Comparison of Effects in Randomized Controlled Trials With Observational Studies in Digestive Surgery

Satoru Shikata, MD,*† Takeo Nakayama, MD, PhD,‡ Yoshinori Noguchi, MD, MPH,§ Yoshinori Taji, MD,† and Hisakazu Yamagishi, MD, PhD*

Objectives: To compare the results of randomized controlled trials versus observational studies in meta-analyses of digestive surgical topics.

Summary Background Data: While randomized controlled trials have been recognized as providing the highest standard of evidence, claims have been made that observational studies may overestimate treatment benefits. This debate has recently been renewed, particularly with regard to pharmacotherapies.

Methods: The PubMed (1966 to April 2004), EMBASE (1986 to April 2004) and Cochrane databases (Issue 2, 2004) were searched to identify meta-analyses of randomized controlled trials in digestive surgery. Fifty-two outcomes of 18 topics were identified from 276 original articles (96 randomized trials, 180 observational studies) and included in meta-analyses. All available binary data and study characteristics were extracted and combined separately for randomized and observational studies. In each selected digestive surgical topic, summary odds ratios or relative risks from randomized controlled trials were compared with observational studies using an equivalent calculation method.

Results: Significant between-study heterogeneity was seen more often among observational studies (5 of 12 topics) than among randomized trials (1 of 9 topics). In 4 of the 16 primary outcomes compared (10 of 52 total outcomes), summary estimates of treatment effects showed significant discrepancies between the two designs.

Conclusions: One fourth of observational studies gave different results than randomized trials, and between-study heterogeneity was more common in observational studies in the field of digestive surgery.

Methods

Search for Meta-Analyses of Randomized Controlled Trials and Selection of Topics

Meta-analyses of randomized controlled trials in digestive surgery that had been published up to April 2004 were selected as topics in this study. Retrieved articles were judged suitable for use as topics only if all the following criteria were met: 1) meta-analysis of randomized controlled trials; 2) investigating digestive surgery; 3) assessing the treatment effects of at least one operative intervention versus any other intervention (operative or nonoperative); and 4) subjects were human. Searches were not limited to English language articles (any language). Studies were excluded if the main purpose was not evaluation of treatment effect, such as diagnosis. A literature search was

From the *Department of Digestive Surgery, Kyoto Prefectural University of Medicine, Kyoto, Japan; †Department of Clinical Epidemiology, Kyoto University Graduate School of Medicine, Kyoto, Japan; ‡Department of Health Informatics, Kyoto University School of Public Health, Kyoto, Japan; §Department of Medicine, Fujita Health University School of Medicine, Aichi, Japan.

Supported by a Health and Labour Sciences Research Grant (Health Technology Assessment) from the Ministry of Health, Labour and Welfare, Japan.

Reprints: Takeo Nakayama, MD, PhD, Department of Health Informatics, Kyoto University School of Public Health, Konoe-cho, Yoshida, Sakyoku, Kyoto 606-8501, Japan. E-mail: nakayama@pbh.med.kyoto-u.ac.jp.

Copyright © 2006 by Lippincott Williams & Wilkins

ISSN: 0003-4932/06/24405-0668
DOI: 10.1097/01.sla.0000225356.04304.bc

(Ann Surg 2006;244: 668 – 676)
performed using the PubMed (1966 to April 2004), EMBASE (1986 to April 2004) and Cochrane Library (Issue 2, 2004) databases. A computer-assisted search was conducted using the following combination of Medical Subject Heading Terms and text words: “surgical procedures, operative,” “digestive system surgical procedures,” “randomized,” “random,” “meta-analysis,” and “review.” A manual search was also performed using references from the retrieved review articles.

Search for Observational Studies for Meta-Analysis

If meta-analyses of both randomized and observational studies had been performed on the same topic in each selected review article, the results could be used for comparison. However, if meta-analysis of observational studies had not been performed, we attempted to perform that by ourselves. Thus, when a meta-analysis of observational studies could not be identified in the selected review article, we needed to search for such meta-analyses while gathering observational studies under the following process.

