因果推論のためのシングルパス連合集約：合成データによる□□億規模の検証

和智大次郎

独立研究者 □ オープンソースエンジニア

□□□□□□□□□□□□□□□□□□□□□□□□□□

要旨

シングルパス連合集約による因果推論が、市販のハードウェア上で計算的に実現可能であることを示し、制御された合成実験において理論

方法□ □□□□□□□□□□□□□□□十分統計量集約（コンパクト実装では約□□□バイト□サイト、現行実装では約□□□バイト□サ

結果□ □□□万患者を□□□□□秒（□□□□□□□□□□□□□□□）で、□□億患者を□□□□分（□□□□□□□□□□□□□□□□□□）で処理（いずれも□□コア）。□種類の推定量を実装：□□□シンプル平均差推定量（□□バイト□サイト、□□□□□付き□□□□□□□□ □□□推定量（約□□□バイト□サイト）。

負の対照として、未調整平均差推定量による□□□万規模の推定が、□□□□□□□□でも系統的バイアスを示すことを確認：□□□□有□ 推定□□□□□□）、□□□□□□有病率では著しい過小推定（埋め込み□□□□ □ □□□□□□）。

真の条件付き□□□をアウトカムモデルから解析的に導出して検証した□□□推定量は、主要な知見を示す：規模拡大が最もまれなサブグ□□□□□、有病率 約□□□□□□）は、□□□万人での□□□□□の□□□□過小推定から□□□万人での□□□□□□□□での□□□□□□）は□□□以内、□□□□□□□□□□□□□（真 □□□□）は□□□□および□□で□□以内。

結論□ シングルパス連合集約は□□□□の計算量（患者ごとの処理は一定）を達成し、市販のハードウェアで□□億患者データセットを□ □億人で□が□□□□□□のときに□□□）、□□□不安定性が持続するより大きなサブグループでは、□に依らず残余バイアスが残る

限界□ 全ての検証は既知の真値を持つ合成データを使用している。実世界への適用可能性は、機関間の共同研究による実証的なテストが必

キーワード□ 連合学習、因果推論、シングルパス集約、□□億規模計算、□□□□スケーラビリティ、陽性度違反

---

1. はじめに

1.1 動機

まれなサブグループ（有病率 □□□□□）における因果推論は、よく知られた統計的課題に直面する：傾向スコアの陽性度違反が、単なる□□□□□。処置群と対照群が共変量空間において重複しない場合、逆確率重み付け □□□ は高い統計的精度を持ちながらも系統的にバ□□□□□。

独立したエンジニアとして、私は私の手の届かない□□□承認と機関合意を必要とする実際の多機関医療データにアクセスできない。代わ

1.2 貢献

単一の主張□ シングルパス連合集約による因果推論が市販のハードウェアで計算的に実現可能であることを示し、合成実験において既知の

裏付けとなる証拠□

- □□ スケーラビリティ□ □□□万患者を□□□□□秒（□□□□□□□□□□□□□□□）□□□億患者を□□□□分（□□□□□□□□□□□□□□）、いずれも□□コア
- □□ 通信効率□ スケールによらず一定のサイトあたりペイロード（現行実装では約□□□バイト□サイト、コンパクト版では約□□□□□）

□□ 規模での検証□ アウトカムモデルから解析的に導出した真の条件付き□□□を用いた□つの規模での□□□推定量。主要な知見□
規模拡大が最もまれなサブグループのバイアスを解消する：□□□□□□□□□□□□□（真□□□□□□）は□□□万人で□□□
□□億人で□□□□□□□□の□□□□□□億人で□□□□□へと改善。より大きなサブグループも良好なキャリブレーション：
□□□□□□）は□□□以内；□□□□□□□□□□□□□（真□□□□）は□□□□□□□□□と□□□□超では収穫逓減

