eBayesMet: on main activities regarding to data-analasis issues

Mateusz Nikodem, CASPolska

28 October 2011

Introduction

Main aims of the eBayesMet project

- Make aware of multiplicity of useful statistical methods
- Give some advises which method should be used in particular cases
- Facilitate using the Bayesian models
- **Create educational materials**

Introduction

Main aims of the eBayesMet project

- Make aware of multiplicity of useful statistical methods
- Give some advises which method should be used in particular cases
- Facilitate using the Bayesian models
- **Create educational materials**

Target groups

We focus mainly on reviewers with some background, e.g.

- Medical data analyst
- Clinicians
- University researchers
- Students

Preparing the input of the platform

Make aware of multiplicity of statistical methods

We need to

- define particular problem (type of data, parameter, etc.),
- review which methods can be used for that and which are used in fact,
- learn on limitations and possible applications of identified methods.

Preparing the input of the platform

Make aware of multiplicity of statistical methods

We need to

- define particular problem (type of data, parameter, etc.),
- review which methods can be used for that and which are used in fact,
- learn on limitations and possible applications of identified methods.

Give advises which method should be used in particular cases

We need to

- define typical data sets (sample sizes, number of studies etc.),
- test all applicable methods on each data set,
- make some more general conclusions.

Preparing the input of the platform

Facilitate using the Bayesian models

We need to

- learn on variety of theoretical models,
- implement and test the WinBUGS codes (programming),
- create the automatic transformation from "pure data" to WinBUGS code,
- create the automatic transformation from results of WinBUGS to more user-friendly form.

Preparing the input of the platform

Facilitate using the Bayesian models

We need to

- learn on variety of theoretical models,
- implement and test the WinBUGS codes (programming),
- create the automatic transformation from "pure data" to WinBUGS code,
- create the automatic transformation from results of WinBUGS to more user-friendly form.

Create educational materials

We need to

- define assumptions, starting level, etc.
- define in details the structure and the content,
- collect interesting examples,
- edit the materials.

Systematic reviews

SR of methods currently used in HTA

- Focus on CRD (DARE and HTA) databases
- Searching limited to records published between January 2009 and April 2010
- In the case of direct comparison just a random sample of 150 publications was screened
- The frequency of using particular methods in HTA reports was assessed.

Systematic reviews

SR of methods currently used in HTA

- Focus on CRD (DARE and HTA) databases
- Searching limited to records published between January 2009 and April 2010
- In the case of direct comparison just a random sample of 150 publications was screened
- The frequency of using particular methods in HTA reports was assessed.

SR of methods potentially avalilable for use in HTA

- Focus mainly on mathematical literature with detailed description of methods
- Searching using e.g. the MathSciNet and MEDLINE databasis

Systematic reviews

SR of methods currently used in HTA

- Focus on CRD (DARE and HTA) databases
- Searching limited to records published between January 2009 and April 2010
- In the case of direct comparison just a random sample of 150 publications was screened
- The frequency of using particular methods in HTA reports was assessed.

SR of methods potentially avalilable for use in HTA

- Focus mainly on mathematical literature with detailed description of methods
- Searching using e.g. the MathSciNet and MEDLINE databasis

11 different methods for (direct or indirect) meta-analysis was identified – some of them include different sub-models

Comparison of statistical methods - background

Comparison of statistical methods - background

Our needs

We want to assess some parameter being an **unknown state of nature**. It can be some relative measure of efficiency comparing two medical treatments, e.g.

Comparison of statistical methods - background

Our needs

We want to assess some parameter being an **unknown state of nature**. It can be some relative measure of efficiency comparing two medical treatments, e.g.

- Risk Ratio (Relative Risk), $RR = \frac{p_A}{p_B}$

Comparison of statistical methods - background

Our needs

We want to assess some parameter being an **unknown state of nature**. It can be some relative measure of efficiency comparing two medical treatments, e.g.

- Risk Ratio (Relative Risk), $RR = \frac{p_A}{p_B}$
- Odds Ratio, $OR = \frac{p_A(1-p_B)}{(1-p_A)p_B}$

Comparison of statistical methods - background

Our needs

We want to assess some parameter being an **unknown state of nature**. It can be some relative measure of efficiency comparing two medical treatments, e.g.

- Risk Ratio (Relative Risk), $RR = \frac{p_A}{p_B}$
- Odds Ratio, $OR = \frac{p_A(1-p_B)}{(1-p_A)p_B}$
- Mean Difference, $MD = \mu_A - \mu_B$

Motivation

Samples

- The experiments (e.g. clinical studies) are the basis for assessing the true parameter of interest.
- Aggregation of results of studies lead us to some estimation of the true parameter. We should remember that this estimation is based just on some random sample.

Motivation

Samples

- The experiments (e.g. clinical studies) are the basis for assessing the true parameter of interest.
- Aggregation of results of studies lead us to some estimation of the true parameter. We should remember that this estimation is based just on some random sample.

