Incomplete Statistical Information Fusion and Its Application to Clinical Trials Data

Jianbing Ma[1], Weiru Liu[1], and Anthony Hunter[2]

[1] School of Electronics, Electrical Engineering and Computer Science,
Queen's University Belfast, Belfast BT7 1NN, UK
{jma03,w.liu}@qub.ac.uk
[2] Department of Computer Science, University College London,
Gower Street, London WC1E 6BT, UK
a.hunter@cs.ucl.ac.uk

Abstract. In medical clinical trials, overall trial results are highlighted in the *abstracts* of papers/reports. These results are summaries of underlying statistical analysis where most of the time normal distributions are assumed in the analysis. It is common for clinicians to focus on the information in the abstracts in order to review or integrate several clinical trial results that address the same or similar medical question(s). Therefore, developing techniques to merge results from clinical trials based on information in the abstracts is useful and important. In reality information in an abstract can either provide sufficient details about a normal distribution or just partial information about a distribution. In this paper, we first propose approaches to constructing normal distributions from both complete and incomplete statistical information in the abstracts. We then provide methods to merge these normal distributions (or sampling distributions). Following this, we investigate the conditions under which two normal distributions can be merged. Finally, we design an algorithm to sequence the merging of trials results to ensure that the most reliable trials are considered first.

Keywords: Normal distribution, Merging statistical data, Consistency analysis.

1 Introduction

Clinical trials are widely used to test new drugs or to compare the effect of different drugs [10]. Overall trial results are summarized in *abstracts* of papers/reports that report the trial details. Given that there is a huge number of trials available and details of reports are very time consuming to read and understand, clinicians, medical practioners and general users mainly make use of this highly summaritive information in the abstracts to obtain an overall impression about drugs of interest. For example, many clinical trials have been carried out to investigate the intraocular pressure-lowering efficacy of drugs, such as travoprost, bimatoprost, and latanoprost, [2,4,9,11,13,14,15,16,18]. When an overview or survey of a collection of clinical trials is required, a *merged or integrated* result is desirable.

When the full details about the statistics used in the trials are available, merging the results from these trials is usually a matter of systematic use of established techniques

H. Prade and V.S. Subrahmanian (Eds.): SUM 2007, LNAI 4772, pp. 89–103, 2007.

from statistics. However, in reality, it is impossible to read all the details about each trial. Most of the time, information in abstracts is most useful for the following reasons. First, it is common that a person reads the abstract of a paper before reading the full paper/report when deciding if the trial is relevant. Second, with more and more information available on the Web, obtaining an abstract is much easier and most of the time it is free while getting a full paper can be more difficult and expensive (one may need to pay a fee). Third, in the field of clinical trials, abstracts often provide sufficient information about trial analysis for a clinician to update their knowledge (such as about the pros and cons of a particular treatment). Therefore, we concentrate here on developing techniques to merging information solely provided in the abstracts.

As a convention, clinical trials usually use normal distributions to record trial results. So it is a natural idea to merge normal distributions to a single one as the integrated result. There is a classical method to merge normal distributions [3]. However, when using this method to merge two identical normal distributions, the merged result is a different normal distribution which is counterintuitive, since we would expect the merged result to be the same as the original distribution. Some other methods have been proposed to integrate probability distributions ([3,12,19]) or to learn the integrated probability distributions ([6,8]). But these methods generally do not lead to a normal distribution as a result, so they are not suitable for our purposes. Furthermore, in some abstracts about clinical trials, information about underlying statistics can be incomplete, e.g., the standard deviations are not given. To deal with this, we need to make use of some background knowledge in order to construct an adequate normal distribution to facilitate merging.

In this paper, we first propose approaches to constructing normal distributions from both complete and incomplete statistical information in the abstracts. We then provide methods to merge normal distributions. We also study how to measure if two normal distributions are in conflict (or consistent), in order to decide if they should be merged. To sequence a merging of multiple trials data, we introduce the notion of reliability to sort the merging sequence. An algorithm is designed to merge trials results based on both reliabilities of trials and consistencies among trials.

The remainder of this paper is organized as follows. Section 2 provides some preliminary knowledge about normal distributions and introduces the notion of degrees of consistency of normal distributions. Section 3 introduces categories of statistical information commonly found in abstracts and how they are related to normal distributions. Section 4 contains our merging methods for merging complete and incomplete statistical information. In Section 5, we give a definition for measuring conflict among normal distributions and how this is used to decide if a merging shall take place. Section 6 investigates how a collection of clinical trials results should be sequenced for merging and an algorithm is designed to implement this. Finally, in Section 7, we conclude the paper.

2 Preliminaries

We start with some basic concepts about normal distributions. We then define the notion of conflict (or consistency) of two normal distributions.

Definition 1. *A random variable X with mean value μ and variance σ^2 is* **normally distributed** *if its* **probability density function** *(pdf for short) f is defined as follows:*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

In statistics, a normal distribution associated with a random variable is denoted as $X \sim N(\mu, \sigma^2)$. For the convenience of further calculations in the rest of the paper, we use notation $X \sim N(\mu, \sigma)$ instead of $X \sim N(\mu, \sigma^2)$ for a normal distribution of variable X. That is, we use a standard deviation rather than a variance because this will greatly simplify mathematical equations in Section 4.

A normal distribution with $X \sim N(0, 1)$ is called a **standard normal distribution**. Any normal distribution $N(\mu, \sigma)$ can be standardized by letting a random variable $Z = \frac{X-\mu}{\sigma}$, then $Z \sim N(0, 1)$ is a standard normal distribution. For $N(0, 1)$, the standard normal distribution table in statistics [20] provides sufficient information for further calculations of probabilities, such as the probability of an interval that the variable falls in.

