Scientific Research

Bias-Variation Dilemma Challenges Clinical Trials: Inherent Limitations of Randomized Controlled Trials and Meta-Analyses Comparing Hernia Therapies

U. Klinge[1*], Andreas Koch[2], D. Weyhe[3], Enrico Nicolo[4], R. Bendavid[5], Anette Fiebeler[6]

[1]Department of General, Visceral and Transplant Surgery, University Hospital of the RWTH, Aachen, Germany
[2]Cottbus, Germany
[3]Department of General and Visceral Surgery, Pius-Hospital Oldenburg, Oldenburg, Germany
[4]Department of Surgery, University of Pittsburgh Medical Center, McKeesport, USA
[5]Shouldice Hospital, Thornhill, Ontario, Canada
[6]Berlin, Germany
Email: *Uklinge@ukaachen.de

Abstract

Purpose: Evaluation of hernia therapies according to the current rules of Evidence Based Medicine is widely reduced to results of RCTs or meta-analyses. RCTs have been accepted as a most important tool to confirm a superior effect of an intervention. Unfortunately, in hernia surgery, comparisons of RCTs and correspondingly their use in meta-analyses, are not, surprisingly often, able to confirm any significant impact of a specific procedure due to intrinsic restrictions in a multi-causal setting with its web of influences. Methods: Based on our own experiences of clinical studies in surgery, the present article outlines several situations, with their respective reasons, which argue the severe limitations of RCTs and meta-analysis to define an optimum treatment. Results: Meta-analyses accumulate the variations of each trial, which then may cover any clear causal relationship. RCTs usually are dealing with subgroups of standard patients thus excluding the majority of our patients. Low statistical power of current cohort sizes restricts the analysis of subgroups or of effects with low incidences. Simple comparisons of means frequently are hampered by nonlinear relationships to outcome. The relevance of a specific variable is difficult to separate from other influences. The limited surveillance period of studies ignores a delayed change in outcome. Randomization cannot guarantee a standardized patient's condition. All the arguments have to be con-

---

*Corresponding author.

sidered as a crucial and fundamental consequence of the bias-variance dilemma or principle of uncertainty in medicine, and underline the many limitations of RCTs to evaluate any specific impact of hernia therapies on e.g. infection, pain or recurrence. Conclusions: Many surgical issues cannot be and should not be investigated by RCTs, in particular, if a marked patients' heterogeneity has to be considered or the low incidences of the outcome readout cannot be addressed with sufficient statistical power without getting lost in the variation mire. Registries with their non-restricted data-acquisition should be regarded as reliable alternatives for postoperative outcome quality surveillance studies.

Keywords

Randomized Controlled Trial, Hernia Surgery, Registry, Meta-Analysis, Clinical Study

1. Introduction

Since their introduction in the 1940s Randomized Control Trials (RCTs) have been considered to be the highest level of evidence in clinical investigations (level 1b Oxford criteria 2013, meta-analysis or systematic review of RCTs is level 1a) [1] [2]. Comparing the outcome between standardized groups, many RCTs test the impact of standard interventions attempting to prove the superiority of a procedure. Randomization is supposed to provide widely homogenous cohorts of standard patients with a standard disease. And heterogeneities are assumed to be controlled by documentation of various variables trying to consider all possible influences, which are suspected to interfere with patients' outcome. However, just assuming the outcome used as readout is a rare event with an incidence of less than 10%, then any study with sufficient power requires more than 1000 comparable patients to be included. Challenging the messages of single RCTs, the meta-analyses with its summaries of several RCTs into one report are regarded as the "holy grail of evidence based medicine" [3]. Despite several limitations [4] [5] both RCTs and meta-analysis of RCTs correspondingly are considered to represent the best of our knowledge with the highest degree of reliability [6]. High external validity of RCT and subsequently of its meta-analysis would indicate that the results are more transferable to other patients than the study population, and thus it is affected by the details of the studies for characterization of randomized patients, outcome measures and follow-up procedures [7]. For control of the patients' individual condition we usually define a list of prognostically relevant variables. When performing a clinical trial we try to get access to all the major details of our patients by using several variables, convinced by the suspicion that the more variables we have, the better the control of our cohort might be.

