# Bayesian "skew-shrinkage" regression: tool for transfer learning across large health databases

Akihiko Nishimura
Department of Biostatistics

Johns Hopkins Bloomberg School of Public Health | Biostatistics

# Observational Health Data Sciences & Informatics

## Map of Collaborators

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

### OHDSI By The Numbers

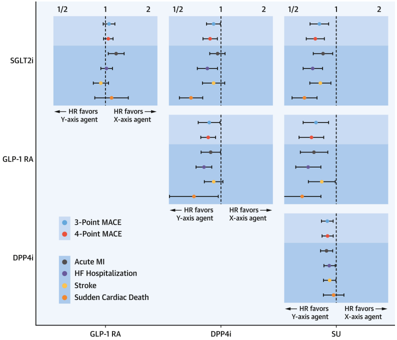
- 534 databases
- 956 million patient records
- 3,758 collaborators

https://ohdsi.org

# LEGEND initiative to bring together health data

"Large-scale Evidence GEneration via Network of Databases"

to compare second-line treatments for type-2 diabetes mellitus:





ELSEVIER

Journal of the American College of Cardiology

Volume 84, Issue 10, 3 September 2024, Pages 904-917

Original Research


## Comparative Effectiveness of Second-Line Antihyperglycemic Agents for Cardiovascular Outcomes: A Multinational, Federated Analysis of LEGEND-T2DM

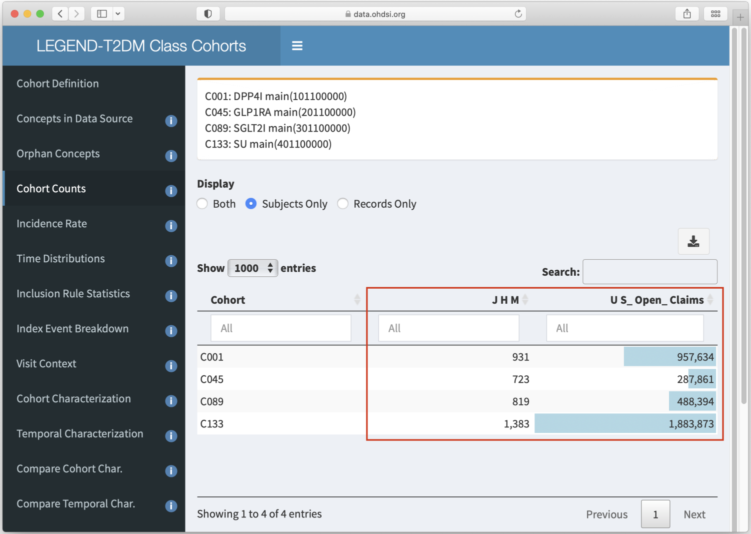

Rohan Khera MD, MS [a b c], Arya Aminorroaya MD, MPH [a], Lovedeep Singh Dhingra MBBS [a], Phyllis M. Thangaraj MD, PhD [a], Aline Pedroso Camargos PhD [a], Fan Bu PhD [d], Xiyu Ding MS [e], Akihiko Nishimura PhD [e], Tara V. Anand BS [f], Faaizah Arshad BS [g], Clair Blacketer MPH [h], Yi Chai PhD [i], Shounak Chattopadhyay PhD [g], Michael Cook BSc [e], David A. Dorr MD, MS [j], Talita Duarte-Salles PhD [k l], Scott L. DuVall PhD [m n], Thomas Falconer MS [f], Tina E. French RN, CPHQ [o p], Elizabeth E. Hanchrow RN, MSN [o p], Guneet Kaur MS [q], Wallis C.Y. Lau BSc, PhD [r s t u], Jing Li MS [v], Kelly Li BS [g], Yuntian Liu MPH [a b], Yuan Lu ScD [a], Kenneth K.C. Man BSc, MPH, PhD [r s t u], Michael E. Matheny MD, MS, MPH [o p], Nestoras Mathioudakis MD, MHS [w], Jody-Ann McLeggon MPH [f], Michael F. McLemore RN [o p], Evan Minty MD, MSc [x], Daniel R. Morales MD [q], Paul Nagy PhD [w], Anna Ostropolets MD, PhD [h], Andrea Pistillo MSc [k], Thanh-Phuc Phan MBA [y], Nicole Pratt PhD [z], Carlen Reyes MD, PhD [k], Lauren Richter MD [f], Joseph S. Ross MD, MHS [b aa], Elise Ruan MD [f], Sarah L. Seager BS [bb], Katherine R. Simon AA [o p], Benjamin Viernes PhD [m n], Jianxiao Yang MS [cc], Can Yin MS [dd], Seng Chan You MD, PhD [ee ff], Jin J. Zhou PhD [g gg], Patrick B. Ryan PhD [f], Martijn J. Schuemie PhD [hh], Harlan M. Krumholz MD, SM [a b ii], George Hripcsak MD, MS [f], Marc A. Suchard MD, PhD [g m jj kk]