In statistics, random samples of individuals are often used as the representatives of the entire group of individuals (often denoted as a population) to estimate the values of some parameters of the population. The mean of variable X of the samples, when the sample size is reasonably large, follows a normal distribution. The **standard error of the mean** (SEM for short), which is the standard deviation of the sample mean, is given by $\mathsf{SEM} = \frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of X of the population and n is the number of samples chosen from the population. We can write $\bar{X} \sim N(\mu, \mathsf{SEM})$. When σ is unknown, the standard deviation s of the samples is often used to replace σ.

To help define the degree of consistency of normal distributions, we introduce the following well-known result.

Let v_1 and v_2 be two vectors. The angle between two vectors can be computed as follows:

$$cos(v_1, v_2) = \frac{< v_1, v_2 >}{\parallel v_1 \parallel_2 \parallel v_2 \parallel_2}$$

where $< v_1, v_2 >$ is the inner product of the vectors and $\parallel v \parallel_2$ is the L_2 norm.

Definition 2. *Let two normal distributions have $f_1(.)$ and $f_2(.)$ as their pdfs respectively. The* **degree of consistency** *of the two normal distributions, denoted as $c(f_1, f_2)$ is defined as follows:*

$$c(f_1, f_2) = \frac{< f_1, f_2 >}{\parallel f_1 \parallel_2 \parallel f_2 \parallel_2}$$

where $< f_1, f_2 >$ is the inner product given by:

$$< f_1, f_2 >= \int_{-\infty}^{+\infty} f_1(x) f_2(x) dx$$

and $\parallel f \parallel_2$ is the L_2 norm given by:

$$\parallel f \parallel_2 = \int_{-\infty}^{+\infty} f^2(x) dx$$

The degree of consistency $c(f_1, f_2)$ defined above is in (0,1]. When f_1 and f_2 are identical normal distributions, $c(f_1, f_2) = 1$, while $c(f_1, f_2) \rightarrow 0$ when $\parallel \mu_1 - \mu_2 \parallel \rightarrow \infty$. Value $c(f_1, f_2)$ increases along with the *closeness* of f_1 and f_2.

3 Statistical Information in Abstracts

In abstracts of papers about clinical trials, information about underlying statistics can be summarized into the following four categories.

- Category I: A normal distribution can be identified when both μ and σ are given.
- Category II: A normal distribution can be identified when only μ is given.
- Category III: A normal distribution can be constructed when a confidence interval is given.
- Category IV: A normal distribution can be constructed if at least two sentences, each of which gives a probability value of the variable in a particular range, are available in the abstract.

After looking through a large collection of abstracts of clinical trials on IOP reductions using different drugs, we believe that the above four categories cover a significant proportion of statistical information in abstracts [2,4,9,11,13,14,15,16,18]. In this paper, we concentrate on how to model and merge these four types of information.

For each category of statistical information, we try to interpret it in terms of a normal distribution. We use X to denote the random variable implied in the context of each sentence.

For the first category, a normal distribution is explicitly give, for example, sentence "Mean IOP reduction at 6 months was -9.3+/-2.9 mmHg in the travoprost group" can be interpreted as follows

$$X \sim N(-9.3, 2.9)$$

For the second category, a normal distribution can be defined with a missing standard deviation. For instance, sentence "There was at least 90% power to detect a mean IOP change from baseline of 2.9 mmHg" can be interpreted as

$$X \sim N(2.9, \sigma)$$

where σ is unknown. To make use of this information, we need to draw on background knowledge about the interval that σ lies. From our investigation, this information can be obtained either through a clinician or from some text books on this specific topic. Therefore, we can assume that this background knowledge is available and can be used during merging.

For the third category of information, a confidence interval $[a, b]$ is given. It is then possible to convert this confidence interval into a normal distribution as follows

$$\mu = \frac{a+b}{2}, \qquad \sigma = \frac{b-a}{2k}$$

As a convention, the presented analysis of clinical trials results usually use the 95% confidence interval. In this case, we have $k = 1.96$. However, if a given confidence

interval is not the usual 95% confidence interval (say, it uses the p-confidence interval), it is possible to use the standardization of the normal distribution as $P(Z \in [-k, k]) = p$. Then value k can be found by looking up the standard normal distribution table.

For example, from sentence "Bimatoprost provided mean IOP reductions from baseline that ranged from 6.8 mmHg to 7.8 mmHg (27% to 31%)", it is possible to get a normal distribution $N(\mu, \sigma)$ with full information.

For the fourth category of information, a sentence like "By month 3, 85% of participants in the bimatoprost group had a mean IOP reduction of at least 20%" can be used to define a probability of the variable in a particular range, such as

$$P(X \geq 0.2b) = 0.85$$

where b is the baseline IOP value.

It is possible to generalize this expression to $P(X \geq x) = p$ and then further to

$$P(\frac{X - \mu}{\sigma} \geq \frac{x - \mu}{\sigma}) = p$$

using the standardization technique.

By looking up the standard normal distribution table, it is possible to determine the value for $(x - \mu)/\sigma$. Similarly, if another sentence is given in the abstract with another range for X, then another equation $(x^{'} - \mu)/\sigma = y^{'}$ can be obtained, therefore, the values of μ and σ can be calculated. In a situation where only one of such sentence is given but μ is provided, a normal distribution can still be constructed. Otherwise, it would be difficult to use this piece of information. From our analysis of abstracts, it seems that it is very rare that only one of these sentences is given, usually, two or more such descriptions are available.

To summarize, from our case study, usually we can get normal distributions from all the four type of information we normally find in abstracts.

4 Merging Normal Distributions

In this section, we discuss how to merge two normal distributions when either full information or partial information about them is available.