For meta-analyses of observational studies, we first searched observational studies for all selected topics. In each topic, the same inclusion criteria used for meta-analysis of randomized controlled trials were used, with the exception of study design. Observational study designs were used if they could be categorized as prospective nonrandomized studies, retrospective cohort studies, case-control studies, case series with control groups, or other unspecified designs (provided a control group was used). A literature search was performed using the PubMed (1966 to April 2004), EMBASE (1986 to April 2004) and Cochrane Library (Issue 2, 2004) databases. PubMed contains no search term for observational studies, so a text-word strategy was used to search for “observational,” “nonrandomized,” “case series,” “case control study,” “cohort,” “retrospective,” and “prospective.” In addition, a manual search was performed using references from the retrieved review articles. We also attempted to contact as many experts from the review articles as possible.

Data Extraction and Selection of Outcomes

All available binary data were extracted from the outcomes of the gathered observational studies. Data extraction was performed after translation of the article into English if the article had not been written in English or Japanese. Up to this point, 2 authors (S.S., T.N.) undertook the literature searches and data extraction independently, and disagreements were resolved by consensus.

For final inclusion of a topic in the present evaluation, binary data for the same outcome had to be available from at least one randomized trial and at least one observational study. When primary outcomes had been defined in the review article, these were used for the main comparison. Whenever the primary outcome was unclear, the outcome that was considered a priori as the most clinically important was selected, using consensus among the data extractors. In digestive surgery, mortality was generally given priority in clinical importance over other outcomes.

Statistical Analysis

For all selected topics, data from observational studies were combined. Generally, the fixed-effects model weighted by Peto’s odds ratio method or the Mantel-Haenszel method was used for data pooling, followed by a test of heterogeneity. Heterogeneity between studies was assessed using Q statistics. Given the low power of this test, a significance level of 0.10 was used, rather than 0.05. If the hypothesis of heterogeneity was accepted, the random-effects model using the DerSimonian-Laird method was used. However, this study sought to compare summary estimates of randomized controlled trials with observational studies under equivalent conditions to the maximum extent possible. Thus, when performing meta-analysis of observational studies, we used the same method that had been used in the meta-analysis of randomized controlled trials. In this study, the quantity $I^2$ was used for assessing heterogeneity between trials in meta-analyses, calculated as: $I^2 = \left[\frac{Q - df}{Q}\right] \times 100$, where $Q$ is the $\chi^2$ statistic and $df$ is the degrees of freedom. A value greater than 50% may be considered indicative of substantial heterogeneity.

Although pooled odds ratio or pooled relative risk could be used as the indicator of summary estimates of outcomes, the present study used the same indicator that had been used in the meta-analysis of randomized controlled trials. In this context, odds ratios and relative risks will inevitably be similar in magnitude, as the rates of outcome events are low. Relative risks were therefore considered as odds ratios in comparisons of summary estimates. Confidence intervals were always calculated at 95%. When one arm of an outcome contained no events, this was considered a “zero cell” in the $2 \times 2$ table. Zero cells create problems in computing ratio measures of treatment effect. This problem was dealt with using a common method of adding 0.5 to each cell of the $2 \times 2$ table for the trial.

To evaluate concordance between the results of randomized and observational studies, the following analyses were performed: 1) assessment of the number of cases in which the summary estimates of the observational studies suggested an effect at least double that of the randomized trials; and 2) evaluation of whether differences in the summary odds estimates of randomized controlled trials and observational studies for the same topic were larger than what would be expected by chance alone. To accomplish this, $Z$ scores were calculated as follows:

$$Z = \left[\ln(OR_{RCT}) - \ln(OR_{OBS})\right] / \left\{\text{var}[\ln(OR_{RCT})] + \text{var}[\ln(OR_{OBS})]\right\}^{1/2},$$

where $\ln(OR_{RCT})$ is the natural logarithm of the odds ratio or relative risk of randomized controlled trials, $\ln(OR_{OBS})$ is the natural logarithm of the odds ratio or relative risk of observational studies, and var is variance. A $Z$ score above 1.96 or less than $-1.96$ suggests a nonrandom difference between randomized controlled trials and observational studies (0.05 level of statistical significance).
All statistical analyses were performed using STATA statistical software version 8.1 (STATA Corporation, College Station, TX).