# LEGEND initiative to bring together health data

**Table 1 | Description of databases from the Observational Health Data Sciences and Informatics network included in the study.**

| Name of database | Abbreviation | Country of origin | Years of exposure included | No of participants |
|---|---|---|---|---|
| **US national databases (claims data)** | | | | |
| IBM MarketScan Commercial Claims and Encounters Data | CCAE | USA | 2011-21 | 265 874 |
| IBM Health MarketScan Multi-State Medicaid Database | MDCD | USA | 2011-20 | 40 064 |
| IBM Health MarketScan Medicare Supplemental and Coordination of Benefits Database | MDCR | USA | 2011-21 | 43 857 |
| Optum Clinformatics Extended Data Mart - Date of Death | OCEDM | USA | 2011-21 | 211 877 |
| Optum de-identified Electronic Health Record Dataset | OEHR | USA | 2011-21 | 299 008 |
| US Open Claims | USOC | USA | 2000-21 | 3 521 191 |
| **US health system databases (electronic health record data)** | | | | |
| Columbia University Irving Medical Centre | CUIMC | USA | 2011-21 | 4561 |
| Johns Hopkins Medicine | JHM | USA | 2016-21 | 3759 |
| Stanford Medicine | STARR | USA | 2011-21 | 2993 |
| Department of Veterans Affairs Healthcare System | VA | USA | 2011-21 | 230 019 |
| **Non-US databases (electronic health record data)** | | | | |
| Australia Longitudinal Patient Database Practice Profile | ALPD | Australia | 2012-21 | 2322 |
| France Longitudinal Patient Database | FLPD | France | 2012-21 | 13 270 |
| Germany Disease Analyser | GDA | Germany | 1992-21 | 32 442 |
| Health Informatics Centre at the University of Dundee | HIC | Scotland | 2011-21 | 5580 |
| HKHA - Hong Kong Hospital Authority | HKHA | Hong Kong | 2011-18 | 4614 |
| UK-IQVIA Medical Research Data | IMRD | United Kingdom | 2011-19 | 25 173 |
| Information System for Research in Primary Care | SIDIAP | Spain | 2011-21 | 61 382 |

# Data from indiv health systems aren't big enough



# Transfer leaning from larger database to smaller one

**Idea:** Inform the model for a smaller “Database B” by transferring insights from the model trained on a larger “Database A.”

# Transfer leaning from larger database to smaller one

**Idea:** Inform the model for a smaller "Database B" by transferring insights from the model trained on a larger "Database A."

As an example, consider a linear model for both Database A and B:

$$\boldsymbol{y}^{(A)} = \boldsymbol{X}^{(A)}\boldsymbol{\beta}^{(A)} + \boldsymbol{\epsilon}^{(A)},$$
$$\boldsymbol{y}^{(B)} = \boldsymbol{X}^{(B)}\boldsymbol{\beta}^{(B)} + \boldsymbol{\epsilon}^{(B)}.$$

# Transfer leaning from larger database to smaller one

**Idea:** Inform the model for a smaller "Database B" by transferring insights from the model trained on a larger "Database A."