4.1 Normal Distributions with Full Information

Let the normal distributions associated with two random variables X_1 and X_2 be as following

$$X_1 \sim N(\mu_1, \sigma_1), \qquad X_2 \sim N(\mu_2, \sigma_2)$$

We want to merge them into a new normal distribution with random variable X as $X \sim N(\mu, \sigma)$. An intuitive idea for merging is to let the merged μ *divide* the two distributions equally. Since in general $\sigma_1 \neq \sigma_2$, we cannot simply let $\mu = \frac{\mu_1 + \mu_2}{2}$. We define the following criterion that μ should satisfy

$$P(X_1 \leq \mu) + P(X_2 \leq \mu) = P(X_1 \geq \mu) + P(X_2 \geq \mu). \tag{1}$$

Indeed, the above equation ensures that the merged μ divides the two distributions equally.

Proposition 1. *Assume we have $X_1 \sim N(\mu_1, \sigma_1)$, $X_2 \sim N(\mu_2, \sigma_2)$, and let μ be the merged result that satisfies (1), then we have*

$$\mu = \frac{\mu_1\sigma_2 + \mu_2\sigma_1}{\sigma_1 + \sigma_2}$$

The proof of this and other subsequent propositions are given in the Appendix.

It is easy to see that if $\sigma_1 = \sigma_2$, then $\mu = \frac{\mu_1+\mu_2}{2}$. In particular, if two normal distributions are the same, then the merged μ should not be changed, which is exactly what we want.

From Proposition 1, we notice that the coefficients of μ_1 (also X_1) and μ_2 (X_2) in calculating μ are $\frac{\sigma_2}{\sigma_1+\sigma_2}$ and $\frac{\sigma_1}{\sigma_1+\sigma_2}$, respectively. So when calculating σ, we still use these two coefficients for X_1 and X_2 and the variance σ^2 should satisfy

$$\sigma^2 = \frac{\sigma_2}{\sigma_1 + \sigma_2}\int_{-\infty}^{+\infty} f_1(X_1)(x-\mu)^2 dx + \frac{\sigma_1}{\sigma_1 + \sigma_2}\int_{-\infty}^{+\infty} f_2(X_2)(x-\mu)^2 dx \quad (2)$$

where $f_1(X_1)$ and $f_2(X_2)$ are the pdfs for X_1 and X_2 respectively.

Proposition 2. *Assume we have $X_1 \sim N(\mu_1, \sigma_1)$, $X_2 \sim N(\mu_2, \sigma_2)$, and let variance σ^2 be the merged result that satisfies (2), then we have*

$$\sigma = \sqrt{\sigma_1\sigma_2(1 + \frac{(\mu_1 - \mu_2)^2}{(\sigma_1 + \sigma_2)^2})}$$

It is easy to check that such a σ satisfies the following properties.

Proposition 3. *Assume we have $X_1 \sim N(\mu_1, \sigma_1)$, $X_2 \sim N(\mu_2, \sigma_2)$, and the merged result of these two distributions is $X \sim N(\mu, \sigma)$, then*

1. *If $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$, then $\sigma = \sigma_1 = \sigma_2$.*
2. *If $\sigma_1 = \sigma_2$, but $\mu_1 \neq \mu_2$, then $\sigma > \sigma_1 = \sigma_2$.*
3. *If $\sigma_1 \neq \sigma_2$, but $\mu_1 = \mu_2$, then $\mathsf{min}(\sigma_1.\sigma_2) \leq \sigma \leq \mathsf{max}(\sigma_1, \sigma_2)$.*

Proof: The proof is straightforward and omitted.

Unfortunately, it does not satisfy the associative property.

Example 1. *The following two normal distributions are constructed from [15,16]. In [15], the baseline IOP (Intraocular Pressure) in the latanoprost 0.005% group is (Mean(SD)) 24.1(2.9) mm Hg. We use X_{NM} to denote the normal distribution of the baseline IOP in the latanoprost 0.005% group, so we get $X_{NM} \sim N(24.1, 2.9)$. Similarly, in [[16]], the corresponding baseline IOP is 23.8(1.7) mm Hg, so we get $X_{PY} \sim N(23.8, 1.7)$.*

Based on Propositions 1 and 2, we get

$$\mu = \frac{24.1 * 1.7 + 23.8 * 2.9}{1.7 + 2.9} = 23.9 \quad \sigma = \sqrt{1.7 * 2.9 * (1 + \frac{(24.1 - 23.9)^2)}{(1.7 + 2.9)^2})} = 2.2$$

So the merged normal distribution is $X_{NMPY} \sim N(23.9, 2.2)$ *and it is closer to* X_{PY} *than to* X_{NM}*. This is natural because* X_{PY} *with a smaller standard deviation means that this normal distribution is more accurate and most of the values will be closer to its mean value. Therefore, the merged result has a mean value that is closer to this distribution.*

There is another well known method for merging two normal distributions [3] which gives

$$\mu = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \qquad \sigma = \sqrt{\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}} \tag{3}$$

The above two equations come from the mathematical result of the distribution of $X_1 + X_2$. A drawback of this equation is that when the two original normal distributions are the same, the merged σ is different from the original one. This is not intuitively what we want to get from a merging. Therefore, we start from the assumption that the mean value μ should divide the two normal distributions equivalently that is how we have obtained the different equations from above to calculate μ and σ.