Meta-analyses

The process of aggregation of results of studies is called 'meta-analysis'.
There are many statistical methods to do that (e.g. some of them are a kind of weighted average, with differently adjusted weights).

Motivation

Samples

- The experiments (e.g. clinical studies) are the basis for assessing the true parameter of interest.
- Aggregation of results of studies lead us to some estimation of the true parameter. We should remember that this estimation is based just on some random sample.

Meta-analyses

The process of aggregation of results of studies is called 'meta-analysis'.
There are many statistical methods to do that (e.g. some of them are a kind of weighted average, with differently adjusted weights).
One can be surprised how different estimation can be achieved just by using different statistical methods **for the same input data**!

What can we measure?

Credibility

Usually, estimating some parameter we give the symmetric 95% confidence interval. We expect that there is (at least) 95% chance that the **true** parameter is in fact within the interval.
But is it true that each method can guarantee that?
How it comes, that some *CI*'s are wider and some are narrower?

What can we measure?

Credibility

Usually, estimating some parameter we give the symmetric 95% confidence interval. We expect that there is (at least) 95% chance that the **true** parameter is in fact within the interval.
But is it true that each method can guarantee that?
How it comes, that some *CI*'s are wider and some are narrower?

Precision

We expect that the Confidence Interval (being a kind of measure of uncertainty) is as narrow as possible. E.g. the result of meta-analysis

$$RR = 0.8(0.6; 1.06)$$

is poorer argument for clinical conclusions than:

$$RR = 0.8(0.64; 0.99)$$

How can we measure? Ideas

searching for a gold standard...

We need a gold standard to have a possibility to compare the results of different calculations with some expected result.

How can we measure? Ideas

searching for a gold standard...

We need a gold standard to have a possibility to compare the results of different calculations with some expected result.
Even if we assume, that we know some gold standards in clinical problems, how to test statistical methods?

How can we measure? Ideas

searching for a gold standard...

We need a gold standard to have a possibility to compare the results of different calculations with some expected result.
Even if we assume, that we know some gold standards in clinical problems, how to test statistical methods?
Each clinical study is just some random sample, its results does not need to be in exact accordance with gold standard!.

How can we measure? Ideas

searching for a gold standard...

We need a gold standard to have a possibility to compare the results of different calculations with some expected result.
Even if we assume, that we know some gold standards in clinical problems, how to test statistical methods?
Each clinical study is just some random sample, its results does not need to be in exact accordance with gold standard!.

...creating a gold standard!

Bradburn M, Deeks J, Berlin J, Localio R. Much ado about nothing: a comparison of meta-analytical methods with rare events. *Statistics in medicine 2007;26:53-77*.

Establishing the gold standard

Characterization of particular data set

- Type of data: (dichotomous / continuous)
- Structure of evidence (direct/indirect/mixed)
- Number of studies *
- Sample size in each group of each study *
- Probability of occurring an end-point in each group
- Statistical heterogeneity (e.g. the value of I^2)

* Can be taken from some existing systematic review

Establishing the gold standard - an example

Characterization of particular data set

- Type of data: **dichotomous**
- Structure of evidence: **direct**
- Number of studies: **4**
- Sample size in each group of each study
  $\mathbf{39/39 + 40/40 + 29/30 + 41/41}$
- Probability of occurring an end-point in each group
  treatment:$\mathbf{p_T = 0.1}$, control: $\mathbf{p_C = 0.05}$
- Statistical heterogeneity: **no; fixed effect model**

Establishing the gold standard - an example

Characterization of particular data set

- Type of data: **dichotomous**
- Structure of evidence: **direct**
- Number of studies: **4**
- Sample size in each group of each study
  $\mathbf{39/39 + 40/40 + 29/30 + 41/41}$
- Probability of occurring an end-point in each group
  treatment:$\mathbf{p_T = 0.1}$, control: $\mathbf{p_C = 0.05}$
- Statistical heterogeneity: **no; fixed effect model**

True odds ratio:

$$OR = \frac{p_T(1-p_C)}{(1-p_T)p_C} = \frac{0.1 \cdot 0.95}{0.9 \cdot 0.05} = \mathbf{2.11}$$

following the example, generating random data

- For each patient of treatment group we independently randomize if he has an outcome (with $p_T = 0.1$) or not
- For each patient of control group we independently randomize if he has an outcome (with $p_C = 0.05$) or not

Then we get first random (fictional) results of studies, e.g:

following the example, generating random data

- For each patient of treatment group we independently randomize if he has an outcome (with $p_T = 0.1$) or not
- For each patient of control group we independently randomize if he has an outcome (with $p_C = 0.05$) or not

Then we get first random (fictional) results of studies, e.g:

Study	treatment: r_T/n_T	comntrol: r_C/n_C
1	$\mathbf{3}/39$	$\mathbf{2}/39$
2	$\mathbf{5}/40$	$\mathbf{1}/40$
3	$\mathbf{3}/29$	$\mathbf{2}/30$
4	$\mathbf{2}/41$	$\mathbf{0}/41$

following the example, generating random data

We independently repeat the process for $j = 1, 2, \ldots 1000$, therefore we generate 1000 (fictional) results of studies. Taking the mean of the frequencies from all 1000 results we are very close* to assumed values of p_T and p_C.