Unfortunately some important limitations give rise to the doubt that we cannot always rely on the mean results of a meta-analysis or a RCT for the best treatment of our specific patients. Uncritical application of RCT results to non-selected patients can even be dangerous and lead to a substantial increase in mortality [8]. Results of underpowered RCT are often mistakenly used as reference in the use of modified therapies [9]. In the field of hernia surgery we are facing similar experiences with lots of underpowered studies or several meta-analyses ending up in comparisons with non-significant differences. It may be time to have a closer look at the limitations of RCT and its meta-analysis with focus on studies done for hernia.

2. Limitation 1: Meta-Analysis Accumulates Variations of Single Trials and Thus Tends to End Up in Comparisons with Non-Significant Differences

e.g. Comparing the Lichtenstein procedure to endoscopic ones in hernia repairs, for example, the mean rate of recurrences in 25 studies was found to be 1.8% for Lichtenstein and 2.8% for TAPP/TEP respectively [10]. The significantly higher recurrence rate after TAPP/TEP disappeared if a single study, which contributed 76% of all recurrences, was excluded. The variance was 30.3 for the endoscopic procedures and 8.4 for Lichtenstein, with a range of 0% to 11% and 0% to 25%, respectively. Obviously the accumulation of several heterogeneous studies did not reduce the variance, as it is expected when repeating experiments in physics. The missing effect, which has been "confirmed" by a meta-analysis, is often the expression that the variation among the trials and studies exceeds any possible effect in some subgroups, and thus the results of a meta-analysis often do not provide a re-

liable basis to treat our patients.

Any compilation of the varying results of RCTs within a meta-analysis generally bears the risk to lose statistical power and to favour "non-significant", "similar" or "comparable" results. A search for PubMed articles on Feb 15th 2013 with the terms "meta-analysis + hernia" listed 51 articles published in 2012, 35 of which dealt with abdominal wall hernias. 71% (25 out of 35) concluded that they could not confirm significant differences, at least for some outcome parameters. In six articles the conclusion called for bigger trials. In 7 articles, only, meta-analysis confirmed significant differences, *i.e.* lower Surgical Site Infection (SSI) with laparoscopic incisional Hernia Repair (IHR) [11] [12]. However, these statements sometimes conflicted with other meta-analyses, e.g. as for a positive effect of antibiotic prophylaxis to prevent SSI [13] [14], no effect [15], or possibly a slight effect [16]. One reason for these controversial results is the fact that most trials are done without assessment of the different patients' risk for the development of an SSI. In fact any prophylactic effect of antibiotics will be influenced by the patients' specific risk for developing a SSI, being more effective in patients at risk; however in patients at low risk for SSI the low incidences will make it barely possible to demonstrate any improvement.

3. Limitation 2: An RCT Includes a Subgroup of Standard Patients with Standard Treatment and Standard Outcome, Which Thus Excludes the Majority of Our Patients

The RCT as basic component of most meta-analyses usually starts with a detailed list of inclusion criteria to define the study group. There is no doubt that patients under extreme conditions should not be included in a trial. However, to adopt the result of a RCT for our own patients, the remaining trial subjects should be widely comparable with most patients out of study centers. But can we be certain that an RCT covers at least the majority of our patients?

Interestingly in 1869, 143 years ago, Theodor Billroth published in the Langenbecks Archives of Surgery that in a cohort of 1000 patients only 25 - 100 may show the same course [17]. In 2005 Rothwell described examples when the rate could even drop below 1% leading to serious limitations of external validity [7]. This effect may explain the discrepancy that, permanently, 10% - 15% of our operations for groin hernia repair are done for recurrences though in many, many trials the rate of recurrences is far below 1%. It is the highly artificial condition of a RCT with highly selected patients that can explain why results differ so greatly in between RCT and epidemiological data bases [7] [18].

4. Limitation 3: Combination of Several Inclusion Variables for Control of Patients Leads to a Minority of Normal Patients, Even If Each Was Considered as "Normal" in 95%

When initiating a clinical trial we would like to exclude patients that are somehow not normal. Often a value is accepted as "normal" when 95% of the patients show this characteristic. However, it means that the number of patients for the study will decrease by at least 5% with every single variable added. The use of 14 variables as control parameters already will reduce a cohort of 1000 patients to 488 patients, and the use of >50 variables will reduce the number of "standard patients" to less than 10%—or the number Billroth estimated 140 years ago, which may be the rational basis for his assumption (**Figure 1**).