4.2 A Special Case Considering the Sample Mean

Now we consider situations where variable X denotes the mean of the samples. From $SEM = \frac{\sigma}{\sqrt{n}}$, we get $n = \frac{\sigma^2}{SEM^2}$. Let X_1 be the mean of m_1 variables whose standard deviation is σ_1 and X_2 be the mean value of m_2 variables whose standard deviation is σ_2. Provided that m_1 and m_2 are reasonably large, X_1 and X_2 both follow a normal distribution as

$$X_1 \sim N(\mu_1, SEM_1), \qquad X_2 \sim N(\mu_1, SEM_2)$$

respectively. When we consider merging two clinical trials results, we need to assume that the populations of the two samples are similar (or even the same), therefore, it is reasonable to assume that $\sigma_1 = \sigma_2$. Under this assumption, we have the following merging result

Proposition 4. *Let* $X_1 \sim N(\mu_1, SEM_1)$ *and* $X_2 \sim N(\mu_1, SEM_2)$*, then for the merged normal distribution, we have*

$$\mu = \frac{\mu_1 * SEM_2^2 + \mu_2 * SEM_1^2}{SEM_1^2 + SEM_2^2}$$

Proposition 5. *Let* $X_1 \sim N(\mu_1, SEM_1)$ *and* $X_2 \sim N(\mu_1, SEM_2)$*, then for the merged normal distribution, we have*

$$SEM = \sqrt{\frac{SEM_1^2 * SEM_2^2}{SEM_1^2 + SEM_2^2}}$$

Although the above merging results happen to be similar to the pair of equations in (3), we need to point out that they are used in different circumstances. Unlike equations in (3) which solve the sum of two normal distributions, Propositions 4 and 5 deal with the merging of the sample means and with the assumption that the standard deviations of the populations of the two samples are equivalent.

Here if the two normal distributions are the same, the SEM^2 will be a half of the original one. This satisfies the property that the variation of the mean is in counter proportion to the sample size, so when the sample size is doubled (after merging), SEM^2 is halved. It is also easy to prove that the above merging method has the associative property.

Example 2. *The mean IOP reduction is a variable which is the mean of the distribution of samples. When the sample size is reasonably large, it follows a normal distribution. In [16], the mean IOP reduction of the travoprost 0.004% group at the end of three months is* $X_{PY} \sim N(9.4, 3.1)$*, while in [11], the corresponding mean IOP reduction at the end of three months with the same drug is* $X_{HS} \sim N(8.7, 3.8)$.

Based on Proposition 4 and Proposition 5, we get:

$$\mu = \frac{9.4 * 3.8^2 + 8.7 * 3.1^2}{3.8^2 + 3.1^2} = 9.1 \qquad SEM = \sqrt{\frac{3.8^2 * 3.1^2}{3.8^2 + 3.1^2}} = 2.4$$

So the merged normal distribution is $X_{PYHS} \sim N(9.1, 2.4)$*. We can see that the merged* SEM *is significantly smaller than the original ones, because SEM decreases when a sample size increases.*

4.3 Normal Distributions with Missing Standard Deviations

We consider situations where one of the two standard deviations (or standard errors of the mean) in two normal distributions is missing. As we have observed, in medical domains there is usually an interval that contains σ or SEM. For example, in the clinical trials, σ for baseline IOP is usually in [1.5, 4.0] mm Hg. We can then use the interval for merging. Without loss of generality, we assume that σ_2 (or the SEM_2) is unknown, but it is in an interval.

Proposition 6. *Let* $X_1 \sim N(\mu_1, \sigma_1)$ *and* $X_2 \sim N(\mu_2, \sigma_2)$ *be two normal distributions where* μ_1, σ_1, μ_2 *are given but* σ_2 *is in interval* $[a, b]$*. Then the merged* μ *based on Proposition 1 is as follows*

If $\mu_1 > \mu_2$*, then* $\mu \in [\frac{\mu_1 a + \mu_2 \sigma_1}{\sigma_1 + a}, \frac{\mu_1 b + \mu_2 \sigma_1}{\sigma_1 + b}]$
If $\mu_1 = \mu_2$*, then* $\mu = \mu_1$
If $\mu_1 < \mu_2$*, then* $\mu \in [\frac{\mu_1 b + \mu_2 \sigma_1}{\sigma_1 + b}, \frac{\mu_1 a + \mu_2 \sigma_1}{\sigma_1 + a}]$

Proposition 7. *Let* $X_1 \sim N(\mu_1, \sigma_1)$ *and* $X_2 \sim N(\mu_2, \sigma_2)$ *be two normal distributions where* μ_1, σ_1, μ_2 *are given but* σ_2 *is in interval* $[a, b]$*. Then the merged* σ *based on Proposition 2 is as follows*

If $\mu_1 = \mu_2$*, or* $b \leq \sigma_1 + \frac{8\sigma_1^3}{(\mu_1 - \mu_2)^2}$*, then* $\sigma \in [\sqrt{\sigma_1 a(1 + \frac{(\mu_1 - \mu_2)^2}{(\sigma_1 + a)^2})}, \sqrt{\sigma_1 b(1 + \frac{(\mu_1 - \mu_2)^2}{(\sigma_1 + b)^2})}]$
If $\mu_1 \neq \mu_2$ *and* $a \geq \sigma_1 + \frac{(\sigma_1 + b)^3}{(\mu_1 - \mu_2)^2}$*, then* $\sigma \in [\sqrt{\sigma_1 b(1 + \frac{(\mu_1 - \mu_2)^2}{(\sigma_1 + b)^2})}, \sqrt{\sigma_1 a(1 + \frac{(\mu_1 - \mu_2)^2}{(\sigma_1 + a)^2})}]$

Proposition 8. *Let $X_1 \sim N(\mu_1, SEM_1)$ and $X_2 \sim N(\mu_2, SEM_2)$ be two normal distributions where μ_1, SEM_1, μ_2 are known but SEM_2 is in interval $[a, b]$. Then the merged μ based on Proposition 4 is as follows*