成果物□ 測定されたパフォーマンスと合成データでの検証済み正確性を持つオープンソース参照実装。実際の多施設データベースでのテ

限界□ 全ての検証は既知の真値を持つ合成データを使用している。実世界への適用可能性は、独立したエンジニアとして私のアクセス範

位置付け□ これは計算的実現可能性を確立し、実装の正確性を検証するものである。医学的発見や政策的推奨ではない。

---

2. 実装

2.1 アーキテクチャ：シングルパス連合集約

主要な革新□ 連合学習において一般的な反復的通信を回避するシングルパス集約□□□□□□□。

各サイト□はローカルで計算する：

- 傾向スコア勾配：□□□□□□□
- 傾向スコアヘッセ行列：□□□□□□□□□□□
- アウトカム回帰統計量：□□□□□、□□□□□
- メタデータ：□□□□□□□□□（サイトあたりの患者数）
- 合計：約□□□バイト□サイト（現行実装、完全行列）；コンパクトな上三角符号化では約□□□バイトに削減可能

中央集約機：

- 統計量の集約：□□□□□□、□□□□□□
- □□□□□□□□□□□□□□□をローカルで反復（追加のサイト通信なし）
- □□□□□□、□□□□□□を使用した逆確率重み付けによる因果効果の計算

「シングルパス」の意味□ 各サイトは十分統計量を正確に□回提供する□繰り返しのサイト通信はない。中央集約機はメモリ制約のある
□利用可能な□□□コア数）でサイトを処理する。各バッチの後、集約機は□回の□□□□□□ステップを実行し、次のバッチの前に□
□□□□□□□□の厳密な「ワンショット」定式化（□□□□□□の更新前に全てのサイトを同時に集約する）とは構造的に異なる。
□□□コア）にのみ成立する。□□ランの場合（□□□□□サイト、□□バッチ）、推定値は□がバッチ間で安定化するにつれて改善さ
□□□□□）；完全な軌跡については補足□□を参照。

数学的特性□ 十分統計量集約を用いた□□□□□□□□□□□□□□□連合学習は、総和の結合性を利用する
□□□□。全サイトが□バッチに収まる場合、連合推定値は中央分析と同一である。バッチ処理する場合（□□ラン）、推定値は□がバッ

通信の利点□ 勾配を繰り返し交換する反復的連合学習とは異なり、各サイトは十分統計量を□回送信するだけで、中央サーバーはさらな

2.2 技術スタック

- 言語□ □□□□□□□□□□□□□□□□□□□□ □□□
- 並列化□ □□□□□□スレッド（os.cpus().lengthコア、自動検出）
- データ形式□ □□□□ □□□ □□□□ □□□□□□□□ テーブル（□□□□□□、□□□□□□□□□□□□□□、□□□□□□□□
- メモリ□ サイトあたり□□□□（ピーク□ □□未満；測定値：□□ラン；□□ランでは□ □□未満）
- 通信□ 約□□□バイト□サイト（現行）、約□□□バイト□サイト（コンパクト符号化）；サイトサイズによらず□□□□

このコードベースの□つの実装（異なるセクションで結果を生成）：

実装	役割	推定量	通信□サイト	出力	使用箇所
□□□（run-polypharmacy.ts）	負の対照（ベースライン）	シンプル平均差	□□バイト	□□□、□□、□□、□□	セクション□□□
testOptimized.js	提案された貢献エンジン	□□□□□□□□□□□□□□□	約□□□バイト	□□□のみ	セクション□□□、□□□

□□□は傾向調整なしのバイアスの大きさを示す（負の対照）。□□□エンジンが提案システムであり、傾向重み付けを適用し、同じ十分

2.3 合成データ生成

□□□□互換データジェネレーターを使用してカスタム多剤併用シナリオを生成した。□□□□□□□の臨床パターン□□□□ を構造的に模倣し、異なる有病率レベルでシステムをテストするための□つの相互作用層を持つ：

層	有病率	埋め込み効果（□□□分□年）	臨床パターン	ラベル
□	□□□	□□□□	□□□ ステージ□□□□	一般
□	□□□□	□□□□	□□□ □□ □ ループ利尿薬	まれ
□	□□□□□□	□□□□	□□□ □□ □ ループ利尿薬 □ 年齢□□□	非常にまれ

適応症による交絡（実際のモデル、米国サイト；confStrength □ □□□）：

logit(P(Treatment)) = -2.0 + 0.02×年齢 □ □□□□□□□□ □ □□□□□□□□□□ □ □□□□□□□□□□ □ □□□□□□□□