Thus, on one hand more variables will enhance the safety that the patient is "normal", but on the other hand it will reduce the number of patients for ready enrolment, and thus reduce the applicability to the majority of our patients and lessen the relevance of the study results for our general patient population.

5. Limitation 4: Number of Parameters in the Study Protocol Will Separate the Study Cohort into Subgroups with Almost Individual Destiny

Multiple variables for risk assessment separate any cohort into subgroups of individual patients. If the patient passed the inclusion criteria then several further variables are used to get access to any further possible risk factor. Almost every study includes records on age, BMI or gender, but usually many more variables are listed up, because they are suspected to have a possible impact on outcome. For analysis most of them get coded as binary digit, e.g. "absent" or "present", "normal" or "pathological", "male" or "female". Others are split up in 3 to 5 levels, e.g. tumor size or status of lymph node in oncology. Even numerical values such as age or BMI often are recoded if a risky condition is suspected, e.g. >75 years or <85 years, normal weight or adipose with BMI > 35.
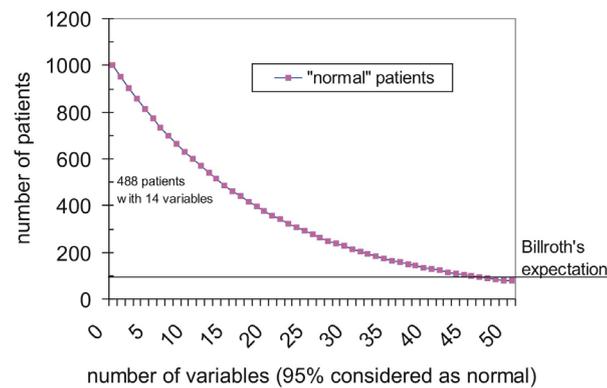
Figure 1. Number of patients with "normal" values in relation to the number of variables used.

The individual patient then can be characterized by the pattern of these variables. However, the exponential increase of the total number of all possible options by adding variables is often underestimated. Whereas 2 binary variables provide only 4 different options (aa, ab, ba, bb), for *n* binary variables there result 2^n options. Correspondingly 10 variables with a binary coding will already provide more than 1000 different conditions; more than 40 variables offer more options than the world has inhabitants. For variables coded with more than 2 levels the number is even dramatically higher. All the different patterns of 17 variables each coded with 4 levels will already exceed the current world population, thereby, in principle, providing an individual pattern for every man. Correspondingly, already some few variables will divide the cohort in so many different patterns that it finally ends up in almost sole individual conditions. In fact, a RCT with its focus on the mean result in a study cohort has to ignore the differences given by the other variables. As it cannot provide an analysis of different patterns, it does not provide a reliable basis to treat subgroups of patients at risk.

6. Limitation 5: Mode of Interference Often Is Not Defined for Causal Variables Showing a Nonlinear Relationship to Outcome

For most of our studies we assume a linear relationship between effect and outcome, e.g. we expect that risk for recurrence roughly correlates with the size of the defect, or with increasing BMI. Unfortunately, often there is no reason for this assumption.

In a study investigating the manifestation of incisional hernia after laparotomy we found an increased risk for patients with a body mass index (BMI) of 25 - 35, but a lower risk for thinner as well as for thicker patients (**Figure 2**) [19]. The relation between risk for incisional hernia and the patients' age shows a similar U-shaped configuration, with a peak for patients between 50 and 70 years of age. Younger as well as older patients both seem to have a reduced risk, the first maybe because of better healing, the second perhaps because of limited survival time. For whatever reason, the risk of BMI or age is not linearly correlated with incisional hernias. This is in line with observations from Rosemer *et al.* who could demonstrate a similar U-shaped correlation between BMI and postoperative complication, with higher risks for BMI < 20 as well as for BMI > 25. A BMI, which is too high, obviously is as bad as when it is too low [20]. Ignorance of this U-shaped relation will likely end up in non-significant results, when comparing means or testing 2 × 2 tables with Chi [2]. In fact, for most of the variables currently used in trials when starting a study we do not know whether there are one or more peaks in relationship with outcome, and which are the corresponding cut-offs. Thus, the inappropriate design of many RCT with a risk-assessment simply using comparison of means tends to result in no significant differences.

