# Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-level Information from External Data

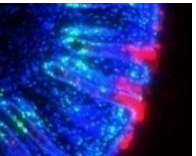
Raymond J. Carroll

**Department of Statistics**

**Institute for Applied Mathematics and Computational Science**

Texas A&M University

**http://stat.tamu.edu/~carroll**



STATISTICS
TEXAS A&M UNIVERSITY

# Acknowledgment


- Based on a paper by Nilanjan Chatterjee (JHU), Yi-Hau Chen (Academic Sinica, Taiwan), Paige Maas (NCI) and RJC, in **JASA**



# Acknowledgment

- Both are powerhouses, and I have been fortunate to write 16 papers with Nilanjan since 2005, 5 of which include Yi-Hau.

- They are astonishingly smart

# Basic Idea

- Increasingly, there are very large "representative" data sets that provide summary statistics but whose **individual data are not accessible**
- We call this the **External Data**
- Generically, we will call the external data Y and X
- **Y** is a response
- **X** is a set of covariates

# Basic Idea

- **Y** is a response
- **X** is a set of covariates
- In some cases, another covariate **Z** is observed
- Examples include the SEER cancer registry and many of the large consortia doing genomics, but the problem is more general

# Basic Idea

- How to use very large external data sets that provide only summary level information is an increasingly relevant problem.

- Individual external data may not be available due to factors such as
  - Storage and computing
  - Ethical reasons, e.g. privacy of study subjects
  - Protection of future research interests of data generating institutions and investigators

# Basic Idea

- For simplicity of presentation, I will also assume that the external study is so large that any distribution/parameters can be treated as know

- We have extended the results to the case of smaller external studies

# Basic Idea

- Individual investigators may then do a study where individual-level data are **accessible**
- Plus, if **Z** is not available in the external study, it is also measured
- For example, in breast cancer risk prediction, Z might be mammographic density, which has not been measured in large cohorts
- We call this the **Internal Study**

# Basic Idea

- Two main points of calibration to an external model and data set

- **Increasing efficiency**: Clear, since we will use the information in an external study not available to an internal study

- **Improve generalizability**: probably more important than efficiency, since we want results to apply to the representative external study

# Basic Idea

- There are two main cases

- **Case 1**: Where (X,Z) is measured in the external population, but summary information is provided for only a simple, crude risk model, e.g., one without interactions

- **Case 2**: Where only X is measured in the external study, but Z is also measured internally

# Basic Idea

- We work initially in the context that the external study does not have individual-level data

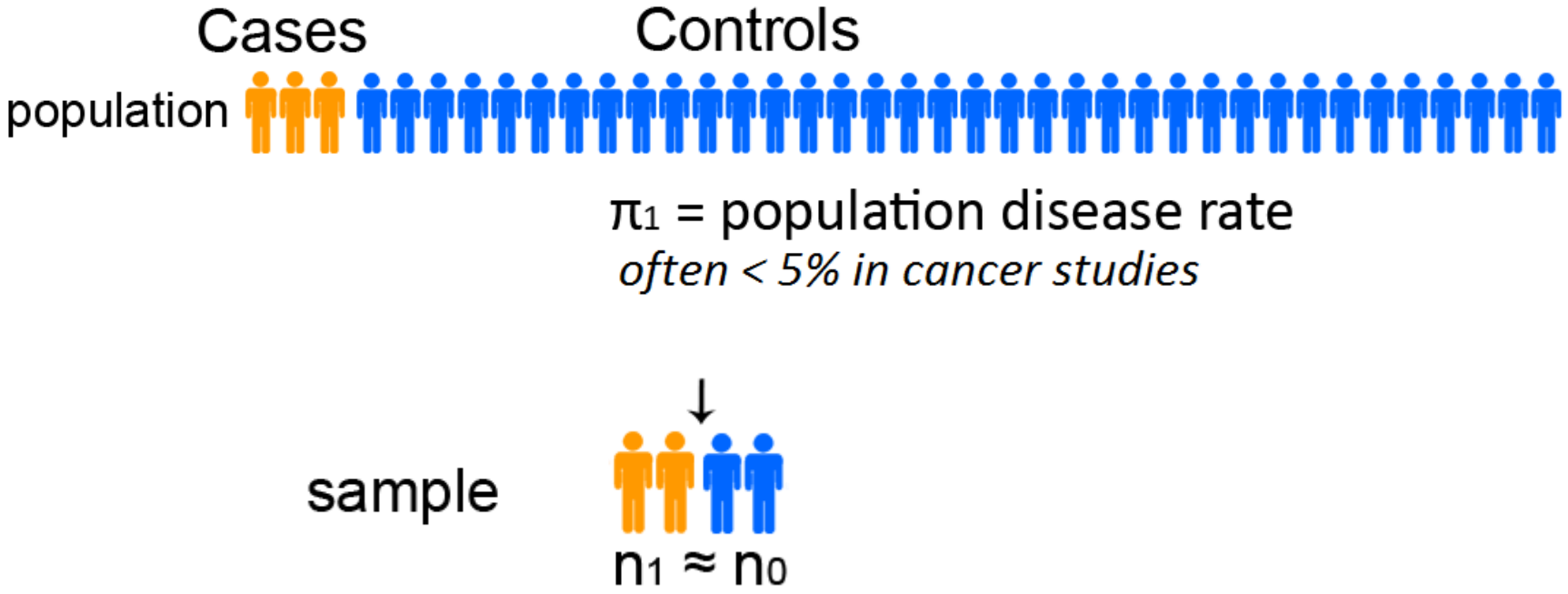
- The subjects in the internal study may or may not be part of the external data

