Analysis of Benefit-risk Balance in Decision-making of the Food and Drug Administration for Premarket Approval of Therapeutic Medical Devices

Yoshihiro Muragaki, *, # Miyuki Uematsu, ** Hiroshi Iseki, Mitsuo Umezuma ***

Abstract Compared to the evaluation of new pharmaceutical drugs, the assessments of the design and results of clinical trials for medical devices are not well established. For medical devices, the definition of the benefit-risk balance assessed during approval by regulatory agencies is not clear, which may result in subjectivity of the decision-making process. It is possible to hypothesize that the newly approved medical device should be superior in both risk and efficacy to the already existing device, which is used as control. To test this hypothesis, we performed an independent analysis of the premarket approvals (PMA) of therapeutic medical devices based on assessment review of reports of a regulatory agency, the Food and Drug Administration (FDA). A total of 74 studies that tested various medical devices for PMA were selected. For each clinical trial, the study design was evaluated with particular emphasis on its nature (retrospective or prospective), presence of a control arm, randomization, and masking. We performed an objective analysis of the benefit-risk balance between effectiveness and safety in the test arm compared to that in the control arm, using an original method for data evaluation. Of the 74 studies, 56 (76%) were prospective, 1 was purely retrospective (1%). 15 were mixed (20%), and 2 (3%) did not specify the nature of study. Only 46 studies (62%) included a comparative control group, 26 of which (57%) demonstrated “equivalence” but not “superiority” of the primary effectiveness measure. Depending on the evaluation criteria (mortality, complications, adverse effects, others) the results of safety assessment revealed advantage of the test arm in only 16-38% of comparative studies. The designs of the protocols for testing therapeutic medical devices and the criteria of objective evaluation during approval for broad clinical practice are not standardized. For PMA approval, FDA does not ultimately require better effectiveness and/or safety of the new device compared to the existing control device.

Keywords : premarket approval, benefit-risk balance, decision-making, regulatory science.


1. Introduction
In recent years, dedicated multifaceted industry-government-academia research groups were established in Japan for the development of issues related to the so-called “regulatory science”. This term was originally related to the requirement of scientific approach in the development of new medications, upon which regulatory agencies operate, but its current use is wider and extends not only to issues related to medical pharmacology, but also to the approval of novel diagnostic and therapeutic equipment.

In the USA, premarket approval (PMA) is the most
difficult and most expensive stage of the development process of medical devices, but this process is still not fully optimized or standardized. Compared to evaluation of new pharmaceutical drugs, assessments of the design and results of clinical trials for medical devices seem not well established [1, 2]. While it is generally considered that benefit-risk balance is the most important evaluating factor during approval of the tested medical devices by regulatory agencies, its strict definition is not clear and methods of its assessment during decision making appears somewhat subjective. At present, different groups of researchers in Japan are trying to define their own strategies to bring new medical equipment into clinical use as fast as possible. Under such conditions, the Japanese regulatory agency attempts to make the approval process transparent and promotes open discussions between representatives of industries, governmental bodies, and academic institutions. Since the decision-making process during development of new medical devices constitutes a very important aspect of the regulatory science [3], our group decided to perform an analysis of approval criteria for new medical equipment used by the regulatory agency, using open access information available through the internet.

Initially we planned to use the Japanese database, but it contains too limited number of approved domestic
devices, which do not permit reliable evaluation of decision-making during the approval process. In order to overcome this problem, we decided to investigate a wide range of different devices as well as similar equipment produced by different manufacturers. Therefore, we performed an independent analysis of the approval of therapeutic medical devices based on a review of the reports of a regulatory agency, the Food and Drug Administration (FDA).

We hypothesized that the newly approved medical device should be superior in both risk and efficacy to the already existing device used as control. Special emphasis was put on the design of the clinical trials. Herein, we report the final results of our investigation [4].

2. Methods

For the purpose of the present study, a retrospective search of FDA database available on the internet (http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cftopic/mda/) was done. This site provides all essential information of recently approved devices under “CDRH Consumer Information.”

2.1 Data selection

Between 2000 and 2008, 549 medical devices received PMA by FDA. However, 295 of them either overlapped (similarly named in different categories) or were designed specifically for diagnostic purposes. Of the remaining 259 devices, 180 had similar mechanisms of action within the same medical field and were therefore excluded. After excluding the above-mentioned 475 PMA reports of medical devices, 74 representative reports were selected for the present analysis (Fig. 1). However, 28 of the 74 reports did not note the use of a comparative control group.

