

遺伝統計学チームの研究ミッション

生命科学（医学や農学）のビッグデータを機械学習・人工知能ベースの遺伝統計学で解析を行い、ヒトの疾患や農業形質などの複雑な現象に寄与する要因を探り出す。今回、我々研究チームの取り組む課題のうち、疾患リスク予測を代表して紹介する。

大規模ゲノムコホートデータへの疾患リスク予測手法の適用

背景

事例	推定対象	人数
Ueki & Tamiya (2016)	アルツハイマー病	700人程度
Takahashi et.al. (2020)	うつ症状 CES-D値	5千人程度
2021年度	生活習慣病 HbA1c等41形質	3万人程度
今年度	生活習慣病 HbA1c等6形質	14万人程度

近年のゲノムデータを活用した疾患リスク予測の有用性が注目を集めている。遺伝統計学チームが開発したSTMGP法は、アルツハイマー病や抑うつ症状で、高い予測精度を示した。疾患リスク予測の実用化に向けて、より多くの疾患を対象にして、さらに大規模なデータでの検証が求められる。

これまでの進捗

[UKバイオバンクデータの解析]

UKバイオバンクの10万人を超える大規模データを解析し、生活習慣病における精度を検証した。その結果、STMGP法が他手法よりも精度が高く、英国人(Caucasian)の解析でも優位性があることを確認した。

[外部要約統計量の活用]