As an example, consider a linear model for both Database A and B:

$$\boldsymbol{y}^{(A)} = \boldsymbol{X}^{(A)}\boldsymbol{\beta}^{(A)} + \boldsymbol{\epsilon}^{(A)},$$
$$\boldsymbol{y}^{(B)} = \boldsymbol{X}^{(B)}\boldsymbol{\beta}^{(B)} + \boldsymbol{\epsilon}^{(B)}.$$

We expect $\boldsymbol{\beta}^{(A)}$ and $\boldsymbol{\beta}^{(B)}$ to be correlated; i.e. the value of $\boldsymbol{\beta}^{(A)}$, if known, provides information on $\boldsymbol{\beta}^{(B)}$:

$$\left(\boldsymbol{y}^{(A)}, \boldsymbol{X}^{(A)}\right) \Rightarrow \boldsymbol{\beta}^{(A)} \Rightarrow \boldsymbol{\beta}^{(B)}.$$

# Transfer leaning from larger database to smaller one

1) Obtain the posterior of $\boldsymbol{\beta}^{(A)} \mid \boldsymbol{y}^{(A)}, \boldsymbol{X}^{(A)}$;

# Transfer leaning from larger database to smaller one

1) Obtain the posterior of $\boldsymbol{\beta}^{(A)} \mid \boldsymbol{y}^{(A)}, \boldsymbol{X}^{(A)}$;

2) Calculate the informed mean $\mu_j^{(B \mid A)}$ and std deviation $\sigma_j^{(B \mid A)}$ for $\beta_j^{(B)}$ according to the assumed correlation structure;

# Transfer leaning from larger database to smaller one

1) Obtain the posterior of $\boldsymbol{\beta}^{(A)} \mid \boldsymbol{y}^{(A)}, \boldsymbol{X}^{(A)}$;
2) Calculate the informed mean $\mu_j^{(B \mid A)}$ and std deviation $\sigma_j^{(B \mid A)}$ for $\beta_j^{(B)}$ according to the assumed correlation structure;
3) Train the model B under prior $\beta_j^{(B)} \sim \mathcal{N}\left(\mu_j^{(B \mid A)}, \sigma_j^{2(B \mid A)}\right)$.

# High-dim, data-driven prediction/causal inference

**Table:** Counts of covariates within each OMOP concept domains.

| **Domains** | **Counts** | |
|---|---|---|
| | Hopkins | PharMetrics |
| Condition | 5,170 | 10,358 |
| Drug | 1,685 | 2,118 |
| Measurement | 1,334 | 940 |
| Procedure | 1,137 | 4,479 |
| Observation | 359 | 876 |
| Device | 105 | 1,194 |
| Race | 6 | 0 |
| Gender | 1 | 1 |
| Ethnicity | 1 | 0 |
| Overall | 9,798 | 19,967 |

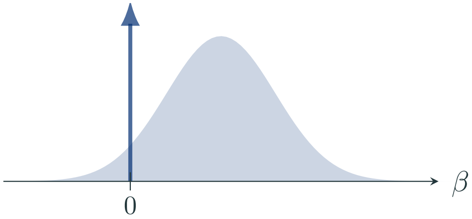
# Skew-shrinkage for high-dim transfer learning

# Skew-shrinkage for high-dim transfer learning

Consider combining an informed prior with shrinkage by setting

$$\beta_{j,\,\mathrm{sh}}^{(B)} = \delta_j \beta_j^{(B)}$$

where $\delta_j = 0$ with probability $p \in [0, 1]$ and $\delta_j = 1$ otherwise.

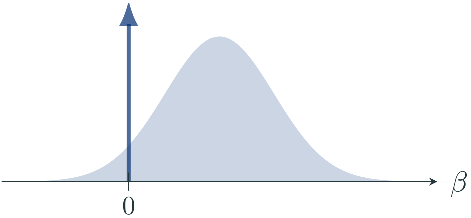


# Skew-shrinkage for high-dim transfer learning

Consider combining an informed prior with shrinkage by setting

$$\beta_{j,\mathrm{sh}}^{(B)} = \delta_j \beta_j^{(B)}$$

where $\delta_j = 0$ with probability $p \in [0, 1]$ and $\delta_j = 1$ otherwise.