（confStrengthはサイト別：米国□□□□、日本□□□□、北欧□□□□、インド□□□□；□□□係数 □ □□□□□□□□□□□□□□ □ □□□）

高い□□□□□、高齢、高い□□□は処置確率を高め、低い□□□□はわずかに低下させる（腎機能が著しく低下している場合の処方減少 = -2.0 - 0.05×年齢 □ □□□）。この年齢媒介の負の交絡により、未調整平均差推定量は治療効果を過小評価する□
処置患者は実際よりも改善が少なく見える（反事実的な自然経過がより悪いため）。

規模の仮定□ □□億の構成は□□□□□サイト □ □□□万患者□サイトを使用する。各□□□万患者のサイトは、大型学術医療センター

一般化を制限する重要な単純化：

- □□ 低次元交絡□ ロジスティックモデルの□共変量
  - 現実：□□□以上の変数が複雑な非線形相互作用を持つ
  - 影響：陽性度違反は低次元シミュレーションのアーティファクトかもしれない
- □□ 理想化された欠損□ □□のランダム欠損
  - 現実：無視できない□□～□□□の欠損（より病気の患者はより多くのデータが欠損している）
  - 影響：実際の傾向スコア推定はより困難になる
- □□ 既知の真値□ 埋め込み効果はプログラムされている
  - 現実：真の効果は未知かつ検証不可能
  - 影響：バイアス検出は真値が既知の場合にのみ可能
- □□ 均質なサイト□ 同一のデータ生成プロセス
  - 現実：サイトは人口統計、診療、測定機器、コーディングシステムで異なる
  - 影響：実際の連合集約は異質性の下で失敗する可能性がある
- □□ 未測定交絡なし□ 全ての交絡因子が観測されている

- 現実：常に未測定交絡因子が存在する（遺伝、服薬遵守、社会的要因）
- 影響：実際の因果推論は追加のバイアス源に直面する

含意□ これらは些細な単純化ではない。実際の医療データ分析とは根本的に異なる問題を表している。結果は概念実証としてのみ解釈す

---

3. 結果

3.1 計算スケーリング：□□□万から□□億患者

ハードウェア□ □□□□□ □□□□□□□□□（□□□□□）、□□論理コア。全てのランは利用可能な全コアを使用（batchSize = os.cpus().length）。□□□□年□月□□日に検証。

規模	サイト	患者□サイト	コア	処理時間	スループット（□□□□秒	通信□サイト	合計通信
□□□万□□		□□□□□□□□	□□	□□□□□秒	□□□□□□□□□□	約□□□バイ	約□□□□□
□億	□□□	□□□□□□□□□□□	□□	□□□秒	□□□□□□□□□□□	約□□□バイ	約□□□□□
□□億	□□□□□	□□□□□□□□□□□□	□□	□□□□分（□□□秒）	□□□□□□□□□	約□□□バイ	約□□□□□

□ 現行実装（完全行列）。コンパクトな上三角符号化では約□□□バイト□サイト（□□億では約□□□□□合計）に削減可能。

スケーラビリティ分析：

- 同一ハードウェアでの□つの規模点（□□コア）□ □□□万人で□□□□□秒 □ □億人で□□□秒 □ □□億人で□□□秒。患者あたりの処理量は□□□□であり、□□□□の総計算量を実現。壁時計時間は□□□□コアでおよそ□□ □ 壁時計□□□倍）はアルゴリズムの劣線形性ではない：□□□利用率の向上を反映している（□□□万人では□バッチで□□□ □ データ□□倍、壁時計□□倍。
- スループットは□億人でピーク（□□□□□ □□□□□）：□バッチにわたって全□□コアが一貫して飽和している場合。□□□
- 一定の通信□ 約□□□バイト□サイト（現行実装、完全行列）でサイトサイズによらず；コンパクトな上三角符号化では約□□□