7. Limitation 6: Specific Relevance of Redundant Variables Is Not Defined

Not least in the field of hernia surgery there are several different factors, which are all assumed to have an impact on outcome. And there are many different read-outs that are suspected as indicative for a successful treatment (**Figure 3**). It generally may be argued whether a single trial can find an answer as to how far a single factor can influence a specific readout without considering all the other contributing influences. Looking at factors, which may influence the outcome after surgery there will be no surprise, if e.g. experience is related to better
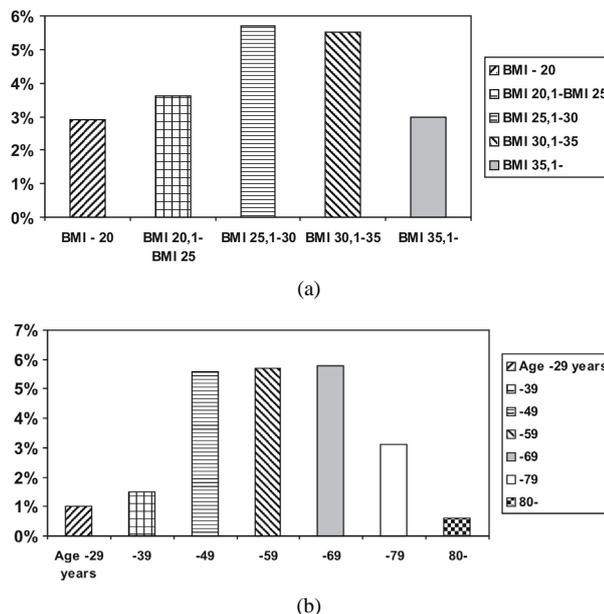
(a)

(b)

Figure 2. U-shaped relationship between risk for incisional hernia and BMI (a) and age (b) (data from Höer J, Lawong G, Klinge U, Schumpelick V. [factors influencing the development of incisional hernia. A retrospective study of 2983 laparotomy patients over a period of 10 years] Chirurg, 2002 May, 73(5), 474-480) [19].
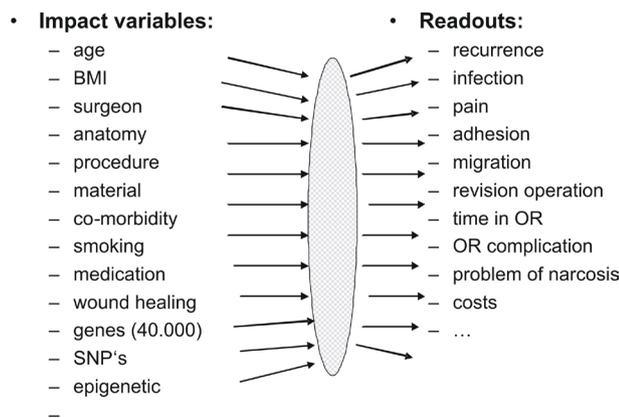

Figure 3. Variables in the field of hernia surgery that are assumed to have an impact on various read outs for outcome, either alone or in combination.

results. But surprisingly, in our data pool of about 3000 laparotomies there was an increased incidence in the patients for an incisional hernia in patients whose laparotomy was closed by the most experienced surgeons, and the lowest risk if treated by residents (**Figure 4(a)**). Should we then try to get treated by residents when we need a laparotomy? In fact the relationship is more complex. When looking at further details we found that the selection of patients for the residents had fewer risks. The experienced surgeons had to deal more often with recurrent incisions, patients with cancer, patients who received blood transfusions or with signs of shock in the OR. All these variables clearly indicate that in this analysis the choice and impact of the surgeon is a consequence of the different patients' selection and their criteria.