2.1.1 Analysis

In the 74 selected reports on PMA of medical devices, the information on animal studies during preclinical testing was initially checked. For clinical trials of each medical device, the study design was evaluated with particular emphasis on its nature (retrospective or prospective), presence of control arm, randomization, and masking.

Detailed analysis was conducted on 46 devices that had undergone controlled clinical trials before PMA. In this group, we conducted an objective analysis of the benefit-risk balance between effectiveness and safety in the test arm compared to that in the control arm. For this purpose, effectiveness was estimated by primary or secondary endpoint, and safety by the presence of adverse events and/or side effects. An original method of evaluation was used. An effectiveness score of “+1” was assigned if the endpoint of the test arm was significantly better than that of the control arm, a score of “0” if no difference was found, and a score of “−1” if the endpoint was worse in the test arm. Correspondingly, if a trial had several endpoints, the scores were summed (for two endpoints both of which demonstrated advantages in the test arm, the score was: [+1] + [+1] = 2). Safety was assessed in the same manner. A safety score of “+1” was assigned if the adverse events and/or side effects were significantly less prominent in the test arm compared to control arm. All devices were classified by the effectiveness and safety scores. Additionally, a “regulatory science

---

Fig. 1  Flow chart of report selection for the present study.
score” (RS score) was defined as the sum of effectiveness score and safety score.

2.1.2 Statistics
Chi-square test was used for statistical comparison of effectiveness, safety, and RS scores. The level of significance was determined for two-tailed P-values less than 0.05.

3. Results
Among 74 reports on PMA of medical devices, information on preclinical animal studies was available in 46 (62%). Overall, 56 studies (76%) were prospective, 1 was purely retrospective (1%), 15 were mixed (20%), and 2 (3%) did not specify the nature of study. There were 29 randomized and 16 masked studies. Eleven of 16 masked studies were double-blind, 4 were single-blind, and 1 was mixed. Overall, 62 studies (84%) were multicenter, 4 were single center (5%), 5 were mixed (7%), and 3 (4%) did not specify. The number of participating centers was provided in 64 reports; the mean number of centers was 20, and the mean number of enrolled patients was 317.

3.1 Studies without a control arm
As mentioned above, only 46 reports (62%) contained a comparative control group. Among 28 reports without comparative control, 13 did not use control group at all, 12 used either historical control or a combination of historical and prospectively designed control groups, whereas 3 studies did not provide any reliable information on this issue. Studies without a control arm were related to testing of a heart valve, cardiac defibrillators, aortic and mitral prostheses, a replaceable heart, a cardiac pacemaker, a ventricular assist device for children (with a clearly defined exemption from the usual study standards), a renal stent, a control system for Parkinson disease, a cardiac ablation device, breast implants, a urinary prosthesis to control incontinence, a spinal cord stimulation system, and a prosthesis for jaw surgery. More than half of all studies without control tested potentially life-saving devices related to cardiac problems.

In studies without a control arm, effectiveness was usually assessed relative to the baseline levels of pretreatment parameters, whereas success was defined as a predetermined level of improvement of these parameters. In some cases, this level was determined from the published data in the literature and case studies. To facilitate the evaluation process, FDA applied Bayesian statistics to provide prospective attainment of the targeted parameters of interest. However, in many occasions, it was not possible to clarify how the required improvement relative to baseline was determined. In such cases, it was often stated that safety was “satisfactory” in terms of complications.

3.2 Studies with a control arm
Statistical analysis was used to compare primary endpoints in 39 of 46 studies with a control arm (85%). Secondary endpoints were evaluated in 31 studies, 25 of which applied statistical analysis to compare their data. Regarding the criterion for primary effectiveness, 26 of 46 (57%) studies with a control arm demonstrated “equivalence”, but not “superiority” (Table 1). Among studies that compared secondary endpoint with a control, 8 devices were approved after demonstration of “equivalence” only (demonstration of “superiority” was not always required by study design).

Results of safety assessment in the 46 studies with a control arm are presented in Table 2. Overall, depending on the evaluation criteria, the test arm demonstrated advantages in 16-38% of studies only.

3.2.1 Evaluation of benefit-risk balance
Overall results of the evaluation of benefit-risk balance are shown in Fig. 2 as a two-dimensional map.

