EDITORIAL

Opportunities and challenges of clinical research in the big-data era: from RCT to BCT

Stephen D. Wang

Founder, Pioneer Bioscience Publishing Company & AME Publishing Company, Hong Kong, China

At the current stage, the main types of clinical research include prospective study, retrospective study, translational research, and evidence-based study, while the prospective randomized controlled trial (RCT) remains the dominating research type (Figure 1). In a typical RCT, the researchers randomly divide the subjects into several groups provided with different interventions, and then compare the results. With a sufficient number of subjects recruited, RCTs can ensure that all the known and unknown confounding factors have the same effects on each group.

Once a primary end point is set, most RCTs select one indicator as the primary study endpoint, type I error, and type II error, and then the sample size is calculated. Based on the research findings, the researchers will conclude whether the expected result(s) has been achieved. This seemingly 'perfect' study design exposes a commonly neglected problem: the statistical analysis is typically based on the different effectiveness of different interventions provided at baseline. For example, in an RCT evaluating whether the efficacy of a novel chemotherapy drug is superior to the conventional ones when used as the first-line treatment for an advanced cancer, the researchers will often compare the overall survival (OS) between two groups of patients administered with different therapies and then draw a conclusion. During this process, the researchers take it for granted that the intervention provided at baseline is the only variable factor to be considered, and any subsequent interventions (e.g., the second-line and even third-line therapy) will not affect the outcomes (e.g., OS). Although it has been frequently reported that certain subsequent interventions will actually be administered throughout the treatment, they will not

be evaluated in the final analysis. Then, what is the truth? In fact, the vital effects of the subsequent interventions on the outcomes have repeatedly been demonstrated. For example, research has confirmed that the second-line treatment can improve the prognoses. Even if no similar evidence is available, this factor should not be neglected in statistical analysis or at least, its effect on the outcomes should be taken into consideration. Regretfully, the current statistical analysis is focused on the relationship between baseline factors and a specific outcome—such a point-to-point analysis cannot satisfactorily reflect the association of the baseline treatment and its subsequent dynamic processes (during which innumerable data, known as 'big data', are produced) with the outcomes.

More often than not, patients receive multiple confounding interventions from disease progression to death. This consideration gives rise to another question: why not directly use the progression-free survival (PFS) as the primary endpoint? Maybe it works. However, will PFS serves an ideal alternative for OS for these patients? After all, improvement of OS is the core issue. Currently, a somehow popular solution is to analyze whether PFS is an independent predictor of OS in a specific patient population based on the previous publications: if the answer is yes, PFS will then be used as an ideal alternative for OS. The problem remains: is such an analytic method scientifically sound? Obviously, if the multiple interventions (and their dynamic progresses) from disease progression to death are not incorporated in the final analysis, this method is fundamentally far from perfect.

Therefore, a more scientifically sound analysis of the findings of the RCT will highly depend on the scientific analysis of the big data.

We are embracing the big-data era where many tools become available for analyzing the data.

Will RCT maintain its dominating position? The answer is No. It is expected that Big-data Clinical Trial (BCT) will reshape the profiles of clinical research (Figure 2), and BCT will take the place of RCT as the dominating research type. Currently, the RCTs are performed in highly selective subjects who are sampled from an overall population. Though there are findings from
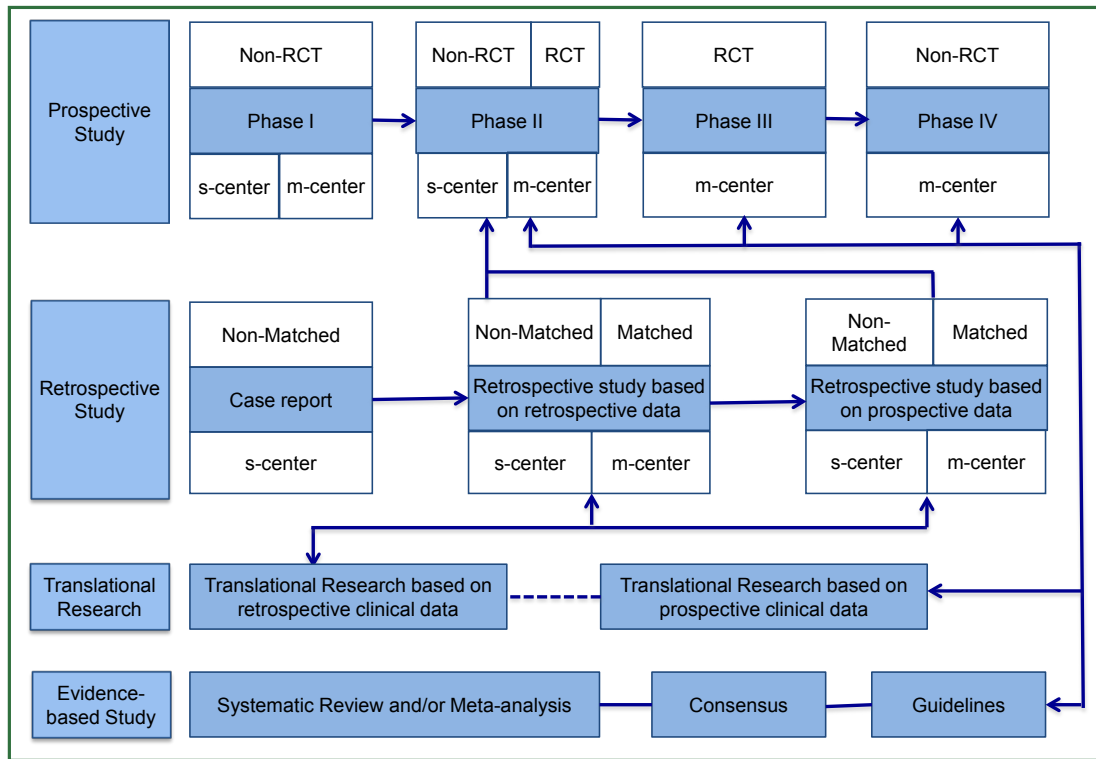
Corresponding to: Stephen D. Wang. Founder, Pioneer Bioscience Publishing Company & AME Publishing Company, FLAT A, 9/F Gold Shine Tower, 346-348 Queen's Road Central, Sheung Wan, Hong Kong, China. Email: dwang@thepbpc.org.