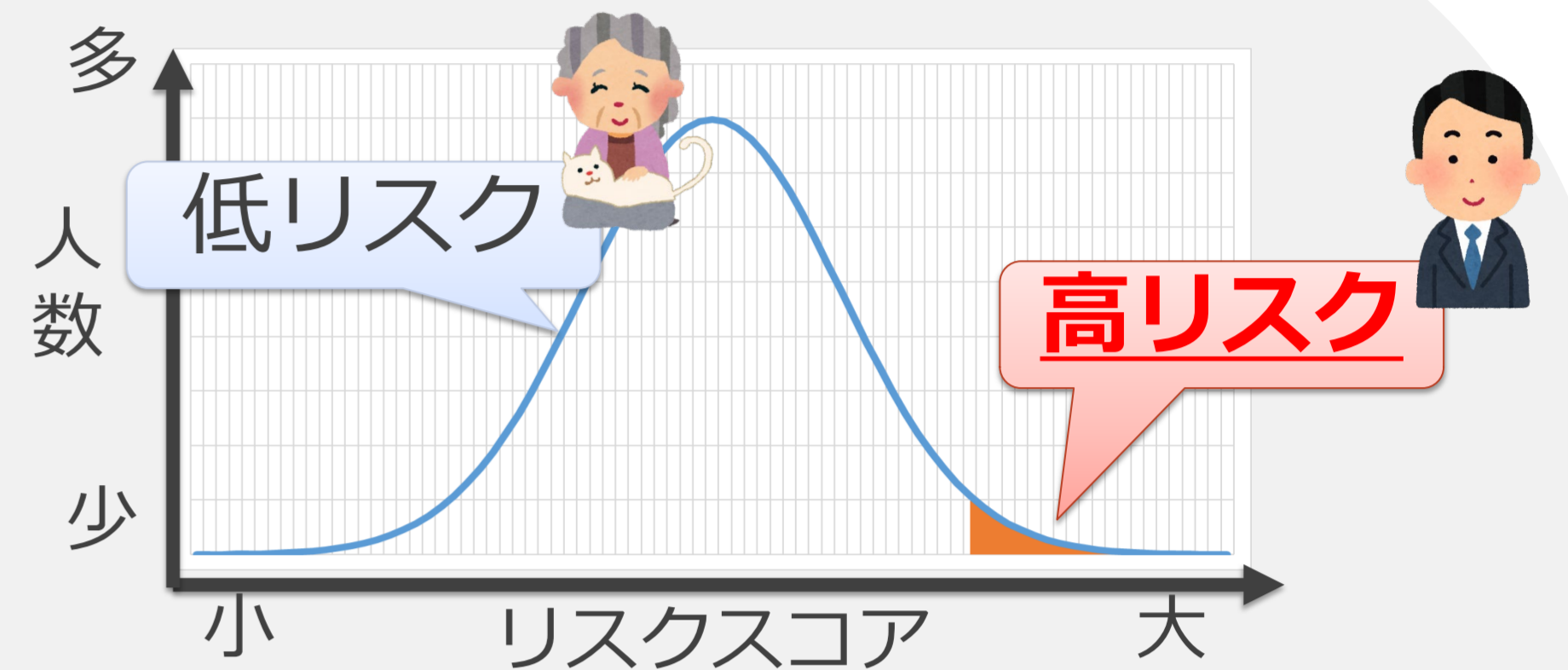
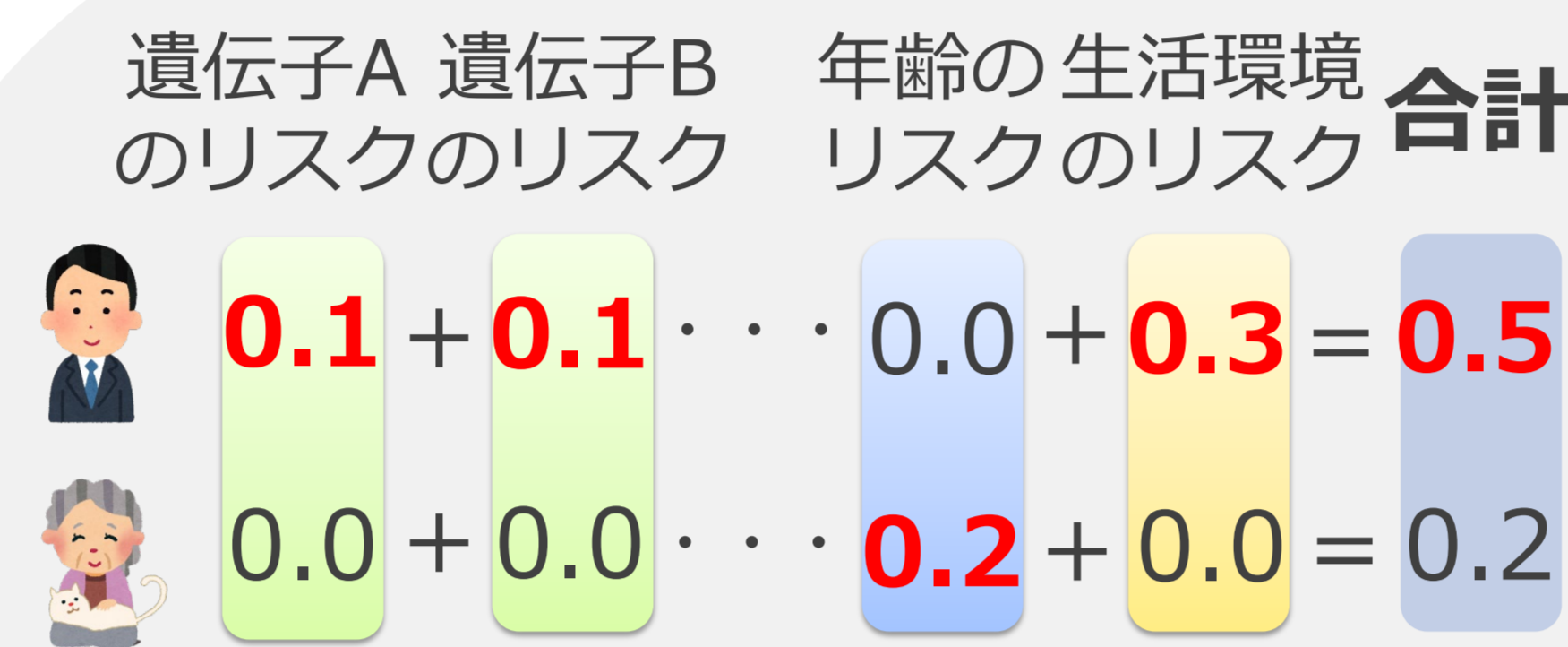
外部に公開されている既存研究(要約統計量)を活用し、解析精度を高める研究が注目されている。STMGP法においても外部要約統計量を活用するロジックを実装した。GLGCコンソーシアムの要約統計量を活用した検証を実施し、東北大学のデータ解析精度を高めることに成功した。