Using our original method to evaluate 46 studies with a control arm as described above, 8 devices (17%) demonstrated superior scores for both effectiveness and safety, 10 (22%) had superior effectiveness score and equivalent safety score, 9 (20%) showed superior effectiveness score but inferior safety score, and 14 (30%) had equivalent scores for both effectiveness and safety. The 14 devices with equivalent effectiveness and safety scores

---

Table 1 Assessment of effectiveness in 46 studies with a control arm which resulted in premarket approval of the therapeutic medical devices.

<table>
<thead>
<tr>
<th>type of endpoints [n(%)]</th>
<th>parameters in the test group [n(%)]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>superior</td>
</tr>
<tr>
<td>primary</td>
<td></td>
</tr>
<tr>
<td>46(100)</td>
<td>20(43)</td>
</tr>
<tr>
<td>secondary</td>
<td></td>
</tr>
<tr>
<td>31(100)</td>
<td>21(68)</td>
</tr>
</tbody>
</table>

n, number of studies.

Table 2 Assessment of safety in 46 studies with a control arm, which resulted in premarket approval of the therapeutic medical devices.

<table>
<thead>
<tr>
<th>evaluation criteria [n(%)]</th>
<th>parameters in the test group [n(%)]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>superior</td>
</tr>
<tr>
<td>mortality</td>
<td></td>
</tr>
<tr>
<td>11(100)</td>
<td>3(27)</td>
</tr>
<tr>
<td>complications</td>
<td></td>
</tr>
<tr>
<td>31(100)</td>
<td>5(16)</td>
</tr>
<tr>
<td>adverse events</td>
<td></td>
</tr>
<tr>
<td>17(100)</td>
<td>6(35)</td>
</tr>
<tr>
<td>other</td>
<td></td>
</tr>
<tr>
<td>8(100)</td>
<td>3(38)</td>
</tr>
</tbody>
</table>

n, number of studies.
showed smaller RS score compared to the other 32 devices ($P=0.035$). In contrast, the group of devices with superior effectiveness score (27 of 46 cases; 59\%) had greater RS score compared to the other 19 devices ($P=0.013$).

4. Discussion

Protocols for testing novel medical devices cannot simply copy the equivalent protocols for the evaluation of new pharmaceutical drugs, which nearly always include a control group and use randomization. Our analysis of PMA of therapeutic medical devices granted by FDA between 2000 and 2008 demonstrated that 20\% of the studies had only single arm, whereas the majority were not blinded or randomized. Such findings are consistent with the previous reports focusing exclusively on cardiovascular tools [1, 2]. The studies without controls involved devices for a variety of health conditions, while more than a half of the studies tested potentially life-saving devices related to cardiac problems. Therefore, it is possible to suspect that in such cases the testing requirements were simplified in order to allow quick approval of the devices with high life-saving potential for patients with heart diseases. In fact, an official statement of such policy was issued for trials of ventricular assist device for children. Nevertheless, testing of some other devices such as breast implants also did not include a control group for no obvious reason, since these devices cannot be considered to save life or maintain health.

Only 16 of the 74 studies analyzed were described as masked or partially masked (double- or single-blinded), while it is possible that such description was missed or omitted in some reports. This aspect is also different from the testing of pharmaceutical drugs for FDA approval, which must be double-blinded as a rule. Blind testing of many devices is evidently not possible. For example, if a placebo (sham) device is used in a trial, it cannot be masked easily. Nevertheless, since the patient is aware of the type of procedure performed in an unmasked fashion, when the patient’s subjective scoring (such as quality of life and pain level) is included in evaluation of the device, the possible source of error should be taken into consideration. Moreover in many cases, use of the novel device requires special skills, and the learning curve may have potential influence on the results of the clinical trial, particularly in the surgical field.

For FDA approval, testing of pharmaceutical drugs requires unequivocal demonstration of superiority in terms of efficacy and/or safety compared to the control group. Meanwhile, our analysis of benefit-risk balance revealed that 30\% of the devices approved by FDA had equivalent scores for both effectiveness and safety compared to control group. While no approved therapeutic device showed inferior effectiveness compared to control, many of the approved devices did not demonstrate “superiority”. In fact, regarding the primary endpoint, 57\% of studies with a control arm demonstrated only “equivalence” compared to the control group. These devices were frequently associated with equal or even inferior safety compared to control, which strongly suggests that FDA mainly regards effectiveness, or even
potential effectiveness, of the device as the primary criterion for approval, and generally considers effectiveness as being more important than safety in terms of complications, adverse events, or other evaluation criteria (certainly, beside mortality). Therefore, a proof of “equivalence” for the safety parameters in the test group compared to control was usually sufficient for approval, and “superiority” was required in a few cases only.