What may be the best variable to reflect the patients' risk in developing an incisional hernia? Is it the experience of the surgeon, the challenge of a recurrent incision, the presence of cancer, the need for blood transfusion
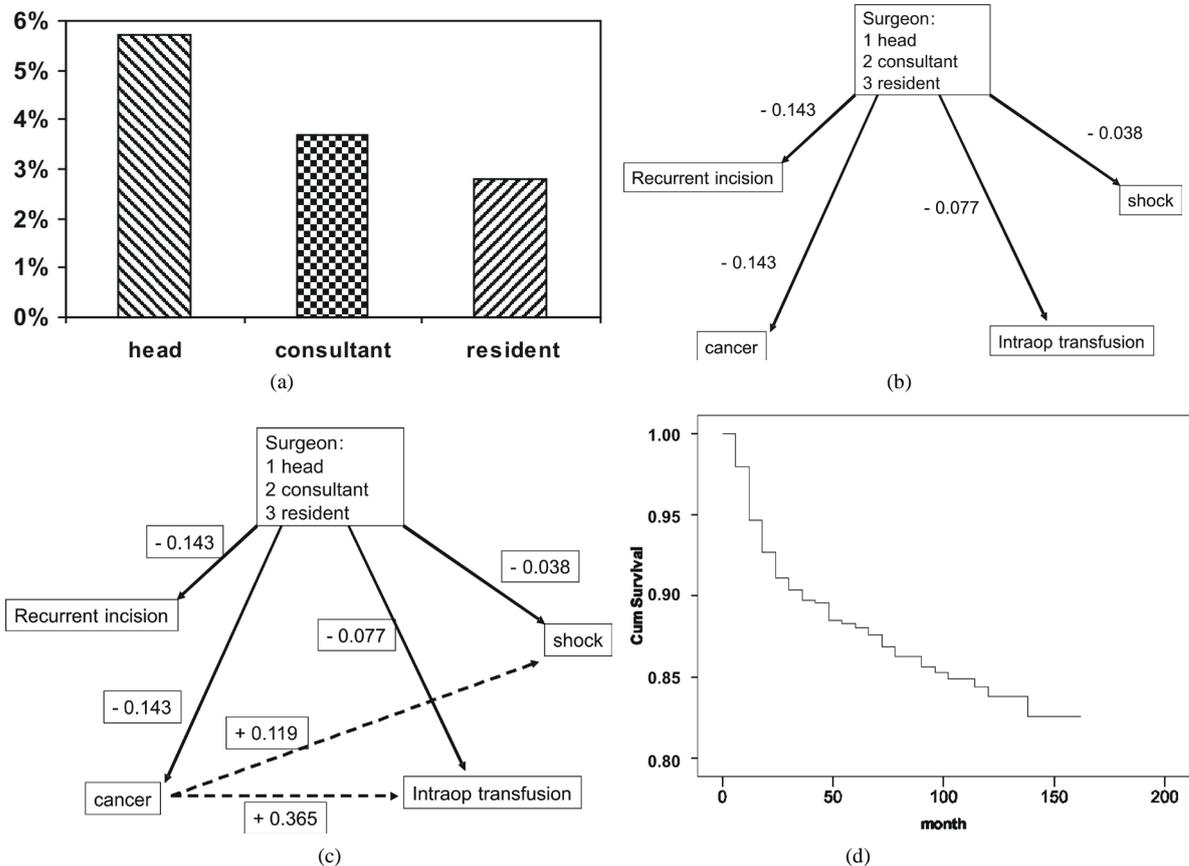
Figure 4. (a) Impact of the surgeons experience on the risk to develop an incisional hernia (data from Höer J, Lawong G, Klinge U, Schumpelick V. [factors influencing the development of incisional hernia. A retrospective study of 2983 laparotomy patients over a period of 10 years] Chirurg, 2002 May, 73(5), 474-480) [19]; (b) Relation between surgeons experience and subsequent variables with impact on the patients risk for developing an incisional hernia (a). Cross-correlation between "cancer" and subsequent variables (b). Negative correlations solid, positive correlations dotted (data from Höer J, Lawong G, Klinge U, Schumpelick V. [factors influencing the development of incisional hernia. A retrospective study of 2983 laparotomy patients over a period of 10 years] Chirurg, 2002 May, 73(5), 474-480) [19]; (c) Kaplan-Meier estimate for development of an incisional hernia after laparotomy (data from Höer J, Lawong G, Klinge U, Schumpelick V. [factors influencing the development of incisional hernia. A retrospective study of 2983 laparotomy patients over a period of 10 years] Chirurg, 2002 May, 73(5), 474-480) [19].