RESULTS

Characteristics of Topics, Observational Studies

A literature search was first performed to select meta-analyses of randomized controlled trials for the topics, identifying 1184 potentially relevant articles. The process finally identified and selected 15 meta-analyses of randomized controlled trials for digestive surgical topics in this research (Fig. 1).7,18–31 Three of the 15 reviews contained two topics.21,30,31 Thus, 18 topics were identified for comparison of summary estimates between randomized controlled trials and observational studies (Table 1).

Meta-analyses of observational studies could not be identified for 10 of the 18 topics (topics 2, 3, 8–13, 17, and 18), so additional meta-analyses were required. Meta-analyses of observational studies had been identified for the remaining 8 topics (topics 1, 4–7, and 14–16), and the results were used for comparisons in this study.

For meta-analyses of observational studies for the 10 topics without existing meta-analyses, a literature search was performed and 111 observational studies were selected from 10,960 articles using the process outlined in Figure 2. Of the 111 selected articles, 17 had not been written in English or Japanese, instead appearing in 7 different languages, and the 2 trial assessors therefore abstracted data from the articles after translation into English by independent translators. A total of 52 common outcomes for both randomized controlled and observational studies were available for comparison in this study.

Using the described processes, 52 outcomes of 18 topics were investigated in 276 original articles (96 randomized trials, 180 observational studies) with a total of 101,170 study patients (Table 1). The 180 observational studies comprised 36 prospective and 144 retrospective studies. Randomized and observational studies on the same topic generally administered treatment in the same way and outcome measures were similarly defined.

Between-Study Heterogeneity

Data on between-study heterogeneity using the $I^2$ statistic were available for all 10 meta-analyses of observational studies that we performed specifically for the present study (topics 2, 3, 8–13, 17, and 18). Conversely, data had not been described in 8 of the remaining meta-analyses that had been reported (topics 1, 4–7, and 14–16). In primary outcomes of 16 topics, significant heterogeneity was noted between randomized controlled trials in 1 of 9 topics (11.1%). Significant between-study heterogeneity was identified between observational studies in 5 of 12 topics (41.7%). There was no significant difference between the rates of heterogeneity ($P = 0.18$ by Fisher exact test).

Comparison of Primary Outcomes

In almost all topics, the primary outcome defined in the review or decided by author consensus was mortality. However, in topics dealing with safety of procedures, such as appendectomy and operation for fissure-in-ano, one of the complications, such as risk of wound infection or persistence...
of fissure, was considered as a more appropriate primary outcome.

In 16 of 18 topics, primary outcomes could be compared between observational studies and randomized controlled trials. These summary estimates and associated 95% confidence intervals are shown in Figure 3. One of 16 primary outcomes displayed a magnitude of effect in the combined observational studies that was outside the 95% confidence interval for the