# Basic Idea

- Also, **crucially and perhaps even typically**, the internal data generally may come from a (hopefully population based) case-control study of a large cohort

# Basic Idea

- Case-control studies are special, biased samples
- Developing methods for them in this context is technically complex



# An Aside: Case-Control Studies

- In case-control studies, a random sample of people with a disease (the cases) is taken, and a random sample of those without the disease is taken (the controls)

- There are subtle differences in the likelihood functions for case-control and simple random samples

# Basic Idea

- **Goal**: When and how can we use the summary-level information in the external study to improve inferences based on the individual data in the internal study?

- We seek simple, computationally scalable methods

# Basic Idea

- We specifically do not work in the two stage context where **both**
  - The internal data are a subset of the external data
  - Individual level data are available in the external data

- That problem has a massive and sophisticated literature

# Empirical Likelihood

- Simple random sampling: our method is a type of empirical likelihood method, and in the case I discuss it is asymptotically equivalent to it.

- Of course, the idea of using the external data in the way we do it is new.

- The likelihood for **case-control** sampling has crucial differences from that for simple random sampling, and our solution is new.

# Framework

- The density of Y given (X,Z) is $f(y|x,z,\beta)$ and the true value is $f(y|x,z,\beta_0)$

- **Base Assumption**: the density of Y given (X,Z) is the same in the external population and in the population from which the internal study is drawn

- See later for a comment about when this assumption is violated

# External Framework

- The density of Y given X is **posited** as $g(y|x,\theta)$ , and the limit value of the posited MLE is $\theta_{ext}$

- **Critical Point**: The **posited** model might differ from the true model for Y given X, denoted as p(y|x).

- For example, $f(y|x,z,\beta)$ might be a logistic model and $g(y|x,\theta)$ might be posited also as logistic, but generally the latter is misspecified if the former is true

# Framework

- **Case 1**: The distribution of (X,Z), F(X,Z), is the same in the internal and external studies
- We will deal with the alternative later
- Our basic approach is to use the external data to impose a set of constraints on $\beta_0$, the parameter in the internal data

# Framework

- Let $S_g(y|x,\theta)$ be the loglikelihood score/estimating function for the posited model, e.g., the derivative of the posited loglikelihood for $g(y|x,\theta)$

- Then, $0 = E\left\{S_g(Y|X,\theta_{ext})\right\}$, even if this model is incorrect

- This is the basic **constraint**

# Framework

- Define $U(x,z,\beta,\theta) = \int S_g(y|x,\theta) f(y|x,z,\beta) dy$
- By algebraic manipulation

$$0 = \int U(x,z,\beta_0,\theta_{ext}) dF(x,z)$$

- **Crucially**, this forms a constraint on $\beta$ that is the same dimension as $\theta$

# Constraint

- For today we deal with the case that the external data are so large that $\theta_{ext}$ is known
- Constraint $0 = \int U(x,z,\beta,\theta_{ext})dF(x,z)$
- The outline is clear now
- Maximize the loglikelihood in the **internal** data subject to the constraint from the **external** data
- We take a Lagrange approach

# Constraint

- Can also use a quasilikelihood or any other criterion function

- I will use the terminology loglikelihood, but it applies more generally.

