Meta-analysis of case-control studies

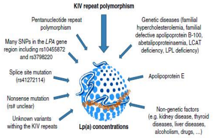
Georgina Bermann

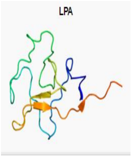
Novartis Pharma AG

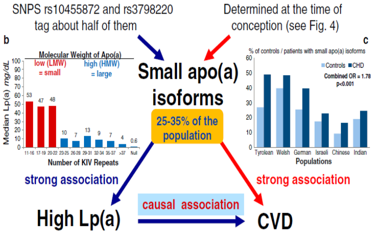
November 21, 2016

Exposure of interest

- Several meta-analyses have identified Lp(a) as a risk factor for cardiovascular diseases, among them 2 different IPD meta-analyses from the *Emerging Risk Factors Collaboration*
- In addition a meta-analysis of genotyped populations using a mendelian randomization approach has identified a causal relation between Lp(a) concentrations and CV events (MI, stroke, CV death)
- The magnitude of the risk appears to be related to concentration, but confounded by quality of assay
- Due to the molecular structure, concentration assays require standardization with respect to the number of *kringle repeats*


KIV repeat polymorphism
Pentanucleotide repeat polymorphism
Genetic diseases (familial hypercholesterolemia, familial defective apolipoprotein B-100, abetalipoproteinaemia, LCAT deficiency, LPL deficiency)
Many SNPs in the *LPA* gene region including rs10455872 and rs3798220
Splice site mutation (rs41272114)
Apolipoprotein E
Nonsense mutation (rs# unclear)
Unknown variants withing the KIV repeats
Lp(a) concentrations
Non-genetic factors (e.g. kidney disease, thyroid diseases, liver diseases, alcoholism, drugs, ...)





SNPS rs10455872 and rs3798220 tag about half of them

Determined at the time of conception (see Fig. 4)

Small apo(a) isoforms

25-35% of the population

strong association

strong association

High Lp(a)

causal association

CVD

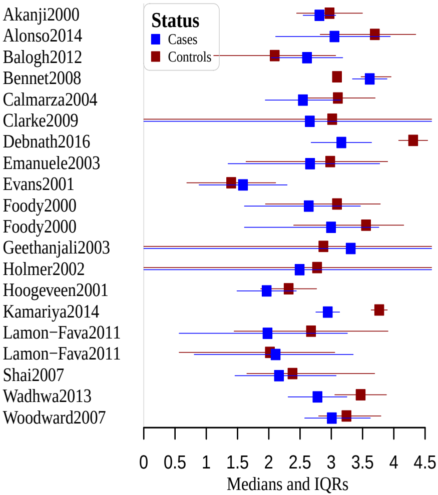
a Lp(a) concentrations

Percentile	mg/dL	HR	95%CI
>95th	>117	2.6	1.6-4.1
90th-95th	77-117	1.9	1.2-3.0
67th-89th	30-76	1.6	1.1-1.2
22nd-66th	5-29	1.2	0.9-1.6
<22nd	<5	reference	

Sampling frame

- The purpose of the data collection was to derive an overall estimate of the distribution of Lp(a) in cases with *Coronary arterial disease*
- Specifically the 3$^{\text{rd}}$ Quartile was of interest
- Due to issues with assay standardization, our sample had to fulfill the following requirements
  - *Early studies lacked proper asssay standardization, making them less reliable, hence study dates were limited to studies performed from 2000*
  - *The cases had to present the disease of interest*
  - *The sample was defined a sample of cases-control studies*
  - *This was later relaxed by including nested case-control studies*