再現コマンド：

```
npm test              # 100万患者 □ □□□□□□□□□□□□
```

層	有病率	□	埋め込み	推定値（□□□ □□）	□統計	□値	偏差	ラベル
□	□□□□	□□□□□□□□	□□□□	□□□□□□ □□□□□□□□ □□□□□□□	□□□□□	□□□□□□□□	□□□□□□□□ □□□□	（符号逆転）まれ
□	□□□□□□□□	□□□□	□□□□	□□□□□□ □□□□□□□□ □□□□□□□	□□□□□	□□□□□□□□	□□□□□□□□ □□□□	（□□□□過小推定）まれ

□□□□年□月□□日に検証（□□□□□□□□、サイト数□□□、患者数□サイト□□□□□□□□、プロファイル□米国）。

再現コマンド：

```
npm run test:tier1
npm run test:tier2
npm run test:tier3
```

解釈：

層□（一般、□□□）□ □□□の過小推定（□□□□□□ 対 埋め込み□□□□）だが符号は正しい。これは適応症による交絡が存在する
0.05×年齢）。未調整の処置□対照比較はそのため見かけ上の治療効果を圧縮する。非常に狭い□□（□□□□□□）はこれが系統的バ

層□（まれ、□□□□）□ 完全な符号逆転。埋め込み効果は有害（□□□□）だが、シンプル推定量は□□□□□□□□で有益（□□□
検出力の失敗ではない。古典的な低検出力による型□エラー □□□ とは異なり、これは処置群と対照群が共変量分布において根本的に異
□□□ループ利尿薬）における構造的交絡から生じる。ループ利尿薬使用は心不全と体液貯留によって引き起こされる□
□共変量傾向モデルにはない危険因子□シンプル推定量が修正できない残余交絡を生み出す。注：セクション□□□では「□□□□□□□

層□（非常にまれ、□□□□□□）□ 著しい過小推定（□□倍小さい）。わずか□□□患者では、処置群と対照群は傾向スコア空間でほ

重要な注意□ これらは検出力の失敗ではない。全ての推定値は狭い信頼区間で□□□□□□□□を達成している。問題は系統的バイアス
推定量の誤った特定（層□）と陽性度違反（層□〜□）から□ランダムなサンプリング変動ではない。高い精度はバイアスの不在を意味し
□□□□□□□。

3.3 提案推定量：□つの規模での□□□□□□□□□□□□□□□□ □□□

役割□ このセクションは主要な貢献を示す□□□億規模の処理をサポートする連合□□□推定量。セクション□□□と同じ十分統計量集

実装□ testOptimized.js、□□□□□□□□□□□□□□□□□□□推定量、約□□□バイト□サイト。□つの規模行（□□□万、□億

□□□ティアとは異なるサブグループ定義□ testOptimized.jsデータジェネレーターは独立したサブグループ基準と埋め込み効果を
□□□□□□□□□□□□□患者で□□□□ □□□分□年）。□□□□□□□□□□□□と□□□□□□□□□□□□はそれぞれ入院と
□□□□ではない。全サブグループの患者は基本□□□□治療効果（□□□□ □□□分□年）を経験し、□□□□□□□□□□□□基準
□ □□□ かつ任意の利尿薬使用）も満たす場合は追加のブーストを得る可能性がある。これはセクション□□□との□の不一致も説明す
□ □□□万 □ □□□□□患者をターゲットにしているが、□□□□□□□□□□□□（raceAsian && age > 75
&& bmi < 22）は実証的に□□□□□□患者（約□□□□）を生成する□全く異なるサブグループ。

注：セクション□□□対□□□の比較□ セクション□□□と□□□は異なるサブグループ定義、異なる推定量、異なる埋め込み効果メカ

シード安定性□ □つの規模全ての結果はベースシード□□を使用する（サイト□はシード□□□□を使用）。□□□万規模での多シード
□□□□□の範囲となり、□□□□□□□□□でのサンプリング変動と一致する。□□□□□□□□□□□□□（□□□□□□）はシード間で

真の条件付き□□□導出□ □□□□アウトカムモデルがegfrChange += treatment × (1.0 + 2.0 × in-
teraction1_flag)であるため、任意のサブグループ□に対する真の条件付き□□□は：