# Constraint

- The loglikelihood of the internal data is $L_{int}(\beta,F)$
- Our proposal is to maximize

$$L_{int}(\beta,F)+\lambda\int U(x,z,\beta,\theta_{ext})dF(x,z)$$

- We treat $(\beta,\lambda)$ as unknown parameters
- In addition, the cdf F(x,z) is **unknown**.
- Remember, we might not observe Z in the external data

# Estimation

- Maximize $L_{int}(\beta,F)+\lambda\int U(x,z,\beta,\theta_{ext})dF(x,z)$

- To deal with the fact that F(x,z) is unknown, we use a semiparametric profile approach

- The idea is to say that (X,Z) has its support at their observed values in the **internal** study, with masses $(\delta_1,\ldots,\delta_n)$, subject to the constraint

$$1=\sum_{i=1}^{n}\delta_i$$

# Estimation

- Thus, we maximize, in $(\beta,\lambda,\gamma,\delta_1,\ldots,\delta_n)$,

$$L_{int}(\beta,F)+\sum_{i=1}^{n}\log(\delta_i)$$
$$+\lambda\sum_{i=1}^{n}\delta_i U(X_i,Z_i,\beta,\theta_{ext})$$
$$+\gamma\left(\sum_{i=1}^{n}\delta_i-1\right)$$

- Terms are
  - Loglikelihood of the internal data
  - Loglikelihood of the masses
  - Constraint from the external data
  - Constraint so that $0=\sum_{i=1}^{n}\delta_i-1$

# Estimation

$$L_{int}(\beta,F)+\sum_{i=1}^{n}\log(\delta_i)$$
$$+\lambda\sum_{i=1}^{n}\delta_i U(X_i,Z_i,\beta,\theta_{ext})+\gamma\left(\sum_{i=1}^{n}\delta_i-1\right)$$

- If the internal data are a simple random sample $(\gamma,\delta_1,\ldots,\delta_n)$ have explicit solutions given the data and $(\beta,\lambda)$

- Thus, we end up needing to maximize a function of the form $G(\beta,\lambda)$
- **Details avoided!**

# Estimation

- If the internal data are a simple random sample we need to maximize $G(\beta,\lambda)$

- The formulae for $G(\beta,\lambda)$ in these cases are messy, but it is easy to derive the score and the Hessian, and then do the maximization

- Besides identifying the constraints, we have extended the work to case-control studies. Technically much more complex.

# Estimation

- In both cases, the function to maximize is
  - Explicit
  - Maximization of the same dimension of $\beta$
  - Does not depend on explicit estimation of F(X,Z)
-

# Asymptotic Theory

- We have a complete asymptotic theory with explicitly computable asymptotic variances.

- The estimate of $\beta_0$ is asymptotically independent of the estimate of $\lambda$

- Not surprisingly, since our method is a constrained semiparametric MLE, it is asymptotically more efficient than using the internal data alone. This has been shown explicitly

# Case-Control Studies

- Many (most?) of the internal data sets are from case-control studies, because the diseases are usually rare

- There is an old result of Prentice & Pyke (1979, *Biometrika*) that for an ordinary CC study, with no constraints and with a logistic regression risk model, ordinary logistic regression can be used

# Case-Control Studies

- Our setting is different.
- The SRS result described above is not true even for a logistic risk model
- We came up with a new profiled likelihood, which is explicit, with an asymptotic theory, etc.