[分割計算の実現]

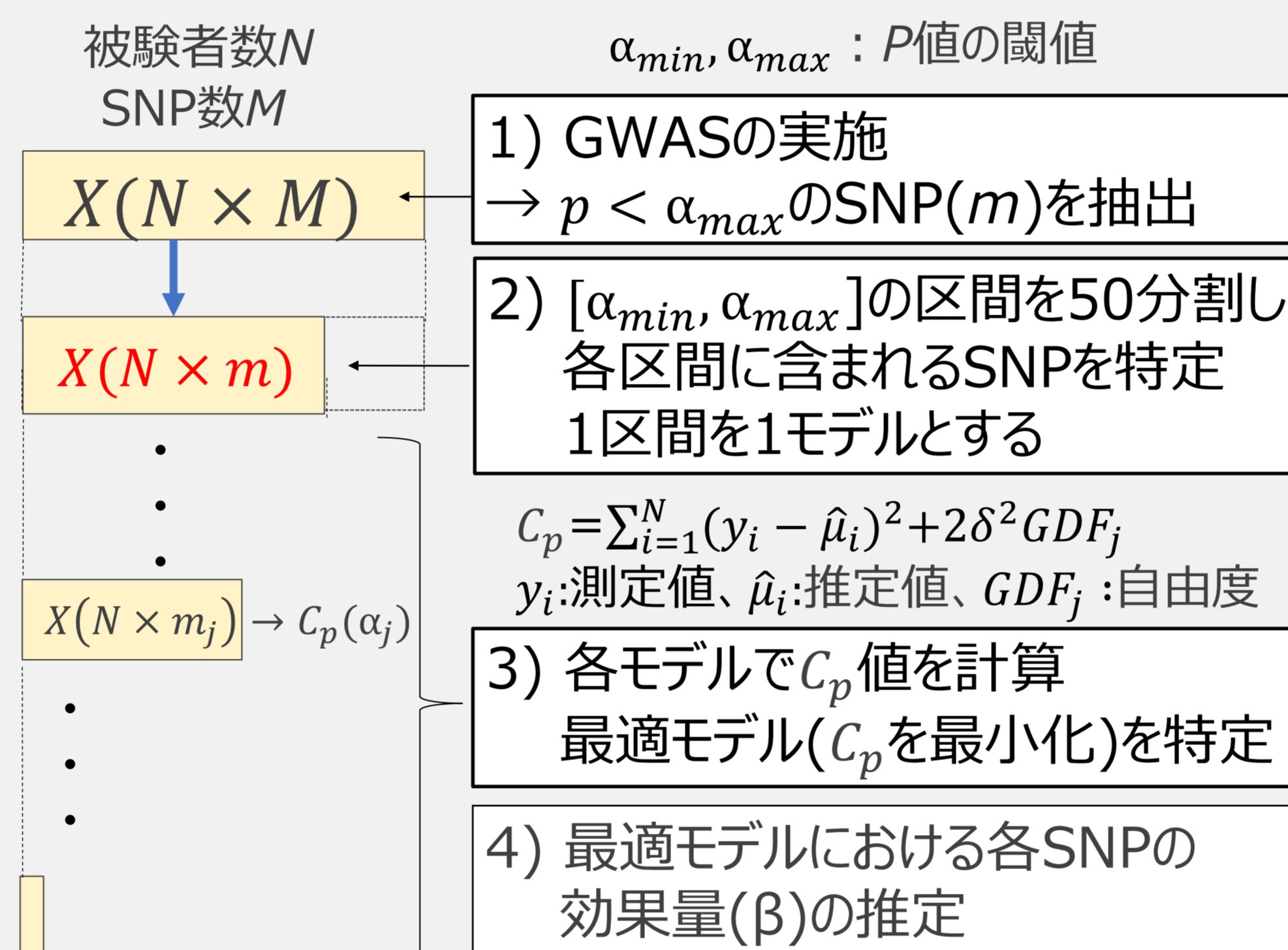
STMGP法による、さらなる大規模なデータ解析を実現するため、STMGP法に分割計算を可能とする機能を考案した。東北大学のデータによる検証では、入力データを半分に分けてもある程度の精度が得られ、他手法より高精度である結果が得られた。