E[eGFR効果 □ 患者 □ □□ □ □□□ □ □□□ □ □□□□□□□□□□□□□□□□基準を満たす □ 患者 □ □□

この重複確率 □ は各ラン中に実証的に測定される。□□□□ランからの結果（□□□サイト、各プロファイル□□サイト□ 最もバランスの取れた推定）：

サブグループ	□□□□□□□□□□□□□□重複□	真の条件付き□□□□ □□□
□□□□□□□□□□□□	□□□□□	□□□□□□ □□□分□年
□□□□□□□□□□□□	□□□□□	□□□□□□ □□□分□年

□□□□□□□□□□□□（□□□□ブーストサブグループ）□ hba1cBaseline > 8.0 かつ任意の利尿薬。明示的な□□□□埋め込 □□□分□年。

規模	サブグループ□	推定□□□（□□□）	真の□□□□効果	偏差
□□□万	□□□□□□□	□□□□□□	□□□□	□□□□□□（□□□□過小）
□億	□□□□□□□□□□	□□□□□□	□□□□	□□□□□□（□□□□過小）
□□億	□□□□□□□□□□	□□□□□□	□□□□	□□□□□□（□□□□過小）

□□□□□□□□□□□□（入院リスクサブグループ）□ raceAsian かつ □□□ □ □□ かつ □□□ □ □□。埋め込み効果は入院（□□□絶対リスク）に対するもので、□□□□ではない。真の□□□□ □□□：□□□□□□ □□□分□年（上記から導出）。

規模	サブグループ□	推定□□□（□□□）	真の□□□□ □□□	偏差
□□□万	□□□□□□	□□□□□□	□□□□□□	□□□□□□（□□□□□□）
□億	□□□□□□□□□	□□□□□□	□□□□□□	□□□□□□（□□□□□□）
□□億	□□□□□□□□□	□□□□□□	約□□□□□□	約□□□□□□ （約□□□□□）

□□□□□□□□□□□□（副作用サブグループ）□ egfrBaseline 30 45 かつ ループ利尿薬 かつ □□□ □ □□。埋め込み効果は副作用（□□□絶対リスク）に対するもので、□□□□ではない。真の□□□□ □□□：□□□□□□ □□□分□年（□□□□推定；□□推定は異なるサイトプロファイル分布のため□□□□□□）。

規模	サブグループ□	推定□□□（□□□）	真の□□□□ □□□	偏差
□□□万	□□□	□□□□□□	□□□□□□	□□□□□□（□□□□□□）
□億	□□□□□□	□□□□□□	□□□□□□	□□□□□□（□□□）
□□億	□□□□□□□	□□□□□□	約□□□□□□	約□□□□□□（約□□□□□）

□□□□年□月□□日に検証（□□□□□□□、ベースシード□□ □□□□□□□；サイト数□□□□□□、患者□サイト□□□□□□

解釈□ 真の条件付き□□□がデータ生成モデルから正確に決定されたことで、□つのサブグループにわたって一貫した状況が明らかにな

□□□□□□□□□□□□（真□□□□、明示的に埋め込み）：□□□□ □ □□□□ □ □□□□の過小推定（□□□万 □ □億 □ □□億）。改善は□□□万 □ □億に集中；□□億超では収穫逓減。

□□□□□□□□□□□□□（真□□□□□□、重複から導出）：□□□は□つの規模全てでほぼ不偏（□□□□□、□□□□□、約□□
見かけ上の誤ったキャリブレーション。真の値は□□□□□□である（□□□□□□□□□□□□□患者の□□□□□が□□□□□□□□□

□□□□□□□□□□□□（真□□□□□□、重複から導出）：劇的な規模依存の改善。□□□万人（□□□□□□）では、□□□推定値
□□□万規模での□□□患者サブグループにおける極端な重みの不安定性に起因する。□億人（□□□□□□□□□）では、推定値は本質的