<table>
<thead>
<tr>
<th>Identification No.</th>
<th>Topic</th>
<th>Randomized Controlled Trial</th>
<th>Observational Study (no. of studies)</th>
<th>No. of Comparable Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Closed postoperative peritoneal lavage vs. no lavage for generalized peritonitis</td>
<td>Leiboff et al (1987)</td>
<td>4 (173)</td>
<td>2/6</td>
</tr>
<tr>
<td>2</td>
<td>Splenorenal shunt vs. endoscopic sclerotherapy in prevention of variceal rebleeding</td>
<td>Spina et al (1992)</td>
<td>4 (310)</td>
<td>0/2</td>
</tr>
<tr>
<td>3</td>
<td>Routine drainage vs. no drainage after elective colorectal surgery</td>
<td>Urbach et al (1999)</td>
<td>4 (414)</td>
<td>0/5</td>
</tr>
<tr>
<td>4</td>
<td>Anal stretch vs. sphincterotomy for fissure-in-ano</td>
<td>Nelson et al (1999)</td>
<td>6 (328)</td>
<td>0/4</td>
</tr>
<tr>
<td>5</td>
<td>Open vs. closed lateral sphincterotomy for fissure-in-ano</td>
<td>Nelson et al (1999)</td>
<td>2 (140)</td>
<td>0/4</td>
</tr>
<tr>
<td>6</td>
<td>Laparoscopic vs. open appendectomy for acute appendicitis</td>
<td>Benson et al (2000)</td>
<td>16 (1703)</td>
<td>3/4</td>
</tr>
<tr>
<td>7</td>
<td>Transthoracic vs. transhiatal resection for carcinoma of the esophagus</td>
<td>Hulscher et al (2001)</td>
<td>3 (138)</td>
<td>3/18</td>
</tr>
<tr>
<td>8</td>
<td>Hand-sewn vs. stapled esophagogastric anastomosis after esophagyectomy</td>
<td>U尔斯el et al (2001)</td>
<td>5 (467)</td>
<td>2/8</td>
</tr>
<tr>
<td>9</td>
<td>Posterior vs. anterior route of reconstruction after esophagyectomy</td>
<td>U尔斯el et al (2001)</td>
<td>6 (342)</td>
<td>0/3</td>
</tr>
<tr>
<td>10</td>
<td>Pyloroplasty vs. no drainage in gastric reconstruction after esophagyectomy</td>
<td>U尔斯el et al (2002)</td>
<td>3 (347)</td>
<td>0/2</td>
</tr>
<tr>
<td>13</td>
<td>Stapled vs. conventional hemorhoidectomy</td>
<td>Sutherland et al (2002)</td>
<td>7 (591)</td>
<td>2/5</td>
</tr>
<tr>
<td>14</td>
<td>Extended vs. limited lymph node dissection for adenocarcinoma of the stomach</td>
<td>McCulloch et al (2003)</td>
<td>3 (1729)</td>
<td>5/8</td>
</tr>
<tr>
<td>15</td>
<td>Open (Hasson type) vs. closed (needle/trocar) access in laparoscopic surgery</td>
<td>Merlin et al (2003)</td>
<td>4 (302)</td>
<td>4/6</td>
</tr>
<tr>
<td>16</td>
<td>Direct trocar vs. closed (needle/trocar) access in laparoscopic surgery</td>
<td>Merlin et al (2003)</td>
<td>3 (665)</td>
<td>0/2</td>
</tr>
<tr>
<td>17</td>
<td>Early vs. delayed open cholecystectomy for acute cholecystitis</td>
<td>Papi et al (2004)</td>
<td>9 (916)</td>
<td>2/16</td>
</tr>
<tr>
<td>18</td>
<td>Early vs. delayed laparoscopic cholecystectomy for acute cholecystitis</td>
<td>Papi et al (2004)</td>
<td>3 (228)</td>
<td>7/9</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>96 (10,493)</td>
<td>180 (90,677)</td>
</tr>
</tbody>
</table>
combined randomized controlled trials (topic 14). In 4 of 16 primary outcomes, summary estimates from observational studies were at least double those from randomized controlled trials (topics 7, 8, 15, and 17). The converse occurred in 3 topics (topics 11, 14, and 16) (exact \( P = 0.45 \) by Wilcoxon test). Evaluation by Z score revealed significant discrepancies between randomized trials and observational studies for 4 of 16 primary outcomes (topic 7, \( Z = -4.28 \); topic 8, \( Z = -2.36 \); topic 11, \( Z = 2.19 \); topic 14, \( Z = 4.34 \)).

**Comparison of All Outcomes**

All summary estimates for 52 outcomes of 18 topics are shown in Table 2. Three types of calculation model were used: random effects calculation using the DerSimonian-Laird method; and fixed effects calculation using Peto’s odds ratio method or the Mantel-Haenszel method. In 21 of 52 outcomes, relative risk was evaluated rather than odds ratio in meta-analyses of observational studies, as the original meta-analyses of randomized controlled trials had used relative risks for evaluations.

In 9 of 52 outcomes, summary estimates from observational studies were at least double those from randomized controlled trials. The converse occurred in 10 outcomes (exact \( P = 0.943 \) by Wilcoxon test). Evaluation by Z score revealed significant discrepancies between randomized trials and observational studies in 10 of 52 outcomes.