If $\mu_1 > \mu_2$, then $\mu \in [\frac{\mu_1 a^2 + \mu_2 SEM_1^2}{SEM_1^2 + a^2}, \frac{\mu_1 b^2 + \mu_2 SEM_1^2}{SEM_1^2 + b^2}]$
If $\mu_1 = \mu_2$, then $\mu = \mu_1$
If $\mu_1 < \mu_2$, then $\mu \in [\frac{\mu_1 b^2 + \mu_2 SEM_1^2}{SEM_1^2 + b^2}, \frac{\mu_1 a^2 + \mu_2 SEM_1^2}{SEM_1^2 + a^2}]$

Proposition 9. *Let $X_1 \sim N(\mu_1, SEM_1)$ and $X_2 \sim N(\mu_2, SEM_2)$ be two normal distributions where μ_1, SEM_1, μ_2 are known but SEM_2 is in interval $[a, b]$. Then the merged SEM based on Proposition 5 is as follows*

$$SEM \in [\sqrt{\frac{SEM_1^2 + a^2}{SEM_1^2 a^2}}, \sqrt{\frac{SEM_1^2 + b^2}{SEM_1^2 b^2}}]$$

In situations where both standard deviations (or the SEMs) are missing, the only method we can use is to let the merged $\mu = \frac{\mu_1 + \mu_2}{2}$ and leave the new σ (or the SEM) still in the interval $[a, b]$.

5 Consistency Analysis of Two Normal Distributions

Merging should take place when two normal distributions refer to the trials that have been undertaken in similar conditions. More specifically, we shall consider the following conditions. First, both trials should be for the same variable (e.g, both for the mean IOP reduction), for the same drug used (e.g, both for travoprost 0.004%), and for the same duration (e.g, both for 12-months). Second, they should be under a similar trial design (e.g, both are cross-over designs) and with similar participants (e.g, the average age should be approximately equivalent). Third, the two distributions from two trials should not be contradict with each other, that is, we need to define a kind of measure to judge how consistent (or conflicting) the two distributions are and give a threshold to indicate whether two distributions can be merged.

Proposition 10. *Let f_1 and f_2 be the pdfs for $X_1 \sim N(\mu_1, \sigma_1)$ and $X_2 \sim N(\mu_2, \sigma_2)$ respectively, then the* **degree of consistency** *of X_1 and X_2 based on Definition 2 is*

$$c(f_1, f_2) = \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} exp(-\frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)})$$

Definition 3. *Let $X_1 \sim N(\mu_1, \sigma_1)$ and $X_2 \sim N(\mu_2, \sigma_2)$ be two normal distributions with f_1 and f_2 as their pdfs respectively. They are* **consistent** *and can be merged if $c(f_1, f_2) \geq t$ holds where t is pre-defined threshold for consistency (such as 0.9).*

The degree of inconsistency (or conflict) can be defined as $1 - c(f_1, f_2)$. The threshold is application dependent and can be tuned to suit a particular application.

When variables X_1 and X_2 denote the means of samples, the above proposition still holds except that we should replace σs with SEMs. In a situation where a standard deviation is missing from one of the normal distributions, we assume the two given normal distributions share similar conditions, so we simply let the missing standard deviation be equal to the existing one. Then the above equation is reduced to:

$$c(f_1, f_2) = exp(-\frac{(\mu_1 - \mu_2)^2}{4\sigma_1^2})$$

When both of the standard deviations are not given, as discussed in Section 4, if we know that $\sigma \in [a, b]$, then we have

$$c(f_1, f_2) \in [exp(-\frac{(\mu_1 - \mu_2)^2}{4a^2}), exp(-\frac{(\mu_1 - \mu_2)^2}{4b^2})]$$

For this case if the given threshold t also falls within this interval, it would be hard to tell whether $t \geq c(f_1, f_2)$ holds. A simple method is to compare t with the middle value of the interval, if t is less than the middle value, a merge shall take place otherwise a merge may not be appropriate.

Example 3. *(Con't Example 1) For the two normal distributions in Example 1, we have* $c(f_1, f_2) > 0.9$*, so these two distributions can be merged.*

If we use the two normal distributions of the baseline IOP of the travoprost 0.004% group in [[11], data collected at 10am] and [16], we have $X_1 \sim N(28.0, 3.1)$*,* $X_2 \sim N(25.4, 3.0)$*, which gives* $c(f_1, f_2) < 0.9$*, so we advise that these two distributions should not be merged. However, if* t *is changed to be* 0.8*, they can be merged. This example also reveals that in our definition of consistency between two normal distributions, the values of means from the distributions play more dominating roles than the standard deviations.*

6 Sequencing the Merge of Multiple Trials Data

When there are more than two (potentially many) clinical trials data to be merged, the sequence of merging is very important because our merging methods of two normal distributions are not associative. For the four categories of information we summarized in Section 3, we can get a normal distributions with full information for three types and for the 2nd category, we get a distribution with a missing standard deviation. Since merging a full distribution with an incomplete distribution results in σ (or SEM) being in an interval, this result will make any subsequence merging more complicated. To address this issue, we merge full and incomplete distributions separately first and then merge the merged results from these two separate sequences.

To decide which trial should be the first data to consider, we consider reliabilities. Unlike the use of reliabilities in the form $\lambda_1 P_1 + \lambda_2 P_2$ where the λ_i, $i = 1, 2$ are used to denote the reliabilities of the sources [1,5,7,17], we use the reliability information to rank clinical trials data. Reliability information is usually provided separately as extra information, for clinical trials, we do not have this information, so we take the number

of samples used in a trial as a measure of reliability. That is, the larger the sample size, the more reliable the trial result.