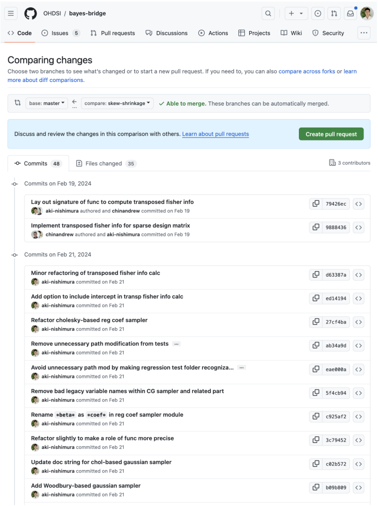


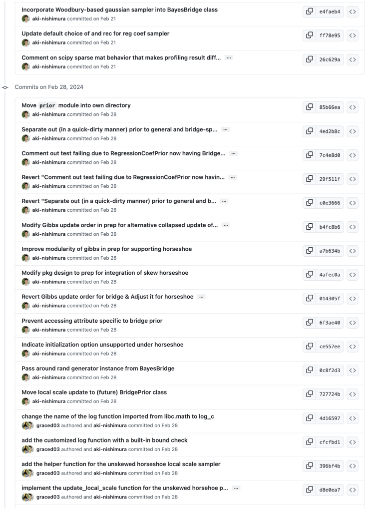

(For computational efficiency, we use a continuous analogue.)

# Demo: new skew-shrinkage feature in BayesBridge

```
skew_mean = np.array([0.1, 1.0, -0.3])
skew_sd = np.array([1.0, 0.1, 0.5])

skew_prior = HorseshoePrior(
  skew_mean=skew_mean, # New!
  skew_sd=skew_sd, # New!
  regularizing_slab_size=1.
)

linear_model = RegressionModel(y, X, family='linear')
bridge = BayesBridge(linear_model, skew_prior)

post_samples, _ = bridge.gibbs(
  n_iter=1000, init={'global_scale': .01}
)
```

# Simple interface change, hard internal work




⋮

⋮

## Application: Hopkins EHR meets LEGEND-T2DM

**Goal:** Compare four classes of second-line T2DM treatment for their cardio-vascular effectiveness and safety.

Here we focus on GLP-1 receptor agonists and DPP-4 inhibitors.

# Application: Hopkins EHR meets LEGEND-T2DM

**Goal:** Compare four classes of second-line T2DM treatment for their cardio-vascular effectiveness and safety.

Here we focus on GLP-1 receptor agonists and DPP-4 inhibitors.

**Data:** IQVIA PharMetrics (source) and Hopkins EHR (destination)

- DPP-4 users: 10,203 in PharMetrics and 1,003 in Hopkins
- GLP-1 users: 9,220 in PharMetrics and 1,032 in Hopkins

# Result of “internal” transfer within IQVIA data

# Result of “internal” transfer within IQVIA data

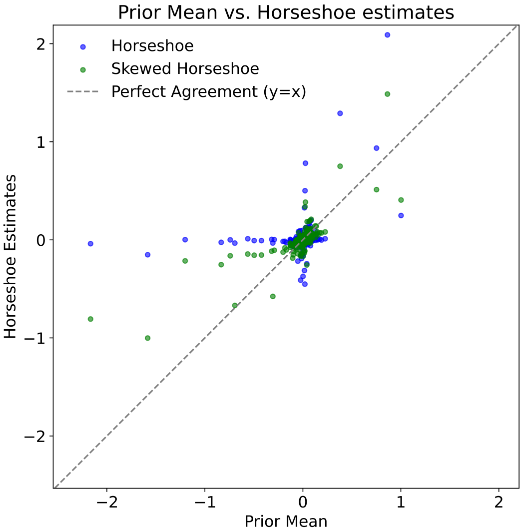
We used 80% of the IQVIA data as “Database A,” 10% as “Database B,” and the rest for calculating out-of-sample AUC:

# Result of “internal” transfer within IQVIA data

We used 80% of the IQVIA data as “Database A,” 10% as “Database B,” and the rest for calculating out-of-sample AUC:

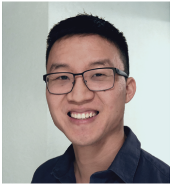
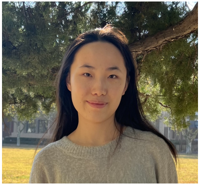
| Data fraction (sample size) | W/o transfer | With transfer |
|---|---|---|
| 10% ($n_B = 1{,}942$) | 0.766 | 0.773 |
| 3% ($n_B = 587$) | 0.724 | 0.747 |
| 2% ($n_B = 387$) | 0.701 | 0.745 |
| 1% ($n_B = 193$) | 0.634 | 0.747 |

# Comparison of estimates w/o vs. with transfer



# Acknowledgments

In collaboration with


and with  (lead), , .