysis Plan

1. Objective

This study aims to forecast the incidence, prevalence, and healthcare burden of Ulcerative Colitis (UC) in Hong Kong over the next ten years (2023–2032) by leveraging historical data from 2014 to 2022. Using time-series statistical modeling techniques, particularly ARIMA models, we seek to predict future trends to support healthcare planning, policy-making, and resource allocation. Understanding the future burden of UC is crucial for effective decision-making, enabling policymakers and healthcare providers to anticipate future demands on the healthcare system. Additionally, estimating the economic impact and unmet treatment needs will help shape innovative treatment strategies and healthcare policies.

2. Rationale

Forecasting the burden of Ulcerative Colitis is essential for healthcare system preparedness and resource allocation. UC is a chronic inflammatory bowel disease (IBD) that requires long-term management, often involving hospitalizations, medication use, and surgical interventions. Predicting future incidence and prevalence trends will help ensure adequate healthcare infrastructure, including specialist services, hospital capacity, and pharmaceutical supply chains. Additionally, estimating the economic burden and identifying gaps in treatment access will facilitate the development of cost-effective healthcare policies and innovative treatment approaches to improve patient outcomes.

3. Study Population and Data Source

The study will utilize data from the Clinical Data Analysis and Reporting System (CDARS), a comprehensive electronic medical record database managed by the Hospital Authority (HA) of Hong Kong. This dataset includes patient demographics, hospital attendance records, diagnoses, prescriptions, and laboratory test results across inpatient, outpatient, and emergency care settings. Additionally, mid-year population data from the Census and Statistics Department of Hong Kong will be used to calculate age-standardized rates and project future disease trends.

Prevalence, incidence, and cost data were aggregated based on analyses performed as part of SCAN-2030 Work Package 1.

4. Statistical Analysis

We employ an AutoRegressive Integrated Moving Average (ARIMA) modeling approach. ARIMA and regression with ARIMA errors are well-established frameworks for analyzing and predicting time-series data, especially when historical patterns, autocorrelations, and potential interventions must be accounted for. For each outcome, candidate models are automatically generated and evaluated using the auto.arima function from the R forecast package, which systematically explores a wide range of plausible ARIMA configurations. To address issues of non-normality and heteroskedasticity commonly observed in healthcare and cost data, both the raw and log-transformed versions of each outcome series are modeled in parallel.

Our modeling strategy is enhanced through the incorporation of exogenous regressors. Binary indicator variables—such as those representing the onset of the COVID-19 pandemic (dum_pan) or significant social events (dum_so)—are included as external covariates (xreg) within the ARIMA framework. These covariates are intended to capture abrupt, non-repeating shifts in the time series that are not explained by endogenous patterns alone. For the historical period (2014–2022), they reflect actual events; for future forecasts (2023–2032), the default assumption is that these events do not recur, and the covariate values are set to zero unless otherwise justified.

4.1 Model Comparison and Selection Criteria

Model selection is guided by a combination of information criteria and error minimization. For each outcome and transformation, ARIMA models are fit to minimize both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), providing a balance between model fit and complexity. All candidate models are extracted and ranked according to their information criterion scores. When the top two models are closely matched (i.e., score gap <2), the model with the higher sum of ARIMA orders (p+d+q) is preferred, as it offers greater flexibility in capturing temporal dependencies and trends—provided that this does not introduce overfitting. If model comparison is inconclusive or model refitting fails, the best default model from auto.arima is retained.

Comprehensive model validation is performed for each selected candidate. Residuals are inspected for randomness and absence of autocorrelation, and the mean squared error (MSE) is computed on the training data. For log-transformed models, fitted values are back-transformed to the original scale prior to MSE calculation to ensure comparability across modeling strategies. The final model selection for each outcome is based on the lowest observed MSE, either on the original or back-transformed scale, thereby prioritizing predictive accuracy. Additionally, historical back-testing is conducted by comparing model predictions within the observed period to actual values, providing a practical check on the model's forecasting fidelity.

5. Study Outcomes

The analytical approach involves data preprocessing, incidence rate calculations, and forecasting using R programming. Diagnosed cases will be grouped by year, and age-specific incidence rates will be computed. Population data will be used to adjust for demographic changes, ensuring accurate age-standardized incidence estimates. The ARIMA forecasting model will predict future trends, with the best-fit model determined through statistical evaluation.

The results will be presented graphically using trend plots and stored in Excel reports for further analysis. These forecasts will provide evidence-based insights to optimize gastroenterology services, improve patient outcomes, and enhance healthcare preparedness for Ulcerative Colitis in Hong Kong over the next decade.