Given a set of trials results that are modeled with incomplete distributions (σ is missing), we rank them based on their sample sizes as (we denote each trial result as μ) $\mu_1, \mu_2, \ldots, \mu_n$. Then the merging of these results are as follows. We first find all the μs that are consistent with μ_1 (the most reliable one) and calculate their average (including μ_1). The result is denoted as μ_1^1. We delete these entries from the above sequence, and we then repeat this procedure for the current most reliable μ in the remaining sequence, and so on. When the initial sequence is empty, we get a new set of μs: $\mu_1^1, \mu_2^1, \ldots, \mu_{n_1}^1$.

When $n_1 = n$, no merging has been taken. That is all trials data are inconsistent with each other. We return μ_1 as the merged result as it is the most reliable one. If $n_1 < n$, we repeat the above merging procedure for the new sequence $\mu_1^1, \mu_2^1, \ldots, \mu_{n_1}^1$.

This merging procedure is described in the following algorithm.

Algorithm Merge(μs)
Begin
$\Psi_1 = \{< \mu_1, 1 >, < \mu_2, 2 >, \ldots, < \mu_n, n >\}$, $\Psi_2 = \{\}$, $m = n$;
//Here $< \mu_i, i >$ means that μ_i is the ith most reliable one.
while $n \neq 1$ **do**
while $|\Psi_1| > 0$ **do**
Let $< \mu_i, i >$ have the minimal i (or, the most reliable one) in Ψ_1, and let $S = \{< \mu_{i_1}, i_1 >, < \mu_{i_2}, i_2 >, \ldots, < \mu_{i_j}, i_j >\}$ containing all the elements in Ψ_1 where the μ_{i_k}, $1 \leq k \leq j$ are consistent with μ_i based on Def 3 (note that μ_i itself is in S), let $\mu_i^{'} = (\sum_{k=1}^{j} \mu_{i_k})(|S|)$, and $\Psi_2 = \Psi_2 \cup \{< \mu_i^{'}, i >\}$.
Let $\Psi_1 = \Psi_1 \setminus S$.
End of while
If $|\Psi_2| = m$, **Return** $\mu_1^{'}$ in Ψ_2 as the result.
Else Let $\Psi_1 = \Psi_2$, $m = |\Psi_2|$, and $\Psi_2 = \{\}$.
End of while
Return μ_1 in the Ψ_1 which has the index 1.

This algorithm stops when no further merging is possible, either because all trials are in conflict or all the results have already been merged into one.

In terms of computational complexity, the number of consistency checks is $O(n^3)$, and the number of arithmetic calculation is $O(n)$. So the complexity of the algorithm is $O(n^3)$.

When we replace the set of trials results in the above algorithm with a set of complete normal distributions $N(\mu_1, \sigma_1), N(\mu_2, \sigma_2), \ldots, N(\mu_n, \sigma_n)$, this algorithm merges these full distributions except that the calculation of averages of μs should be replaced by the equations in Proposition 1 and Proposition 2.

Finally, we merge the results of these two separate sequences to obtain a final result.

7 Conclusion

In this paper, we investigated different types of statistical information implied in abstracts (of papers/reports) about clinical trials. We summarized four types of statistical

information and three out of these four types would enable us to get a full normal distribution about a trial result. The 2nd category provides us with only incomplete distributions. Based on this, we developed methods to merge these types of information. We also defined how to measure the degree of consistency between two distributions. An algorithm was designed to sequence multiple merges.

There are a number of issues we will further look at. First, the threshold used in consistency checking would have an effect on the final result of merging, we will experiment with different threshold values to see how much effect they have. Second, the algorithm divides trials results based on whether a distribution is complete. There can be other sequences for merging which may be able to merge consistent results (currently in the two separate sequences) at an earlier stage. We will need to experiment on this to see what sequence provides the most suitable merging and what conditions are required. Third, we will consider some necessary background knowledge in order to select trials from a large collection of trials data in order to perform a merge.

Acknowledgement. This work is funded by the EPSRC projects with reference numbers: EP/D070864/1 and EP/D074282/1.

References

1. Boussion, N., Soulez, G., Guise De, J., Daronat, M., Qin, Z., Cloutie, G.: Geometrical accuracy and fusion of multimodal vascular images: a phantom study. Med. Phys. 31(6) (2004)
2. Chiselita, D., Antohi, I., Medvichi, R., Danielescu, C.: Comparative analysis of the efficacy and safety of latanoprost, travoprost and the fixed combination timolol-dorzolamide; a prospective, randomized, masked, cross-over design study. Oftalmologia 49(3), 39–45 (2005)
3. Catherine, M., Alison, C., Christophe, M., William, J.: Experimental issues of functional merging on probability density estimation. Artificial Neural Networks, Conference Publication No. 440 pp. 7–9 (1997)
4. Cantor, L.B., Hoop, J., Morgan, L., Wudunn, D., Catoira, Y.: Bimatoprost-Travoprost Study Group, Intraocular pressure-lowering efficacy of bimatoprost 0.03% and travoprost 0.004$ in patients with glaucoma or ocular hypertension. Br J Ophthalmol 90(11), 1370–1373 (2006)
5. Delmotte, F., Borne, P.: Modeling of reliability with possibility theory. IEEE Trans. SMC 28(1), 78–88 (1998)
6. DasGupta, S.: Learning mixtures of Gaussians. In: Proc. IEEE Foundations of Computer Science (1999)
7. Elouedi, Z., Mellouli, K., Smets, P.: Assessing sensor reliability for multisensor data fusion within the transferable belief model. IEEE Trans. on SMC-Part B 34(1), 782–787 (2004)
8. Freund, Y., Mansour, Y.: Estimating a mixture of two product distributions. In: Estimating a mixture of two product distributions, ACM Press, New York (1999)
9. Gracia-Feijo, J., Martinez-de-la-Casa, J.M., Castillo, A., Mendez, C., Fernandez-Vidal, A., Garcia-Sanchez, J.: Circadian IOP-lowering efficacy of travoprost 0.004$ ophthalmic solution compared to latanoprost 0.005%. Curr. Med. Res. Opin. 22(9), 1689–1697 (2006)
10. Greenhalgh, T.: How to Read a Paper: The Basics of Evidence-Based Medicine. BMJ Press (1997)
11. Howard, S., Silvia, O.N., Brian, E., John, S., Sushanta, M., Theresa, A., Michael, V.: The Safety and Efficacy of Travoprost 0.004%/Timolol 0.5% Fixed Combination Ophthalmic Solution. Ame J. Ophthalmology 140(1), 1–8 (2005)