# Simulations

- **Case 1**: Where (X,Z) is measured in the external population, but summary information is provided for only a simple, crude risk model, e.g., one without interactions

- **Cases 2**: Where only X is measured in the external study

- Simulations done for simple random sampling and case-control sampling, with logistic regression models

# Simulations

- Binary Y, bivariate standard normal with correlation 0.3 for (X,Z)

- The full model in the internal study is

$$\text{logit}(Y|X,Z)=\beta_0+X\beta_X+Z\beta_Z+XZ\beta_{XZ}$$

- The relative risks for the main effects were ~1.50 and for the interaction ~1.25

# Simulations

- In these simulations, using logistic regression alone in the internal data is a legitimate approach, even for case-control studies **Prentice and Pyke**, 1979

- As expected coverage probabilities achieve their nominal level

- For the constrained maximum likelihood, the estimators also achieve nominal coverage

# Simulations: Underspecified Model

- Internal study $\text{logit}(Y|X,Z)=\beta_0+X\beta_X+Z\beta_Z+XZ\beta_{XZ}$
- External study $\text{logit}(Y|X,Z)=\theta_0+X\theta_X+Z\theta_Z$
- The relative risks for the main effects were ~1.50 and for the interaction ~1.25
- For $(\beta_X,\beta_Z)$ we see enormous gains in efficiency for the constrained MLE, sometimes MSE efficiency > 20
- Not much for the interaction $\beta_{XZ}$, which makes sense

# Simulations: Missing Z

- Internal study $\text{logit}(Y|X,Z)=\beta_0+X\beta_X+Z\beta_Z+XZ\beta_{XZ}$
- External study $\text{logit}(Y|X,Z)=\theta_0+X\theta_X$
- The relative risks for the main effects were ~1.50 and for the interaction ~1.25
- For $\beta_X$ MSE efficiency ~ 6.0 for the constrained fit
- Not much gain for $\beta_X$ or $\beta_{XZ}$

# Simulations: Measurement Errors

- Binary Y, bivariate standard normal with correlation 0.3 for (X,Z).
- Y is independent of X given Z (called nondifferential measurement error)
- Internal study $\text{logit}(Y|X,Z)=\beta_0+Z\beta_Z$
- External study $\text{logit}(Y|X,Z)=\theta_0+X\theta_X$
- MSE efficiency gain ~ 3.0

# Distributions of (X,Z) Different

- Suppose that the distributions of (X,Z) in the internal study, $F_{int}(x,z)$, is different from that of the external study, $F_{ext}(x,z)$

- What are the consequences of using constrained likelihood?

- What are the solutions?

# Distributions of (X,Z) Different

- What are the consequences of using constrained likelihood?
- Since the constraints in the external study are not the same as the constraints in the internal study, potential for bias.
- We have seen this in simulations. The constraints of course have to be pretty badly violated to make anything matter

# Distributions of (X,Z) Different

- What can be done?

- Consistent and more efficient estimation is possible if the external study summary information provides information about its distribution function $F_{ext}(x,z)$

- This is relevant for our case of the **underspecified model**, where (X,Z) is obtained in the external model but summaries are only available for a crude risk model

# Distributions of (X,Z) Different

- Assume that $F_{ext}(x,z)$ is known: the case that it is estimated can also be handled
- It is sometimes possible to get a **subsample** of the external covariates
- Hence, the original study cannot be scooped, etc.

# Distributions of (X,Z) Different

- Assume that $F_{ext}(x,z)$ is known: the case that it is estimated can also be handled

- The original approach was to maximize

$$L_{int}(\beta,F)+\lambda\int U(x,z,\beta,\theta_{ext})dF(x,z)$$

- We maximize some version of

$$L_{int}(\beta,F)+\lambda\int U(x,z,\beta,\theta_{ext})dF_{ext}(x,z)$$

# Synthetic MLE

- We maximize $L_{int}(\beta, F_{int}) + \lambda \int U(x, z, \beta, \theta_{ext}) dF_{ext}(x, z)$
- If the internal study uses **simple random sampling**, this is just a constrained loglikelihood in the parameters $(\beta, \lambda)$
- Technically, $F_{int}(x, z)$ factors out of the loglikelihood and can be ignored

# Synthetic MLE

- If the internal study is a case-control sample, things are more complex
- The loglikelihood of a case-control study is

$$\begin{aligned}
&\mathrm{pr}(X=x,Z=x|Y=y) \\
&= \frac{f(y|x,z,\beta)F_{int}(x,z)}{\mathrm{pr}(Y=y)} \\
&= \frac{f(y|x,z,\beta)F_{int}(x,z)}{\int f(y|x_*,z_*,\beta)dF_{int}(x_*,z_*)}
\end{aligned}$$

