

# 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



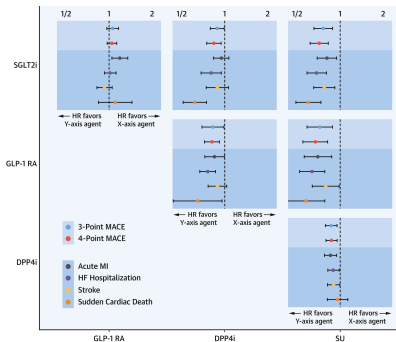
### OHDSI By The Numbers

- 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:



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

# 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
<b>US national databases (claims data)</b>				
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
<b>US health system databases (electronic health record data)</b>				
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
<b>Non-US databases (electronic health record data)</b>				
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

LEGEND-T2DM Class Cohorts

Cohort Definition

Concepts in Data Source

Orphan Concepts

Cohort Counts

Incidence Rate

Time Distributions

Inclusion Rule Statistics

Index Event Breakdown

Visit Context

Cohort Characterization

Temporal Characterization

Compare Cohort Char.

Compare Temporal Char.

C001: DPP4I main(101100000)  
C045: GLP1RA main(201100000)  
C089: SGLT2I main(301100000)  
C133: SU main(401100000)

Display  
 Both  Subjects Only  Records Only

Show 1000 entries Search:

Cohort	J H M	U S_ Open_Claims
All	All	All
C001	931	957,634
C045	723	287,861
C089	819	488,394
C133	1,383	1,883,873

Showing 1 to 4 of 4 entries Previous 1 Next

# Transfer learning 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 learning 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:

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

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

# Transfer learning 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:

$$\begin{aligned}\mathbf{y}^{(A)} &= \mathbf{X}^{(A)}\boldsymbol{\beta}^{(A)} + \boldsymbol{\epsilon}^{(A)}, \\ \mathbf{y}^{(B)} &= \mathbf{X}^{(B)}\boldsymbol{\beta}^{(B)} + \boldsymbol{\epsilon}^{(B)}.\end{aligned}$$

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  $\boldsymbol{\beta}^{(B)}$ :

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



## Transfer learning from larger database to smaller one

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

# Transfer learning from larger database to smaller one

- 1) Obtain the posterior of  $\beta^{(A)} \mid \mathbf{y}^{(A)}, \mathbf{X}^{(A)}$ ;
- 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 learning from larger database to smaller one

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

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

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

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

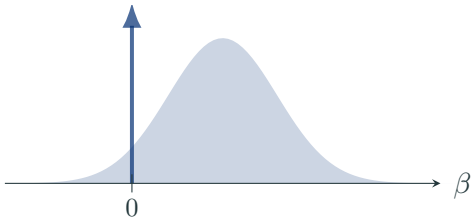
# Skew-shrinkage for high-dim transfer learning

# Skew-shrinkage for high-dim transfer learning

Consider combining an informed prior with shrinkage by setting

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

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

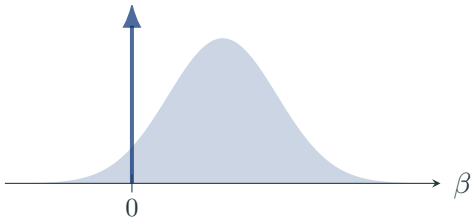


# Skew-shrinkage for high-dim transfer learning

Consider combining an informed prior with shrinkage by setting

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

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

```
1 skew_mean = np.array([0.1, 1.0, -0.3])
2 skew_sd = np.array([1.0, 0.1, 0.5])
3
4 skew_prior = HorseshoePrior(
5     skew_mean=skew_mean, # New!
6     skew_sd=skew_sd, # New!
7     regularizing_slab_size=1.
8 )
9
10 linear_model = RegressionModel(y, X, family='linear')
11 bridge = BayesBridge(linear_model, skew_prior)
12
13 post_samples, _ = bridge.gibbs(
14     n_iter=1000, init={'global_scale': .01}
15 )
```



# Simple interface change, hard internal work

The screenshot shows the GitHub pull request interface for the 'bays-bridge' repository. At the top, there are navigation tabs for Code, Issues, Pull requests, Discussions, Actions, Projects, Wiki, and Security. Below this is a section titled 'Comparing changes' with a sub-header 'base: master' and a dropdown menu showing 'compare: skew-shrinkage'. A green button labeled 'Create pull request' is visible. The main content area is titled 'Commits on Feb 19, 2024' and lists several commits with their titles, authors, and commit hashes. The commits include: 'Lay out signature of func to compute transposed fisher info', 'Implement transposed fisher info for sparse design matrix', 'Minor refactoring of transposed fisher info calc', 'Add option to include intercept in transp fisher info calc', 'Refactor cholesky-based reg coef sampler', 'Remove unnecessary path modification from tests', 'Avoid unnecessary path mod by making regression test folder recogniza...', 'Remove bad legacy variable names within CD sampler and related part', 'Rename ~beta~ as ~coef~ in reg coef sampler module', 'Refactor slightly to make a role of func more precise', 'Update doc string for chol-based gaussian sampler', and 'Add Woodbury-based gaussian sampler'.

The screenshot shows a list of commits for the 'bays-bridge' repository, sorted by date. The commits are listed in a table-like format with columns for the commit title, author, and commit hash. The commits include: 'Incorporate Woodbury-based gaussian sampler into BaysBridge class', 'Update default choice of and rec for reg coef sampler', 'Comment on scipy sparse mat behavior that makes profiling result diff...', 'Move priors module into own directory', 'Separate out (in a quick-dirty manner) prior to general and bridge-sp...', 'Comment on test failing due to RegressionCofPrior now having Bridge...', 'Revert "Comment on test failing due to RegressionCofPrior now havin...', 'Revert "Separate out (in a quick-dirty manner) prior to general and b...', 'Modify Gibbs update order in prep for alternative collapsed update of...', 'Improve modularity of gibbs in prep for supporting horseshoe', 'Modify plg design to prep for integration of skew horseshoe', 'Revert Gibbs update order for bridge & Adjust it for horseshoe', 'Prevent accessing attribute specific to bridge prior', 'Indicate initialization option unsupported under horseshoe', 'Pass around rand generator instance from BaysBridge', 'Move local scale update to (future) BridgePrior class', 'change the name of the log function imported from libc.math to log\_c', 'add the customized log function with a built-in bound check', 'add the helper function for the unskewed horseshoe local scale sampler', and 'implement the update\_local\_scale function for the unskewed horseshoe p...'. The commits are dated 'Commits on Feb 20, 2024'.

## 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

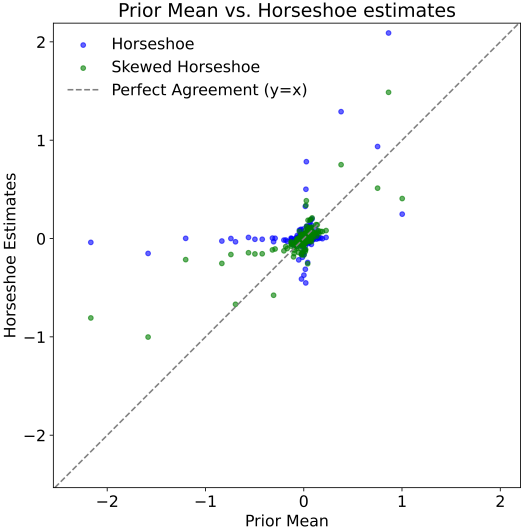
We used 80% of the IQVIA data as “Database A,” 10% as “Database B,” and the rest for calculating out-of-sample AUC:

## 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),