12. Molina, C., Niranjan, M.: Pruning with replacement on limited resource allocating networks by F-projections. Neural Computation 8, 345–356 (1996)
13. Michael, T., David, W., Alan, L.: Projected impact of travoprost versus timolol and latanoprost on visual field deficit progression and costs among black glaucoma subjects. Trans. Am. Ophthalmol Soc. 100, 109–118 (2002)
14. Noecker, R.J., Earl, M.L., Mundorf, T.K., Silvestein, S.M., Phillips, M.P.: Comparing bimatoprost and travoprost in black Americans. Curr. Med. Res. Opin. 22(11), 2175–2180 (2006)
15. Nicola, C., Michele, V., Tiziana, T., Francesco, C., Carlo, S.: Effects of Travoprost Eye Drops on Intraocular Pressure and Pulsatile Ocular Blood Flow: A 180-Day, Randomized, Double-Masked Comparison with Latanoprost Eye Drops in Patients with Open-Angle Glaucoma. Curr. Ther. Res. 64(7), 389–400 (2003)
16. Parmarksiz, S., Yuksel, N., Karabas, V.L., Ozkan, B., Demirci, G., Caglar, Y.: A comparison of travoprost, latanoprost and the fixed combination of dorzolamide and timolol in patients with pseudoexfoliation glaucoma. Eur. J. Ophthalmol. 16(1), 73–80 (2006)
17. Rogova, G., Nimier, V.: Reliability in information fusion: literature survey. In: Proc. of Information Fusion, pp. 1158–1165 (2004)
18. Stefan, C., Nenciu, A., Malcea, C., Tebeanu, E.: Axial length of the ocular globe and hypotensive effect in glaucoma therapy with prostaglandin analogs. Oftalmologia 49(4), 47–50 (2005)
19. Arora, S., Kannan, R.: Learning mixtures of arbitrary Gaussians. In: STOC(STOC 2001), pp. 6–8 (2001)
20. Standard probability Table: `http://onlinepubs.trb.org/onlinepubs/nchrp/cd-22/v2appendixc_files/image002.gif`

Appendix

Proof of Proposition 1: From $P(X_1 \leq \mu)+P(X_2 \leq \mu) = P(X_1 \geq \mu)+P(X_2 \geq \mu)$ and $P(X_1 \leq \mu) + P(X_2 \leq \mu) + P(X_1 \geq \mu) + P(X_2 \geq \mu) = 2$, we get:

$$P(X_1 \leq \mu) + P(X_2 \leq \mu) = 1.$$

By using the standardization of the normal distributions, we get

$$P(\frac{X_1 - \mu_1}{\sigma_1} \leq \frac{\mu - \mu_1}{\sigma_1}) + P(\frac{X_2 - \mu_2}{\sigma_2} \leq \frac{\mu - \mu_2}{\sigma_2}) = 1.$$

So it is equivalent to say: $\frac{\mu-\mu_1}{\sigma_1} + \frac{\mu-\mu_2}{\sigma_2} = 0$. Therefore, we have

$$\mu = \frac{\mu_1\sigma_2 + \mu_2\sigma_1}{\sigma_1 + \sigma_2}$$

Proof of Prop 2: From $f(X) = \frac{\sigma_2}{\sigma_1+\sigma_2} f_1(X_1) + \frac{\sigma_1}{\sigma_1+\sigma_2} f_2(X_2)$, we get

$$DX = \frac{\sigma_2}{\sigma_1 + \sigma_2} D_1X + \frac{\sigma_1}{\sigma_1 + \sigma_2} D_2X$$

Now Let us compute D_1X first. Let $z = \frac{x-\mu_1}{\sigma_1}$,

$$
\begin{aligned}
D_1X &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_1} exp(-\frac{(x-\mu_1)^2}{2\sigma_1^2})(x - \frac{\mu_1\sigma_2 + \mu_2\sigma_1}{\sigma_1+\sigma_2})^2 dx \\
&= \int_{-\infty}^{+\infty} \frac{\sigma_1^2}{\sqrt{2\pi}} exp(-\frac{z^2}{2})(z + \frac{\mu_1-\mu_2}{\sigma_1+\sigma_2})^2 dz \\
&= \frac{\sigma_1^2}{\sqrt{2\pi}}(\int_{-\infty}^{+\infty} exp(-\frac{z^2}{2})z^2 dz + 2\frac{\mu_1-\mu_2}{\sigma_1+\sigma_2}\int_{-\infty}^{+\infty} exp(-\frac{z^2}{2})z dz \\
&\qquad + (\frac{\mu_1-\mu_2}{\sigma_1+\sigma_2})^2 \int_{-\infty}^{+\infty} exp(-\frac{z^2}{2})dz) \\
&= \frac{\sigma_1^2}{\sqrt{2\pi}}(\sqrt{2\pi} + 0 + (\frac{\mu_1-\mu_2}{\sigma_1+\sigma_2})^2\sqrt{2\pi}) \\
&= \sigma_1^2(1 + (\frac{\mu_1-\mu_2}{\sigma_1+\sigma_2})^2)
\end{aligned}
$$