Overall, these data suggest that about one fourth of observational studies gave different results than randomized trials.

**DISCUSSION**

Using data from 276 articles in 18 topics, summary estimates were compared between randomized controlled trials and observational studies in digestive surgery. Significant between-study heterogeneity occurred more often between observational studies than between randomized controlled trials. One fourth of the summary estimates of treatment effects in randomized controlled trials and observational studies differed significantly from each other. From this study, observational studies in digestive surgery tend to have similar results to those by randomized controlled trials. At least, they do not tend to overestimate or underestimate more than randomized controlled trials.

Our findings support the conclusions of earlier evaluations in the 1970s and 1980s. In 2001, Ioannidis et al investigated 45 diverse pharmacologic and surgical topics in 408 articles and concluded that observational studies tend to indicate larger treatment effects (28 of 45 topics vs. 11 of 45 topics) and between-study heterogeneity is more frequent among observational studies than among randomized controlled trials (41% vs. 23%). On the other hand, previous studies by Benson and Hartz and Concato et al reached the opposite conclusion. Benson and Hartz investigated 19 diverse pharmacologic and surgical treatments in 136 articles and found little evidence of larger or differing estimates of treatment effects in observational studies compared with randomized controlled trials. Concato et al evaluated 5 clinical topics and 99 articles, concluding that well-designed observational studies do not systematically overestimate the magnitude of treatment effects when compared with randomized controlled trials on the same topic.

All these previous studies have made substantial contributions toward identifying the problems caused by differing study designs. However, conclusions have inevitably been in the form of general statements, as the studies addressed diverse topics in various clinical fields. The present study was limited to a single clinical field, digestive surgery and thus offers two advantages over previous studies: a more exhaustive search is possible in studies of discrete clinical fields; and higher applicability to clinical practice than a general statement.

In 25% of digestive surgical topics, summary estimates of treatment effects in observational studies yielded different results than randomized trials, but both designs reached similar results in the remaining topics. This may be attributable to various factors. First, quality of surgical randomized controlled trials is low according to some review articles and...
may be so low that the essential contents of randomized trials do not differ from those of observational studies. Second, for most topics, sample sizes may be too small to detect clinically important differences between the results of two types of study. Actually, 12 of 18 topics used fewer than 500 randomized patients. Combined with the use of a rare end-point, mortality, we could expect to see very large confidence intervals in the randomized evidence. The wide confidence intervals mean that demonstrating any significant discrepancy between the two designs will be very difficult.

This study examined not only primary outcomes, but also the secondary outcomes. Generally, results about