サブグループ全体の要約□ 規模拡大は□□□精度を大幅に改善する。最大の利得は□□□□□□□□□□□□にあり、サブグループが□

再現コマンド：

```
npm run run:100m     # 1億患者 □ □□□□□□□□□□□□□□
```

□□ 既知の真値□ 埋め込み効果はプログラムされている□ 現実：真の効果は未知かつ検証不可能□ 影響：バイアス検出は既知の真値で

□□ サイトの均質性□ 同一のデータ分布□ 現実：サイトは人口統計、診療、機器で異なる□ 影響：連合集約は異質性の下で失敗する可

□□ 未測定交絡なし□ 全ての交絡因子が観測されている□ 現実：常に未測定交絡因子が存在する□ 影響：実際の応用では追加のバイア

□□ 単純化されたアウトカム□ 線形関係□ 現実：複雑な疾患進行、時間変動性交絡□ 影響：実際のアウトカムモデリングは大幅により

含意□ これらの単純化は実際の医療データとは根本的に異なる問題を表している。合成結果は技術的実現可能性を示すが、実世界のパフ

技術的限界：

- 単一マシンシミュレーションのみ□ この研究の全ての「連合」サイトは単一マシン上の並列□□□□□□スレッドとして実行され
  いずれもここでは測定されていない。通信サイズ（約□□□バイト□サイト）は理論的なプロトコルオーバーヘッドを表し、測定さ
- 市販ハードウェアのみ（クラウド□□□□ベンチマークなし）
- フォールトトレランスやセキュリティ分析なし
- □□□加速実装との比較なし
- 重みのトリミングや安定化なし□ □□□推定量は傾向重みを制限せず、傾向スコアのテールにおける極端な重み値に対して脆弱な
  □□□□。この合成実験では影響は小さいように見える□真の条件付き□□□がアウトカムモデルから正しく導出されると、全□

4.3 実データを必要とする未解決の問題

□ 実証的な検証なしには不明：

- 実際の薬物警戒システムがまれなサブグループで陽性度違反を経験するかどうか
- □□億規模が実際のまれなサブグループに十分かどうか（実際の有病率と交絡に依存）
- 合成データで観察されたバイアス解消が実際の異質なサイトに一般化するかどうか
- 連合アプローチが既存の中央集権的システムに対して実際の優位性を提供するかどうか
- □□□□統合が実際の観察データベースでシームレスに機能するかどうか

4.4 なぜオープンソースにするのか

主要な目標□ データアクセスを持つ研究者のための参照実装として動作するコードを共有する。

意図されたユーザー：

- □□□□形式のデータベースを持つ薬物警戒研究者
- 連合因果推論を探索する方法論研究者
- まれなサブグループ仮説をテストするヘルスケア情報科学者

期待すること：

□□ 実際の多施設データを持つ誰かがまれなサブグループで陽性度違反が起こるかどうかをテストする
□□ それらが存在すれば、既存の方法との比較評価
□□ 有用であれば、実装へのコミュニティによる改善
□□ 有用でなければ、否定的な結果にも科学的価値がある

現実的な期待□ ほとんどの研究者はこれを必要としないだろう。数人が有用と感じるかもしれない。それで十分である。

4.5 将来の研究（共同研究が必要）

実際のデータで□□億規模の知見を検証するには、以下が必要である：

□□ データアクセス□ 実際の□□□分析のための多機関契約
□□ □□□承認□ 観察研究のための倫理委員会の許可

□□ 比較基準□ 既存の薬物警戒方法へのアクセス
□□ 実世界での検証□ 規模でのバイアス解消が異質な実際のサイトに一般化するかテスト
□□ 方法の比較□ □□□□、マッチング、操作変数との比較 □□□
□□ 専門家との共同研究□ 疫学者と臨床医からのドメイン専門知識
□□ 重みの安定化□ トリミング□安定化された□□□重みの実装と評価（例：第□□第□□パーセンタイルでのトランケーション □□□□）；この合成設定では影響は小さいように見える（□億人以上で全サブグループが□□□以内）が、実世界の高次元交絡
□□ 実際の連合展開□ 実際のネットワーク境界にわたってプロトコルをテストし、実際の通信オーバーヘッド、遅延、機関的制約下で

私は訪問研究者や技術的貢献者としての共同研究に開放的である。

---

5. 結論

シングルパス連合集約は□□□□の総計算量（患者ごとの処理は一定）を達成し、□□コアで□億患者を□□□秒で、□□億患者を□□□

これらの知見は計算的実現可能性を確立し、合成データでの実装の正確性を検証する。この種の規模依存のバイアスパターンが実際の薬物

コード□ □□□□□

- 設定（有病率、埋め込み効果、サイトパラメータ）
- 連合推定値（□□□、□□、□□、□統計、□値）
- パフォーマンスメトリクス（時間、スループット、通信）

□□□結果：

- □□□万人□ `results/test_optimized_run/final_results.json`
- □億人□ `results/optimized_100million_run/final_results.json`
- □□億人□ `results/optimized_1billion_run/final_results.json`
- 多シード□ `results/multi_seed_test/summary.json`