解析手法

疾患リスク予測

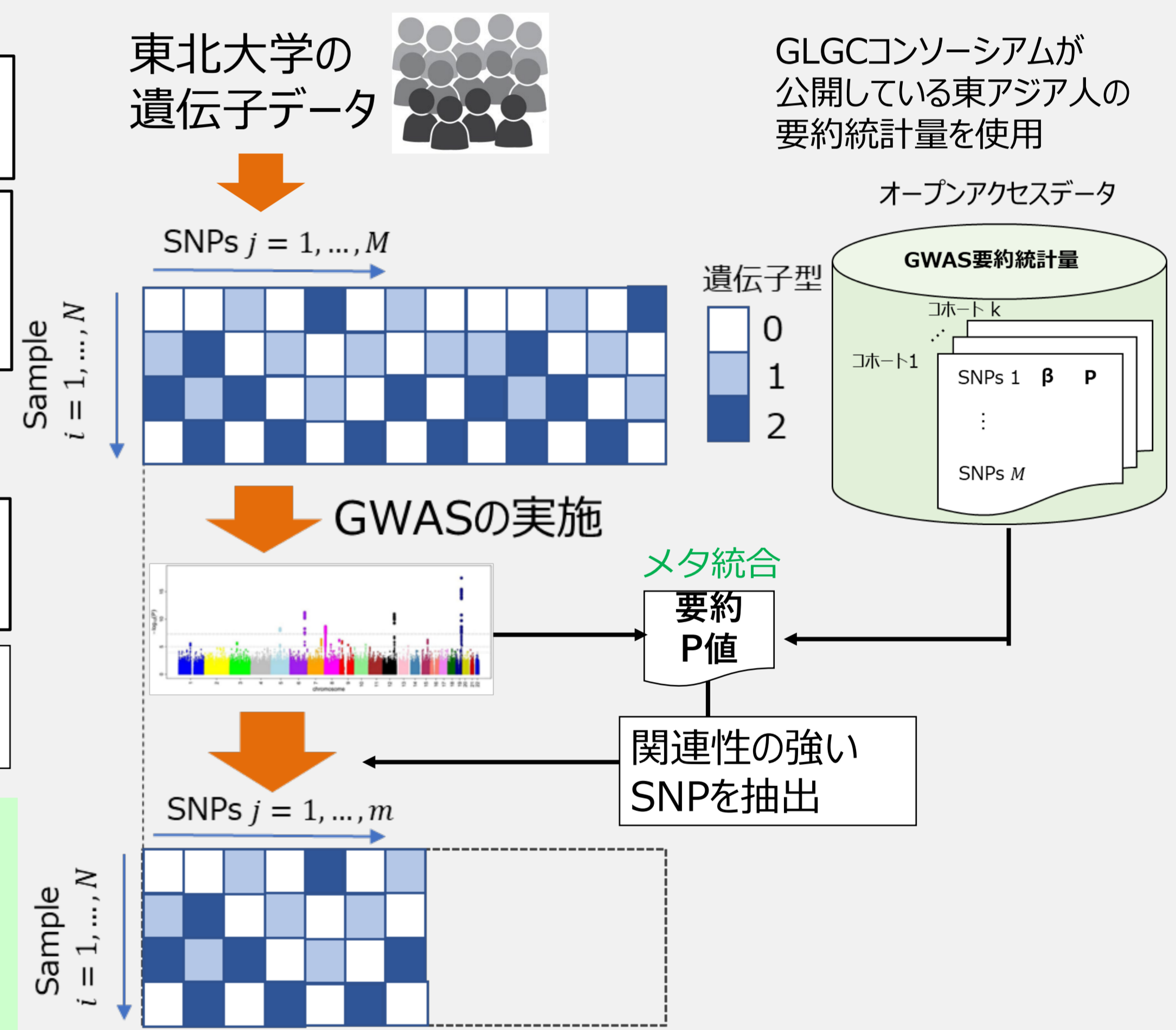


STMGP法

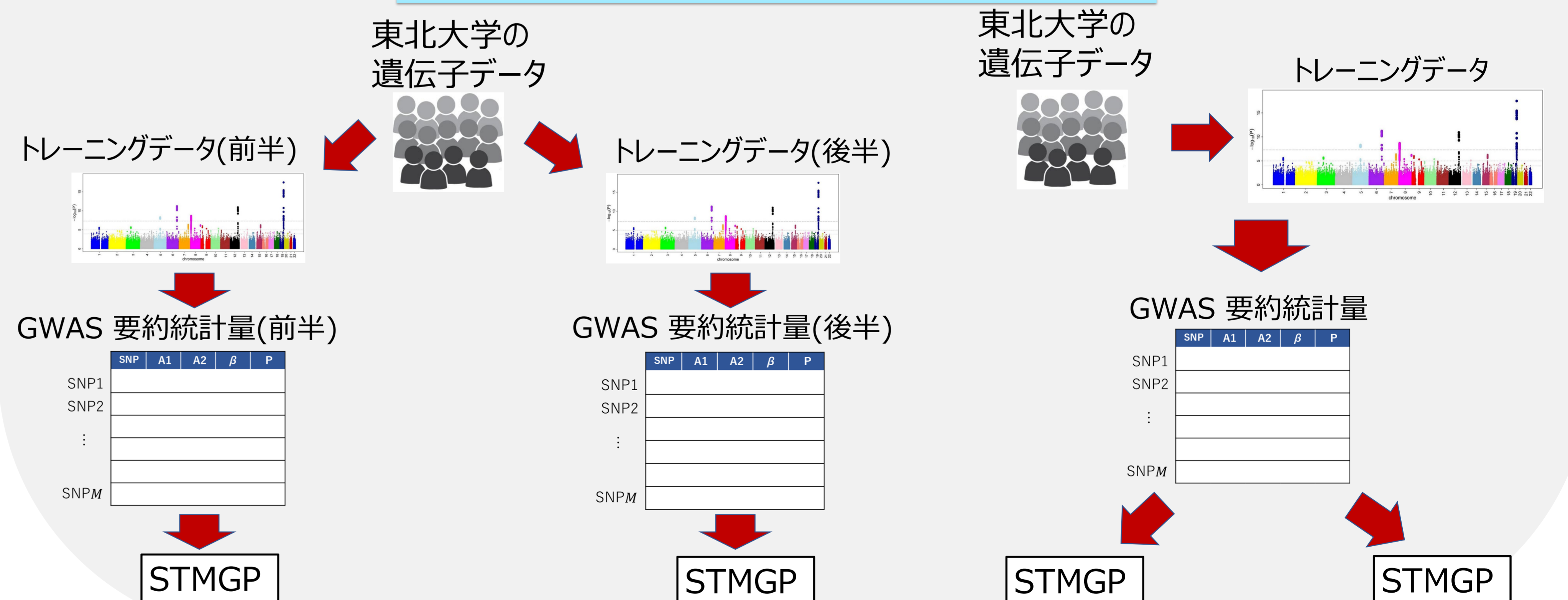


STMGP法の特徴(メリット)
最適な計算モデル・パラメータが自動選択されるため、検証工程(バリデーション)が不要。
近似なしで計算モデルを厳密に構築可能。

(重点研究1) 外部要約統計量の活用