![FIGURE 3. Comparison of primary outcomes between observational studies and randomized controlled trials. This figure is based on data from 13 review articles and 10 meta-analyses of observational studies by the authors. OR, odds ratio; RR, relative risk; CI, confidence interval. *Outcome reporting relative risk rather than odds ratio.](image)
<table>
<thead>
<tr>
<th>Identification No.</th>
<th>Topic</th>
<th>Outcome</th>
<th>Randomized Controlled Trial</th>
<th>Observational Study</th>
<th>Calculation Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No. of Studies</td>
<td>Summary Estimate OR (95% CI)</td>
<td>No. of Studies</td>
<td>Summary Estimate OR (95% CI)</td>
</tr>
<tr>
<td>1</td>
<td>Closed postoperative peritoneal lavage vs. no lavage for generalized peritonitis</td>
<td>Mortality</td>
<td>4</td>
<td>0.65 (0.30–1.40)</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>Splenorenal shunt vs. endoscopic sclerotherapy in the prevention of variceal rebleeding</td>
<td>Rebleeding*</td>
<td>4</td>
<td>0.16 (0.10–0.27)</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Routine drainage vs. no drainage after elective colorectal surgery</td>
<td>Mortality, Anastomotic leak, Pulmonary complication, Wound infection, Persistence of fissure*</td>
<td>4</td>
<td>1.70 (0.87–3.30)</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Anal stretch vs. sphincterotomy for fissure-in-ano</td>
<td>Flatus incontinence*</td>
<td>4</td>
<td>6.63 (2.06–21.3)</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>Open vs. closed lateral sphincterotomy for fissure-in-ano</td>
<td>Persistence of fissure*</td>
<td>4</td>
<td>2.05 (0.52–8.11)</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>Laparoscopic vs. open appendectomy for acute appendicitis</td>
<td>Wound infection</td>
<td>16</td>
<td>0.30 (0.19–0.47)</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>Transthoracic vs. transhiatal resection for carcinoma of the esophagus</td>
<td>Mortality*, Cardiac complication*, Pulmonary complication*, Anastomotic leak*, Vocal cord paralysis*, 3-yr survival*</td>
<td>3</td>
<td>1.20 (0.84–1.42)</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>Hand-sewn vs. stapled esophagogastric anastomosis after esophagectomy</td>
<td>Mortality, Anastomotic leak, Pulmonary complication, Cardiac complication*, 3-yr survival*</td>
<td>3</td>
<td>1.20 (0.34–4.25)</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>Posterior vs. anterior route of reconstruction after esophagectomy</td>
<td>Mortality, Anastomotic leak, Pulmonary complication, Cardiac complication*, 3-yr survival*</td>
<td>3</td>
<td>1.20 (0.84–1.42)</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>Pyloroplasty vs. no drainage in gastric reconstruction after esophagectomy</td>
<td>Pulmonary complication*</td>
<td>2</td>
<td>0.69 (0.42–1.44)</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>Primary repair vs. fecal diversion for penetrating colon injuries</td>
<td>Mortality, Morbidity, Intraabdominal infection, Wound infection</td>
<td>5</td>
<td>1.70 (0.51–5.70)</td>
<td>25</td>
</tr>
<tr>
<td>12</td>
<td>Stapled vs. hand-sewn methods for colorectal anastomosis surgery</td>
<td>Pulmonary complication*</td>
<td>2</td>
<td>0.69 (0.42–1.44)</td>
<td>2</td>
</tr>
</tbody>
</table>

(Continued)
concordance of different studies may vary depending on whether primary or secondary outcomes are examined. Discrepancies may be less apparent for secondary outcomes than for primary outcomes because secondary events are likely to be too uncommon to show any significant difference between arms except in extremely large trials (mega-trials). 17

One possible explanation for the greater frequency of between-study heterogeneity in observational studies than in randomized trials is that each observational study usually includes a wide spectrum of subjects from the population at risk. In contrast, randomized trials use specific inclusion criteria and may not be representative of populations seen in clinical practice.

All topics examined in this study were comparisons in the form of A versus B. Generally, A represented a new procedure while B represented an accepted method, but deciding which was newer was difficult in some topics. Most trials in medicine estimate the benefits of pharmacologic effects, whereas 50 of 52 outcomes in this study estimate risks of operations, such as mortality and morbidity. Discrepancies in summary estimates were estimated accordingly between randomized trials and observational studies. For example, the greatest statistical discrepancy between the two types of study design was topic 14 (mortality), comparing extended and limited lymph node dissections for adenocarcinoma of the stomach (Z = 4.34).

In this topic, although the summary estimate from observational studies was one fourth that from randomized trials (0.63 vs. 2.39), this represented an underestimation of risks, not of benefits.

The authors revealed that one fourth of observational studies gave different results to randomized trials and between-study heterogeneity was more common in observational studies in the field of digestive surgery. Furthermore, even if clinical applicability is improved by combining a large number of observational studies, estimations of treatment effect sometimes differ from those obtained from randomized controlled trials. The present study confirmed such tendencies in the well-defined area of digestive surgery. However, observational studies offer several advantages over randomized controlled trials, including lower cost, greater timeliness, and a broader range of patients. 34 These benefits remain worthy of attention in real clinical settings, particularly where random allocation is not easily accepted by either clinicians or patients. In the field of digestive surgery, large
observational studies may actually be more reliable than small underpowered randomized controlled trials. To clarify how to interpret the findings of observational studies and randomized controlled trials, further analyses in other fields are eagerly awaited.

REFERENCES