Similarly, we get $D_2X = \sigma_2^2(1+(\frac{\mu_1-\mu_2}{\sigma_1+\sigma_2})^2)$. So after some simple calculation, we have $DX = \sigma_1\sigma_2(1 + (\frac{\mu_1-\mu_2}{\sigma_1+\sigma_2})^2), \sigma = \sqrt{DX} = \sqrt{\sigma_1\sigma_2(1 + (\frac{\mu_1-\mu_2}{\sigma_1+\sigma_2})^2)}$

Proof of Prop 4: $\mu = \frac{\mu_1 * m_1 + \mu_2 * m_2}{m_1+m_2} = \frac{\mu_1 * \frac{\sigma^2}{SEM_1^2} + \mu_2 * \frac{\sigma^2}{SEM_2^2}}{\frac{\sigma^2}{SEM_1^2} + \frac{\sigma^2}{SEM_2^2}} = \frac{\mu_1 * SEM_2^2 + \mu_2 * SEM_1^2}{SEM_1^2 + SEM_2^2}$

Proof of Prop 5: $SEM = \frac{\sigma}{\sqrt{m_1+m_2}} = \frac{\sigma}{\sqrt{\frac{\sigma^2}{SEM_1^2} + \frac{\sigma^2}{SEM_2^2}}} = \sqrt{\frac{SEM_1^2 * SEM_2^2}{SEM_1^2 + SEM_2^2}}$

Proof of Proposition 6: If $\mu_1 = \mu_2$, it is straightforward that $\mu = \mu_1$. The remaining part of the proposition is equivalent to prove that when $\mu_1 > \mu_2$, $\mu = \frac{\mu_1\sigma_2 + \mu_2\sigma_1}{\sigma_1+\sigma_2}$, denoted as $g(\sigma_2)$, is an increasing function of σ_2, while when $\mu_1 < \mu_2$, a decreasing function. As the differential of $g(\sigma_2)$ is $g^{'}(\sigma_2) = \frac{(\mu_1-\mu_2)\sigma_1}{(\sigma_1+\sigma_2)^2}$, the result is straightforward.

Proof of Proposition 7: Let $g(\sigma_2)$ denote $\sigma_1\sigma_2(1 + (\frac{\mu_1-\mu_2}{\sigma_1+\sigma_2})^2)$, then σ is an increasing or decreasing function of σ_2 is equivalent to say that $g(\sigma_2)$ is an increasing or decreasing function of σ_2. The differential of $g(\sigma_2)$ is $g^{'}(\sigma_2) = \sigma_1(1 + (\frac{\mu_1-\mu_2}{\sigma_1+\sigma_2})^2) - 2\sigma_1\sigma_2\frac{(\mu_1-\mu_2)^2}{(\sigma_1+\sigma_2)^3}$.

It is obvious that if $\mu_1 = \mu_2$, $g^{'}(\sigma_2) = \sigma_1 > 0$. If $\mu_1 \neq \mu_2$, the $+/-$ sign of $g^{'}(\sigma_2)$ is equivalent to the $+/-$ sign of $\sigma_1(\sigma_1+\sigma_2)^3 + \sigma_1(\sigma_1+\sigma_2)(\mu_1-\mu_2)^2 - 2\sigma_1\sigma_2(\mu_1-\mu_2)^2$, and consequently equivalent to the $+/-$ sign of $(\sigma_1+\sigma_2)^3 - (\sigma_2-\sigma_1)(\mu_1-\mu_2)^2$. When condition $b \leq \sigma_1 + \frac{8\sigma_1^3}{(\mu_1-\mu_2)^2}$ holds, if $\sigma_2 < \sigma_1$, obviously the sign of $g^{'}(\sigma_2)$ is $+$; moreover, if $\sigma_2 \geq \sigma_1$, then $(\sigma_2-\sigma_1)(\mu_1-\mu_2)^2 \leq (b-\sigma_1)(\mu_1-\mu_2)^2 \leq 8\sigma_1^3 \leq (\sigma_1+\sigma_2)^3$, the sign of $g^{'}(\sigma_2)$ is still $+$.

When $\mu_1 \neq \mu_2$ and condition $a \geq \sigma_1 + \frac{(\sigma_1+b)^3}{(\mu_1-\mu_2)^2}$ holds, we have $(\sigma_2 - \sigma_1)(\mu_1 - \mu_2)^2 \geq (a - \sigma_1)(\mu_1 - \mu_2)^2 \geq (\sigma_1 + b)^3 \geq (\sigma_1 + \sigma_2)^3$, so the sign is $-$.

Proof of Proposition 8: The proof is similar to the proof the Proposition 6, except that

$$g^{'}(SEM_2) = \frac{2(\mu_1 - \mu_2)SEM_1^2 SEM_2}{(SEM_1^2 + SEM_2^2)^2}$$

Proof of Proposition 9: Simply notice that $\frac{SEM_1^2+SEM_2^2}{SEM_1^2 SEM_2^2}$ is an increasing function of SEM_2.

Proof of Proposition 10: It is easy to computer that

$$\| f_1 \|_2 = \sqrt{\frac{1}{2\sqrt{\pi}\sigma_1}}, \| f_2 \|_2 = \sqrt{\frac{1}{2\sqrt{\pi}\sigma_2}}$$

and

$$< f_1, f_2 >= \frac{\sqrt{a}exp(-c)}{2\sqrt{\pi}\sigma_1\sigma_2},$$

where

$$a = \frac{2\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, c = \frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}$$

Therefore

$$c(f_1, f_2) = \frac{< f_1, f_2 >}{\| f_1 \|_2 \| f_1 \|_2} = \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} exp(-\frac{(\mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)})$$