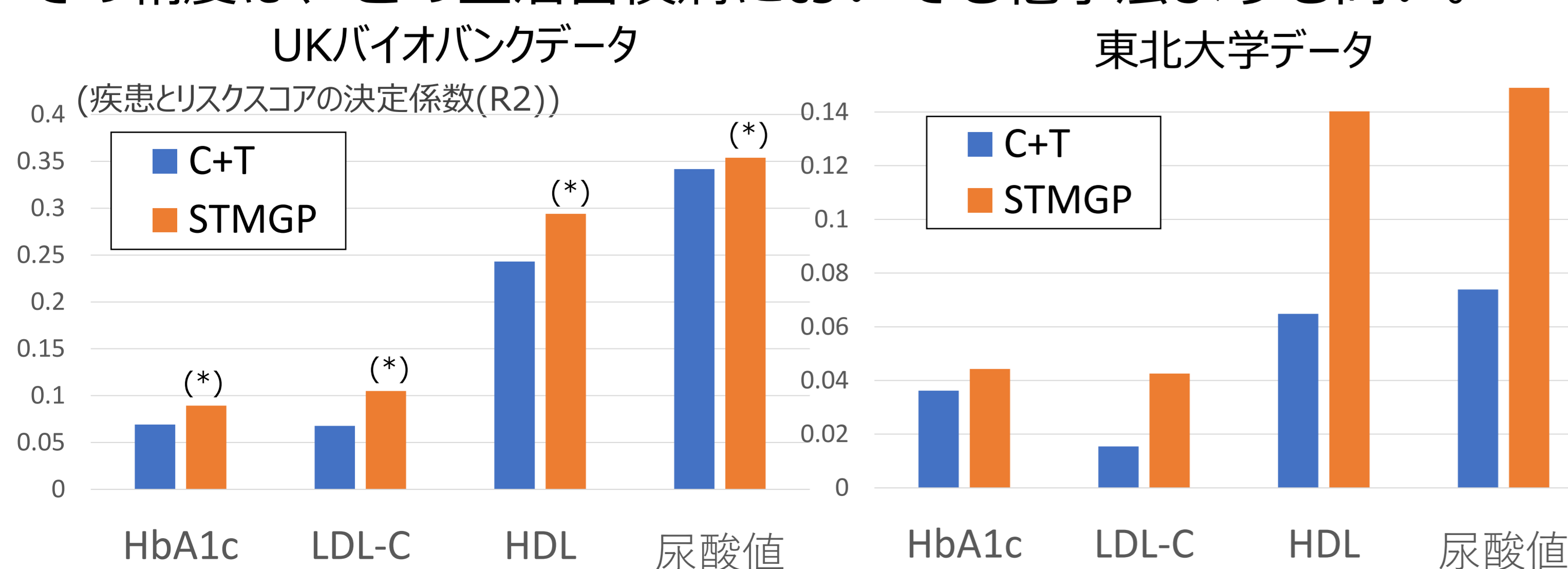


(重点研究2) 分割計算の実現



結果I. UKバイオバンクデータの解析

STMGPがUKバイオバンクの大規模データ解析に対応。その精度は、どの生活習慣病においても他手法よりも高い。



結果II. STMGP法の分割計算の検証

対象人数が多くなると計算量が増え、使用メモリ量が指数的に増加するため、メモリ量を削減する工夫を行った。