Ideally, as it was stated in the primary hypothesis of our study, the novel therapeutic device should be superior in both risk and efficacy to the already existing device used as control (Fig. 3). However, the evaluation process of clinical trials may be rather complicated. Primary end points as well as the optimal number for such testing are not well established. To obtain objective evaluation of the benefit-risk balance, we presented an original method of data evaluation in the present study. Our method takes into consideration both comparative effectiveness and safety, and bases on calculation of the composite RS score.

A limitation of the present study is that only FDA-approved therapeutic devices were assessed. The inability to analyze cases in which PMA was declined does not permit examination of all details of the decision-making process. Moreover, the primary and secondary end points were not separated during objective analysis of the available data. Finally, it should be emphasized that it is difficult to determine precisely the effectiveness and safety of new devices and their influence on the national health system based on the PMA trials only. Recently FDA initiated postmarket surveillance of approved devices, but this aspect was not addressed in the present study. Further investigations should take these concerns into consideration.

5. Conclusion

The designs of protocols for testing therapeutic medical devices and the criteria of objective evaluation during approval for broad clinical practice are not standardized. According to the results of our investigation, FDA does not ultimately require superiority in effectiveness and/or safety of new devices compared to existing controls for PMA approval.

Acknowledgement

The authors are thankful to Dr. Mikhail Chernov for help with manuscript preparation.

References


Yoshihiro Muragaki

He received the M.D. degree from Kobe University in 1986 and Ph.D. degrees in Medical Science from Tokyo Women's Medical College in 1997. He was a certified neurosurgeon from 1992 at the department of Neurosurgery, Tokyo Women's Medical University and an associate professor from 2009 at the Faculty of Advanced Techno-Surgery, Tokyo Women's Medical University. He has been a Professor in the same faculty from 2011. He engages in brain tumor treatment, intraoperative MRI, and regulatory science. He is a member of Japanese Neurosurgical Society, the Japan Society for Neuro-Oncology, and Japanese Society for Medical and Biological Engineering.

Fig. 3 Schematic representation of the proposed method for evaluation of novel therapeutic devices.
Miyuki UEMATSU
Miyuki UEMATSU received the B.Eng., M.Eng., and Ph.D. degrees in Biomedical Engineering from Waseda University in 2002, 2004 and 2007, respectively. She became a JSPS Research Fellow DC2 in 2006, and an Assistant Researcher for Welfare Engineering in Waseda University in 2007. She was a Researcher for Medical Devices from 2007 to 2010 in National Institute of Health Sciences in Japan. Since 2011, she has been a Senior Researcher in the same institution. She engages in application of virtual reality to medicine, development of surgical navigation systems for thracoabdominal region, evaluation and simulation on implantable medical devices for cardiovascular surgery. She has been a member of the society of IEEE, the Japanese Society for Medical and Biological Engineering, and the Japanese Society of Computer Aided Surgery.

Hiroshi ISEKI
Hiroshi ISEKI received the M.D. degree from Tokyo University in 1974 and the Ph.D. degree in medical science from Tokyo Women’s Medical College in 1982. He has been certified from Japanese Board of Neurological Surgery in 1983. He was an associate professor and director of Faculty of Advanced Techno-Surgery (FATS), Institute of advanced Biomedical Engineering & Science, Graduate School of Medicine, Tokyo Women’s Medical University in 2001 and a professor of FATS in 2006. He is a professor of Joint Graduate School of Tokyo Women’s Medical University and Waseda University from 2010. He engages in computer aided surgery, Intelligent operating theater, and Robotic surgery. He is a member of Japan Neurosurgical Society, Japan Society of Computer Aided Surgery, and Japan Society for Medical and Biological Engineering.

Mitsuo UMEZU
He is a biomedical engineer in the cardiovascular research area. He received two PhDs; Doctor of Engineering from Waseda University and Doctor of Medical Science from Tokyo Women’s Medical University. He was appointed as a research associate and laboratory head of a department of Artificial Organs at the National Cardiovascular Center Research Institute, Osaka between 1979 and 1988. Then, he worked as the first project leader of Australian Artificial Heart Program at Sydney St.Vincent’s Hospital. He has been a professor of the Department of Mechanical Engineering, Waseda University since 1992. Now, he is a director of Center for Advanced Biomedical Sciences, TWIns, and a department head of Joint Graduate School with Tokyo Women’s Medical University. His recent research interest includes development and evaluation of artificial organs and regulatory science for medical devices.