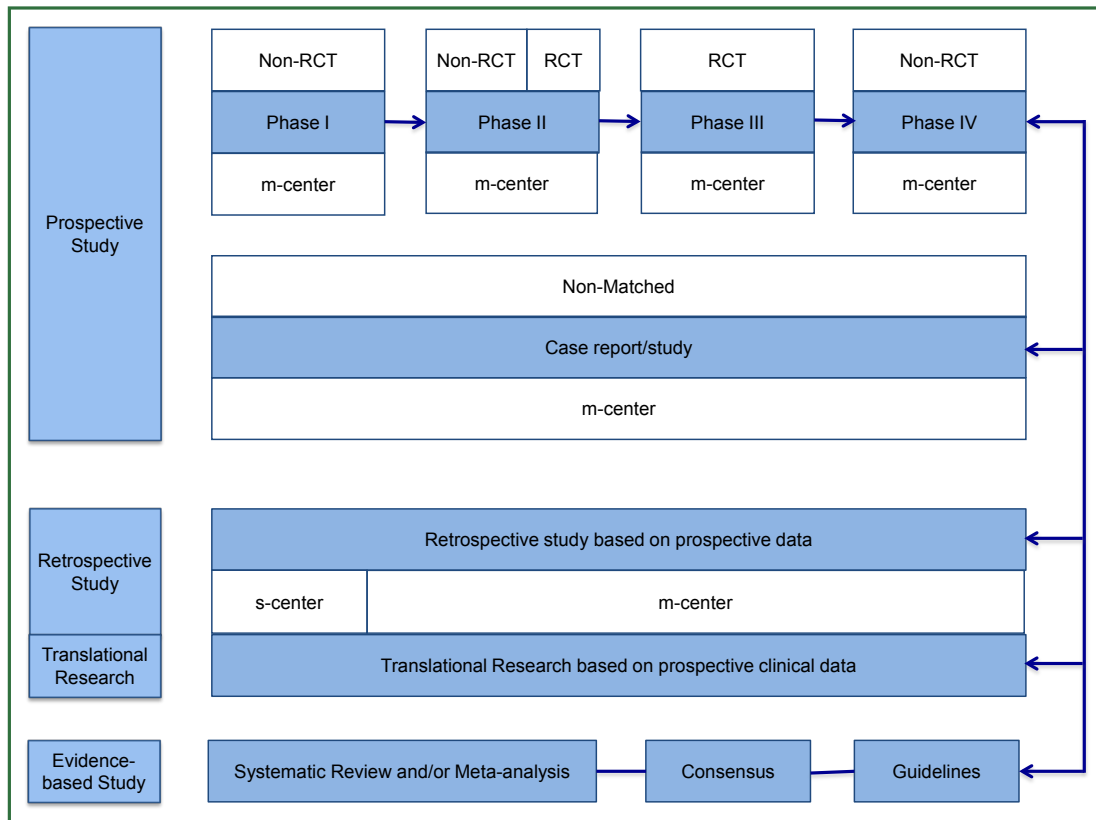**Figure 1.** At the current stage, the main types of clinical research include prospective study, retrospective study, translational research, and evidence-based study. RCT, randomized controlled trial; s-center, single-center; m-center, multi-center.

Figure 2. In the big-data era, the Big-data Clinical Trial (BCT) will reshape the profiles of clinical research and succeed as the dominating research type while RCT will take a back seat to play guiding role. RCT, randomized controlled trial; s-center, single-center; m-center, multi-center.

these small-sample subjects that guide the clinical diagnosis and treatment of a larger sample of population, the coming big-data era entails more analyses conducted on the overall sample (total subsets), and therefore the conclusion drawn according to the corresponding findings based on a more comprehensive setting.

Will RCT disappear for good? The answer is also No. In the big-data era, RCT will play a guiding role. For an interesting finding from a specific RCT, data from a big sample of population will be analyzed for verification before drawing a final conclusion. Therefore in the big-data era, studies based on 'case study' will receive more attention and be elevated to a new level. Many statistical analyses will be based on data obtained from the whole process of diagnosis and treatment of each case prospectively. The quality of data collection during this process will directly affect the reliability of BCT.

In the big-data era, the mysteries of many chronic diseases will be revealed. For instance, research on hypertension currently is still focused on the effectiveness of a specific antihypertensive agent in controlling blood pressure, whereas the so-called control is mainly based on the measurements obtained at several time points. Based on the values of these measurement points,

the researchers linked the blood pressure with a specific long-term outcome (e.g., stroke) and then concluded that drug A, compared with drug B, can effectively control blood pressure and reduce the incidence of stroke. However, an ideal blood pressure control, rather than depending on the measurements at several time points every week or every month, should be the maintenance of a stable blood pressure every day, every hour, or even every minute. The verification of this hypothesis will of course depend on the future BCT.

Of course, BCT will also face many challenges, for example, how to define the BCT? How to define the right to data? How to be a qualified architect for a big-data project? How to prepare for the threats of big data?

In conclusion, in the big-data era, the BCT will reshape the profiles of clinical research and succeed as the dominating research type while RCT will take a back seat to play guiding role.

Acknowledgements