ハードウェア要件

- □□□□ マルチコア（テスト済み：□□コア；`os.cpus().length`を自動使用）
- □□□□ 最低□□ □□（テスト済み：□□ □□、ピーク使用量 □□ □□）
- ストレージ□ コード □ 依存関係で約□□ □□、結果で約□□ □□
- □□□ □□□□□□□□□□□□□□□□□□□□□□で□□□□□□□ □□□□

期待される実行時間

- クイックテスト（□□□患者）：□秒未満
- 全層比較（□ □ □□□□万患者）：合計約□秒（ハードウェア依存；□□コアで測定）
- 多シードテスト（□ □ □□□□万患者）：合計□秒未満（`node src/multiSeedTest.js`）
- □億人ラン：約□秒（測定：□□コアで□□□秒）
- □□億人ラン：約□□□□分（測定：□□コアで□□□秒；コア数でスケール）

謝辞

以下に感謝する：

- 合成ジェネレーター設計に情報を提供した臨床データモデリングの概念を提供した□□□□□□□オープンソースコミュニティ
- □□□□ □□□標準とドキュメントの□□□□□コミュニティ
- 並列処理を可能にする□□□□□□スレッド□□□の□□□□□□□チーム

競合利益

なし。これは商業的利益のない自主的な独立研究である。

データとコードの可用性

ソースコード□ □□□□□□□□□□□□□□□□□□□□

- ジェネレーター□ packages/core/src/causal/omop-synthetic.ts
- □□□□ packages/cli/src/commands/causal/generate-omop-data.ts
- □□□□□ □□□□□コホート定義と互換性あり

実際の患者データなし□ この研究は合成データのみを使用する。□□□承認は必要ない。

再現可能性□ 全ての結果は上記の再現可能性セクションのコマンドで再現可能。期待される実行時間：層テストは約□秒；□億人ランは約

参考文献

□□ □□□□□□ □□ □□□□□□ □□ □□□□□□ □□□□□ □□□□□□□□□□□□□□ □□□□□□□□ □□□ □□□□□ □ □□□□□□□ □□□ □□□□□ □ □□□□□□□□□□□□□ □□□□□□□ □□□□□□□□ □□□□□ □□ □□□□□□□□□□□□□□□ □□□□□□□□□ □□□□□□□□□□□□□□□

□□□□□□□ □□□□□□□□□ □□□□□□□□ □□□ □□□ □□□□□□□□□ □□□□□ □□□□□□ □□□□□□ □□□□ □□□□□□□ □□□□□□ □□□□□□□□□

コンポーネント	説明	コンパクト（上三角）	現行実装（完全）
メタデータ（□□□□□□□□□□）	患者数	約□バイト	約□バイト
サイトあたり合計	シングルパスサイトあたり	約□□□バイト	約□□□バイト

現行実装は簡単のため完全な行列を送信する。上三角を使用するコンパクト符号化では約□□□バイトに削減できる。どちらも患者数によ

主要な利点□ 勾配を繰り返し交換する反復的連合学習□□□□□□□（典型的には□□□以上のラウンド）とは異なり、このアプローチ

□□□万患者、□□サイトの計算□

- 連合（コンパクト）□ 約□□□バイト□サイト □ □□サイト □ 約□□□ □□
- 連合（現行実装）□ 約□□□バイト□サイト □ □□サイト □ 約□□□ □□
- 中央集権的□ □□□万患者 □ □□□バイト□患者 □ □□□ □□
- 削減□ □□□ □□ □ □□□ □□ □ □□□□□□倍（コンパクト）または □□□ □□ □ □□□ □□ □ □□□□□□倍（現行）