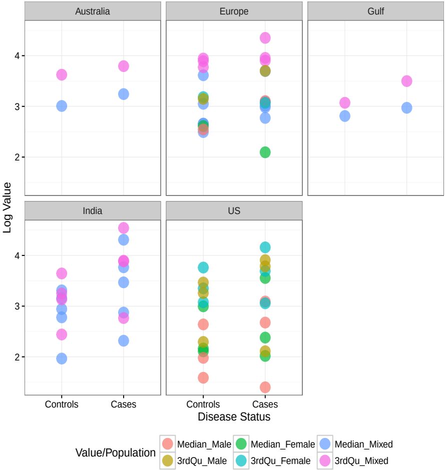
Data in cases and controls



- Available data were either means and standard deviations, or medians and interquartile ranges
- In some cases only medians were available
- Means and standard deviations were converted medians and interquantile ranges under the assumption of a log normal distribution
- When only medians were available, the missing interquantile range values have been depicted as $(0, \log(100))$

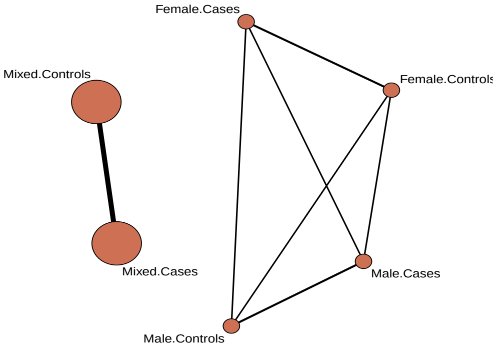
Additional features: Regions and populations

- The study populations were sampled in different regions
- In addition, populations were sampled with preference mixed or by sexes in the regions
- Concentrations were related to the region in which the study was conducted
- Additionally, the levels were related to the study populations
- On the other hand, the study populations did not differ by age, with samples clustering at around the age of 60 for Western countries, and younger (around 40) for the South Indian studies
- The reason for the rather young ages is that the events need to be observed before other competing causes of death censor them



Populations sampled

- Most samples came from mixed males and females
  - *A usual mix in cardiac disease cases is of $>$ 75% male populations*
- A few reported each group separately
- Still, 2 were either all male or all female
- For the *treatments*, or disease status, the network studies always contained both
- From the point of view of the combination of sex and disease status, the network was disjoint
- All in all, 18 studies, but the study label was repeated for the 2 studies reporting males and females separately



A model for the observed data

- Observed data are medians and quartiles.
- For the mean of the log of the observations under the assumption of a lognormal distribution results in a linear model as

$$\log(y) = \mu + \sigma\Phi^{-1}(p) \tag{1}$$

  where p corresponds to the quantile probability
- As a meta-analysis problem, the *disease status* is the design factor for which we want to distinguish or compare the observations
- We can say that these are our *treatments*
- As we have mentioned, other design factors affecting the observed levels are the populations (male, female, mixed) and the regions
- In summary a bivariate normal meta-regression model
- The complication is that all Medians were observed, but the Quartiles were missing in six cases
- The Quartiles had information about the scale parameter of the distribution, but not the Medians

Additional data issues

constraining estimation

- The estimation was conducted in the software `rjags`
- The JAGS Manual warns that *multivariate nodes cannot be partially observed*
- This of necessity turns our specification into two *univariate* regression equations
- In addition, the standard errors of the measured outcomes are not reported
  - *This requires the derivation of a theoretical expression for the log transformed log normal distribution*
  - *Using the delta method and the theoretical expression of the errors in the quantiles we obtain*

$$SE_Y = \frac{\sqrt{e^{\sigma^2} - 1}\sqrt{p(1-p)}}{\sqrt{n}\phi(z(p))} \tag{2}$$

with the standard deviation of the log transformed variable and density function of the standard normal distribution

Data issues (continued)

- A complication that we will ignore is that the scale parameter of interest σ is also included in the standard error
- Standard errors however are meant in this case to inform about the weighting of the studies
- A complication that we will not ignore is that some of the standard errors are missing due to missing quartile information
- I implement the proposal of Stevens (2011), using a likelihood for the variances of the medians and quartiles (as two independent variances)