The compared statistical methods

Now we have to establish which of statistical methods we want to compare. In case of eBayesMet analysis for direct comparisons it were:

The compared statistical methods

Now we have to establish which of statistical methods we want to compare. In case of eBayesMet analysis for direct comparisons it were:

- Mantel-Haenszel,
- Inverse Variance,
- DerSimonian Laird,

The compared statistical methods

Now we have to establish which of statistical methods we want to compare. In case of eBayesMet analysis for direct comparisons it were:

- Mantel-Haenszel,
- Inverse Variance,
- DerSimonian Laird,
- Peto,
- Bayes I (Bayesian fixed effect model based on individual studies *OR*'s),
- Bayes II (Bayesian fixed effect model based on pure source data),
- Bayes Random-effect,
- Bayes including observational studies ("naive", "equivalent" and "sceptical" models),

The compared statistical methods

Now we have to establish which of statistical methods we want to compare. In case of eBayesMet analysis for direct comparisons it were:

- Mantel-Haenszel,
- Inverse Variance,
- DerSimonian Laird,
- Peto,
- Bayes I (Bayesian fixed effect model based on individual studies *OR*'s),
- Bayes II (Bayesian fixed effect model based on pure source data),
- Bayes Random-effect,
- Bayes including observational studies ("naive", "equivalent" and "sceptical" models),

Then for each method we conduct the meta-analysis for each of 1000 randomly generated fictional results of studies and measure the credibility and precision.

Measuring credibility and precision

- **Bias**: (standarized) difference between the mean of ORs calculated for a 1000 generated results and the true OR.
  If the method is highly biased, it means that there is a regular discrepancy between result of calculations and the true value of estimated parameter.

Measuring credibility and precision

- **Bias**: (standarized) difference between the mean of *OR*s calculated for a 1000 generated results and the true *OR*.
  If the method is highly biased, it means that there is a regular discrepancy between result of calculations and the true value of estimated parameter.
- **Confidence interval coverage**, *CIC*: Percent of results of meta-analysis for which 95% confidence interval covers the true value of *OR*.
  If it's low then there is high probability that estimated CI does not cover the exact value of parameter. It can be related with the bias, or it can be caused because the width of *CI* is to small.

Measuring credibility and precision

- **Bias**: (standarized) difference between the mean of *OR*s calculated for a 1000 generated results and the true *OR*.
  If the method is highly biased, it means that there is a regular discrepancy between result of calculations and the true value of estimated parameter.
- **Confidence interval coverage**, *CIC*: Percent of results of meta-analysis for which 95% confidence interval covers the true value of *OR*.
  If it's low then there is high probability that estimated CI does not cover the exact value of parameter. It can be related with the bias, or it can be caused because the width of *CI* is to small.
- **Statistical power**, *SP*: Percent of results of meta-analysis for which the result is statistically significant in assumed direction (i.e. the lower bound of 95% confidence interval is greater than 1). Notice that methods of high *SP* often have low *CIC*.

Some results

Method	Bias	*CIC*	*SP*
Mantel-Haenszel	0.085	93.3%	49.2%
Inverse-Variance	0.011	98.0%	19.6%
DerSimonian-Laird	0.045	99.1%	16.8%
...	...	...	...

More testing...

In eBayesMet analysis we test in total 204 sets of different characteristics, including problematic cases like:

- small sample size,
- low and very low event rate,
- between-studies heterogeneity,
- unbalanced sample sizes in groups of comparison,
- substantial discrepancy between event rates in groups of comparison,

and wide spectrum of more "regular" data.
124 of this sets are related to direct comparison, 80 to indirect/mixed models.

More testing...

In eBayesMet analysis we test in total 204 sets of different characteristics, including problematic cases like:

- small sample size,
- low and very low event rate,
- between-studies heterogeneity,
- unbalanced sample sizes in groups of comparison,
- substantial discrepancy between event rates in groups of comparison,

and wide spectrum of more "regular" data.
124 of this sets are related to direct comparison, 80 to indirect/mixed models.

Results/conclusions in eBaeysMet

Main conclusions for each main type of characteristics of data are (quite systematically) included in a form of a guidelines. The most interesting and important results are also described in a form of presentations.

Conclusions

- There is no universal model which is optimal for each characteristics of input data.
- For some characteristics of data, there are several acceptable methods, but for some other (e.g. very low event-rate) each of tested method has some drawbacks.
- In case of more advanced structure of data the Bayesian models have some advantage on frequentialist ones.

Further research

- Another type of data (e.g. diagnostic tests accuracy),
- More complex attitude to assessment (including issues of evaluating quality of data).

Thank you