S2. アルゴリズムの擬似コード

```
// 各サイト□はローカルで計算する（生データは共有しない）
function computeSiteStatistics(patients: Patient[], beta: number[]): SiteStats {
  const gradient = computeGradient(patients, beta); // 5つの□□□□□□□値
  const hessian = computeHessian(patients, beta);   // 5×5の□□□□□□□□（□□値；□□の上三角）
  const { XWX, XWY } = computeWeightedStats(patients, beta); // 6×6 + 6の□□□□□□□
  const metadata = { nPatients: patients.length };
  // ペイロード□ 約□□□バイト（完全行列）または約□□□バイト（コンパクト上三角符号化）
  return { gradient, hessian, XWX, XWY, metadata };
}

// 中央集約機□ メモリ制約のあるバッチでサイトを処理
function federatedCausalInference(sites: Site[]): CausalEffect {
  let beta = initialize();              // 傾向係数
  const allOutcomeStats: OutcomeStats[] = [];

  // 傾向□ バッチあたり□回の□□□□□□ステップ（追加のサイト通信なし）
  for (const batch of inBatchesOf(sites, BATCH_SIZE)) {
    const batchStats = batch.map(s => computeSiteStatistics(s.patients, beta));

    const G = sum(batchStats.map(s => s.gradient));
    const H = sum(batchStats.map(s => s.hessian));
    const delta = inv(H) * G;           // Newtonステップ
    beta = beta + delta;                //  が安定化（□□億人ランのバッチあたり□□□□□□□□□□□□□ □ □□□

    allOutcomeStats.push(...batchStats);
  }

  // アウトカム□ 全サイトにわたるシングルパス集約（全バッチにわたって収集）
  const XWX_total = sum(allOutcomeStats.map(s => s.XWX));
  const XWY_total = sum(allOutcomeStats.map(s => s.XWY));
```

```
  const theta = inv(XWX_total) * XWY_total;
  return { ate: theta[1] };
  // se: sqrt(inv(XWX_total)[1][1])  // testOptimized.jsではまだ実装されていない
}
```

S3. ハードウェア仕様

CPU: 24論理コア（□□□□□□□上の□□□□、□□□□□□□□□□□□□□□□による全コア利用）
□□□ □□□□□ □□
□□□□□□□□ □□□□

□□□□ □□□万人シミュレーションでピーク□□ □□；□□億人シミュレーションで□□ □□
ストレージ□ 結果に約□□ □□（生の患者データは保存されない；ストリーミング生成）

測定パフォーマンス（□□□□年□月□□日）：
□□□万患者（□□サイト）□ □□□□□秒、□□□□□□□□□ □□□□□、□□コア
□億患者（□□□サイト）□ □□□□□秒、□□□□□□□□□□ □□□□□、□□コア
□□億患者（□□□□□サイト）□ □□□秒、 □□□□□□□□□□ □□□□□、□□コア

著者情報

和智大次郎

独立研究者 □ オープンソースソフトウェアエンジニア

メール□ □□□□□□□□□□□□□□□□□□□□□□□□□□ □□□□□□□ □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□
所在地□ 日本

経歴□

- ヘルスケアデータツールを構築する□□□エンジニア
- 機関所属や学術的立場なし
- 自己資金による独立研究

現在の制約□

- 実際の多機関医療データへのアクセスなし
- ヒト被験者研究の□□□承認なし
- 比較のための独自薬物警戒システムへのアクセスなし

共同研究への開放性□ 訪問研究者ポジション、□□□□データベースアクセスを持つ共同プロジェクト、またはグラント助成研究ポジシ

S4. 10億人ランにわたる傾向スコア□の安定性

□□億人ランは□□□□□サイトを□□の□□バッチで処理する。各バッチの後、中央集約機は□を更新する□回の□□□□□□ステッ
□ 約□バッチ）での□を記録する（測定されたランから）。

処理済みサイト	□□（切片）	□□（年齢）	□□（□□□□）	□□（□□□□□□□）	（□□□□□）	□□□□□	□□□□□□
□□□	□□□□□□	□□□□□	□□□□□	□□□□□	□□□□□□	□□□□□	□
□□□	□□□□□□	□□□□□	□□□□□	□□□□□	□□□□□□	□□□□□	□□□□□