Selected model Parametrization 1

- The model (of the mean of the log transformed observations) that was selected, using *DIC*, checks of trace plot, and *Rhat* values to determine stationarity is

$$\mu_{Y[1:2]} = \alpha + \beta_1 \chi_{\text{Male}} + \beta_2 \chi_{\text{Female}} + \beta_3 \chi_{\text{Cases}} + \beta_4 \chi_{\text{Cases}} \chi_{\text{Male}} + \beta_5 \chi_{\text{Cases}} \chi_{\text{Female}} + v_{\text{Study, Status}} + \gamma_{\text{Region}} \Phi^{-1}(p[1:2]) \quad (3)$$

  where the probability is 0.5 or 0.75 depending on whether the value is the Median or the the Quartile.

- $v_{\text{Study, Status}}$ is a random effect of treatment within study, but studies with 2 populations are treated as 2 different studies

Implied covariance matrix for the mean model

- To describe the implied covariance matrix I start with the design matrix
- Our model for the bivariate observartions can be described as a system of equations where some of the regressors are in common, and some of the regressors are not the same (SUR)

$$[\boldsymbol{X} : \boldsymbol{V}] = \begin{pmatrix} \boldsymbol{X_{11}} & \boldsymbol{X_{21}} & \boldsymbol{V_1} \\ \boldsymbol{X_{21}} & \boldsymbol{X_{22}} & \boldsymbol{V_2} \end{pmatrix} \tag{4}$$

  where matrices $\boldsymbol{X_{2j}}$ represent the regressors not in common, and $\boldsymbol{V}$ are the random effects comon to both

- The resulting covariance matrix of the estimated parameters is

$$\left([\boldsymbol{X}]' \hat{\boldsymbol{W}}^{-1} [\boldsymbol{X}] \right)^{-1} \tag{5}$$

  and

$$\hat{\boldsymbol{W}} = \Sigma_V' \cdot \begin{pmatrix} S_{11} & 0 \\ 0 & S_{21} \end{pmatrix} \cdot \Sigma_V \tag{6}$$

  with the estimated sample variances were derived ignoring the covariances, hence the only implied covariances between the estimated means are generated by the random effects

Ignore the covariance structure of the data?

- This is a case in which the covariance structure is induced by the random effects
- It corresponds to the cases discussed by Riley (2009) in which the between-study variation (induced by the random effect) and the within-study variation are not separately identified
- It would be possible to change the specification completely, and maybe use information on the within-study variation
- This maybe a less attractive approach when some of the correlation has been created by the derivation of the parameter rendering the *standard deviation of the observations*
- Alternatively we can try to include a correlation in the parameter specification
- This works if we discard some of the parameters in the saturated mean specification

Selected model Parametrization 2

- The model specification that performed best (again using DIC, Rhat, etc) replaces the coefficients β_1 and β_2 used to represent the difference of males and females with respect to the reference distribution of a *mixed population* with a new parameter θ having 3 population components, with a bivariate structure at each of the population levels

$$\mu_{Y[1:2]} = \alpha + \theta_{pop,[1:2]} + \beta_3 \chi_{\text{Cases}} + \\ \beta_4 \chi_{\text{Cases}} \chi_{\text{Male}} + \beta_5 \chi_{\text{Cases}} \chi_{\text{Female}} + v_{\text{Study, Status}} + \gamma_{\text{Region}} \Phi^{-1}(p[1:2]) \tag{7}$$

- In order for the model to converge, one of the 3 levels must be set to reference
- The convergence statistics are in fact better (and slightly smaller DIC as well) for the model in which the reference is the level *male*, hence we retain the model using the male population as reference