# Synthetic MLE

- Let the sample sizes be $N_0$ controls and $N_1$ cases
- Let $\mu_y = N_y / pr(Y=y)$
- From Prentice and Pyke (1979), the profiled likelihood of a case-control study is known to be

$$f_{profile}(Y|X,Z,\beta,\mu_1) = \frac{\mu_Y f(Y|X,Z,\beta)}{\sum_{y=0}^{1} \mu_y f(y|X,Z,\beta)}$$

- This is easily seen from the Lagrange argument as well

# Synthetic MLE

- Let $\beta_{cc}=(\beta,\mu_1)$
- Then we maximize

$$\sum_{i=1}^{n}\log\left\{f_{profile}(Y_i|X_i,Z_i,\beta_{cc})\right\}+\lambda\int U_{cc}(x,z,\beta_{cc},\theta_{ext})dF_{ext}(x,z)$$

- Once again, an explicit solution, and for a case-control study

# Synthetic MLE

- In the earlier simulation, constrained MLE ignoring the difference between the internal and external differences between the distribution of (X,Z), we saw bias and lack of coverage for $\beta_X$
- Synthetic MLE removed the bias and made the coverage match the nominal level
- This matched the asymptotic theory, which was also derived

# Example

- We have examples that basically repeat what the simulations indicate
- Here we take a different example where the assumption that the risk models are the same in the two populations may be (is) violated
- We believe that the constrained likelihood will indicate this, and shrink some of the coefficients towards their correct values.

# Illustration

- The **external** data are from the **Breast and Prostate Cancer Cohort Consortium (BPC3)**

- BPC3 is based on 10 large prospective North American cohorts, and thus might be considered "representative"

- There are multiple variables, but I focus on (a) number of 1$^{st}$ degree relatives with breast cancer and (b) whether age at menarche < 12

# Illustration

- The **internal** data are from the **Breast Cancer Detection and Demonstration Project (BCDDP)**

- BCDDP is a much smaller cohort then BPC3. It is also from women screened 1973-1978

- It has the same risk factors as BPC3, but it also measured mammographic density on a **case-control** subset of size 2817.

# Example

- The results were surprising

| | Internal | | Constrained | |
|---|---|---|---|---|
| | Estimate | s.e. | Estimate | s.e. |
| Age menarche < 12 | **0.50** | 0.12 | **0.17** | 0.04 |
| | | | | |
| Num Relatives | **0.65** | 0.09 | **0.30** | 0.03 |
| | | | | |
| MD | **0.43** | 0.04 | **0.44** | 0.05 |

- The shrinkage of the 2 covariates towards 0 is more than can be described by chance

# Example

- Here are the results without mammographic density, logistic regression

| | BCDDP | | BCP3 | |
|---|---|---|---|---|
| | Estimate | s.e. | Estimate | s.e. |
| Age menarche < 12 | **0.38** | 0.12 | **0.08** | 0.03 |
| | | | | |
| Family History | **0.72** | 0.10 | **0.35** | 0.03 |

- We see the same shrinkage, suggesting that (a) risk models may be different; and (b) CML may be getting closer to the true results in BCP3

# Example

- This example serves as a cautionary (?) note about the use of large external data sets
- In our case, BCP3, a large combination of cohort studies of Caucasians in the U.S., is clearly (?) more representative of the U.S. population.
- In building a risk model using MD, it is important to notice that the internal study may not be representative or population based

# Conclusions

- How to use very large external data sets that provide only summary level information is an increasingly relevant problem.

- Individual external data may not be available due to factors such as
  - Storage and computing
  - Ethical reasons, e.g. privacy of study subjects
  - Protection of future research interests of data generating institutions and investigators

# Conclusions

- We have described two ways to use external summary level information and the assumptions needed to exploit it
- Examples included using internal data to
  - Refine crude models
  - Exploiting a new variable not in the external data set for improved risk prediction
- The idea of summary-level external data leads to a host of interesting new questions about how to exploit it