Tianjin Diabetes and Health Cohort Study

Document Date：2025-05-07

ClinicalTrials.gov ID：NCT06913153

Statistical Analysis Plan

Specific processes for analyzing data

1. Define the purpose of the study;
2. Identify the study population;
3. Identify intervention, exposure factors;
4. Identify outcome indicators;
5. Clean and organize data;
6. Study grouping; Missing data handling.

Statistical methods used for each analyzed item:

1. Propensity Score Matching (PSM)
2. Multivariate Parametric Proportional Hazards Model
3. Gompertz Proportional Hazards Model
4. Logistic Regression
5. Cox Proportional Hazards Model
6. Kaplan-Meier Curve
7. Autoregressive Integrated Moving Average (ARIMA) Forecasting Model
8. Bayesian Structural Time Series (BSTS) Model
9. GXboot ， Support Vector Machines， Decision Trees， Random Forests
10. Health economics： Cost-Effectiveness Analysis(CAE)， Cost-Utility Analysis(CUA), Cost-Benefit Analysis(CBA)

The way key assumptions are verified:

1. Propensity Score Matching (PSM)：PSM required balanced covariates (SMD<0.1) and overlapping propensity score distributions for the exposed and control groups after matching. Balance will be verified by standardized mean difference (SMD), propensity score distribution plots, and between-group tests (p>0.05).

2. Multivariate Parametric Proportional Hazards Model：The model assumes that the risk ratio is constant over time (proportional riskiness) and that the parameter distribution (e.g., Weibull distribution) accurately describes the time dependence. The proportional hazards assumption will first be verified by the Schoenfeld residual test (p>0.05), and the goodness of fit of the model with different parameters will be compared by AIC/BIC.

3. Gompertz Proportional Hazards Model：The model assumes that the risk function follows a Gompertz distribution (exponentially varying) and satisfies proportional riskiness. The distributional assumptions will be verified by plotting the log cumulative hazard linearly versus time and comparing with other parametric models (e.g., Weibull) using the likelihood ratio test.

4. Logistic Regression ： Logistic regression requires

continuous variables to be linearly related to the logit, free of multicollinearity and independent of observations. The assumption of linearity will be verified by the Box-Tidwell test, variance inflation factor (VIF<10) will be calculated to exclude multicollinearity, and data will be checked for independent samples.

11. Cox Proportional Hazards Model

12. The core assumptions of the Cox model are proportional riskiness and linear association of covariates with log risk. These two assumptions will be tested using the Schoenfeld residual test (global p>0.05) and Martingale residual plots, respectively.

13. Kaplan-Meier Curve

14. Kaplan-Meier analyses assumed that censoring was independent of event risk and that the mechanism of censoring was consistent between groups.Randomness of censoring will be indirectly assessed by comparing baseline characteristics of censored and non-censored subjects and by log-rank test results. If censoring is not random, data will be processed using competing risk models or multiple imputation.

15. Autoregressive Integrated Moving Average (ARIMA) Forecasting Model

16. The ARIMA model requires the time series to be smooth and the residuals to be white noise. Stationarity will be verified by the ADF test, and residual autocorrelation will be checked using ACF/PACF plots and the Ljung-Box test ($p>0.05$).
17. Bayesian Structural Time Series (BSTS) Model
18. The BSTS model assumes that the temporal structure (trend, seasonality, etc.) is well set and that the prior distribution matches the data characteristics. The model will be validated by posterior predictive checks, information criteria (WAIC/DIC), and residual normality diagnostics.