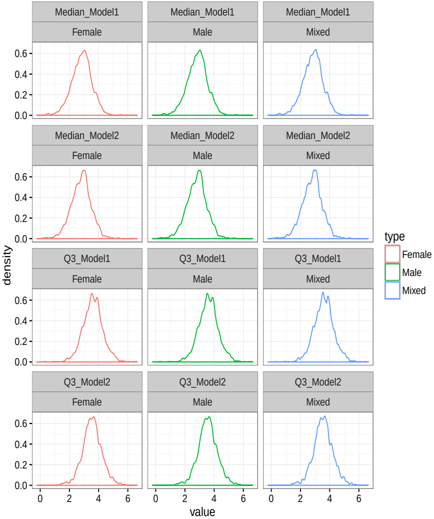
Comparison between models

- Model comparisons are obtained using DIC
- Additionally the penalized deviance calculated in RJAGS
- It is unclear whether using p_D is so meaningful due to the absence on any prior information used in the estimation
- It is possible to compare predictions, by drawing samples from the posterior and creating values that follow the model assumptions
- We examine cases, and subset by populations
- The predictions are generated retaining the same mix of Regions as in the observed data
- The predictions show longer tails, compared to the observed data
- In addition, also medians are overestimated compared to the observed data

Comparisons of predictions

Predictions of Cases

- A simple comparison of the 95th and the 99th percentiles in the observed data of sampled Cases, resulted in more extreme values for *Model 1* than for *Model 2*
- Both models overestimated, with *Model 2* larger by 23% on the 99th percentile of the Median, and twice as large for the 3rd Quartile
- Overestimation in *Model 2* was somewhat less
- For the low percentiles, *Model 2* underestimates more than *Model 1*, with the corresponding 0.01 percentile being almost 60% of the observed value

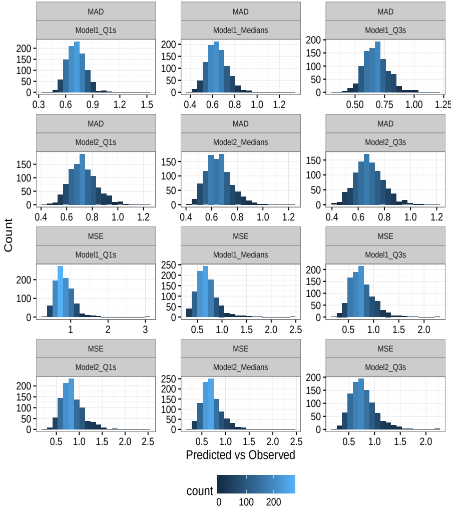


First summary

- The differences between models are very small
- Differences in Model assessment measures were also very small between models
- On the other hand, model predictions are not very reliable
- Given the small sample, the use of a log normal distribution as a model for the data is less adequate

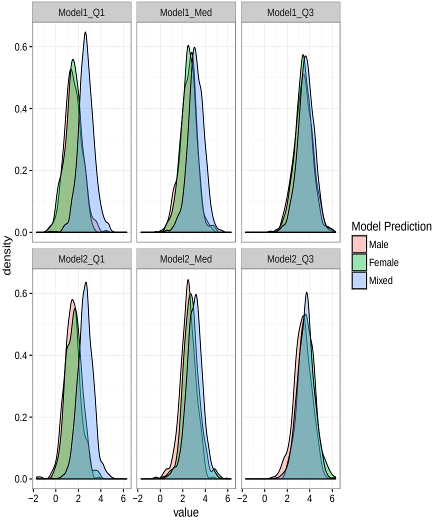
A quick fix?

- One potential improvement can be quickly obtained
- Using all the data available, including information on the 1st Quartile may improve predictions
- Again, now we do not need to redo the model search efforts, but simply expand one dimension
- Both models now are used and we repeat the prediction based comparisons
- A plot of histograms for samples of 1,000 from the posterior using *Mean Absolute Deviation* and *Mean Square Error*



NOVARTIS

Assessment of predictions for cases



- Again, we still have the under, and overestimation in the tails
- However, the overshoot is less dramatic
- Both models perform similarly in predicting
- From the point of view of the model assessment measure the *penalized Deviance* by Plummer (2008) prefers the second model

Conclusions

- Use all your data
- Be prepared to understand the trade-off between a strict parametric assumption to obtain a model and deriving predictions for a sample with extreme observations

References I