or the happenstance of shock in the OR? As all these variables are cross-linked by positive correlations (**Figure 4(b)**), it leads to a marked over fit and overestimate if all five variables are considered to model the specific risk for these patients. Facing the complexity of all the interferences among variables there is no clear way to identify and select the best variable or to weigh their specific impact when trying to model outcome or the patients' individual risk. Any trial that intends to control all relevant impact factors will be doomed to failure by the complexity of this web of interactions.

8. Limitation 7: Limited Surveillance Period Ignores Any Delayed Manifestation of the Readout Parameter

In the study protocol of a RCT there usually is a precise definition of the time point for follow up investigations. But for example, if you are interested in the development of an incisional hernia shall we then survey our patients after 6 month, 1 year or even later?

In the data pool of our previous study we found a rate of 2% when taking an interval of 6 months after operation for follow-up, of 4% after 1 year and 5% after 2 years [19]. However, considering different follow-up periods according to Kaplan-Meier analysis estimates a rate of 19% after 13 years (**Figure 4(c)**), which confirms

the frequently delayed manifestation. Thus, any RCT with short term surveillance of only 1 to 2 years may reflect real outcome only partially, and is insufficient when targeting long-term complications, e.g. hernia recurrences or mesh related infections, which usually leads to mesh explantation 2 years after implantation [21].

9. Limitation 8: Randomization Cannot Exclude the Influence of Confounders and Thus Cannot Control the Patient's Condition

In many trial settings slight heterogeneities are supposed to be homogenized and controlled by randomization of the patients. Overall, the groups to be compared should include patients with almost similar characteristics, which, however, often is an illusion. In a computer model we formed two groups with hypothetical patients, every case defined by one hypothetical variable, and by using random figures, which were coded with either "1" or "2". It was intended that 20% of the "patients" should show a "1". In the first group of 50 cases (A) 9 cases finally were coded with "1", whereas this was true for 13 in the second group (B). As expected there was a slight difference that was not statistically significant. This procedure was repeated for another 9 variables, and again, as expected, resulted in quite similar distribution without any significant differences (**Figure 5(a)**). Obviously, all variables are similarly distributed between the two groups, which then seem to be comparable—at first glance. In the mean every single variable seem to be similarly distributed, but how many of the fifty patients in group A may have a pattern of variables that fits exactly to patients in group B?

Looking at the individual pattern of each case revealed that only 13 out of 50 cases (26%) have a corresponding partner with an identical pattern for the ten variables in the second group (**Figure 5(b)**). 37 of the 50 cases showed at least one different value in the list of 10 variables. Looking at the different patterns with "1" and "2" for all 10 variables, in principle $2^{10} = 1024$ different patterns were possible, of which in these 100 cases 69 have been realized, meaning 69% showed a unique individual pattern. Conclusively, randomization of 10 variables provides similarity in the mean but cannot prevent an enormous heterogeneity of the individual cases. As we cannot exclude that any of these differences may have a relevant impact on the outcome, and when we consider the thousands of genes and proteins as additional variables, which may influence the patients' outcome as well, there is no doubt that randomization will not be able to control the comparability of the cohorts. Thus the results of a RCT may be considered rather as controlled randomization than as randomized control of the trial population.

10. Limitation 9: Modelling of RCT Do Not Reflect a Multifactorial Pathogenesis

A simple and strict unicausal relationship between intervention and outcome usually is the exception and can hardly be applied to most of our questions. In the field of hernia surgery, which often deals with recurrences chronic pain surgery, surgery with biological and non-biological devices all these factors are expected to have an impact on the patients' outcome (**Figure 6(a)**). Consecutively, testing the impact of surgery has to consider that some failures are independently caused by poor healing of the patient or alternatively by a poor design of a medical device. This will reduce the estimated impact of any surgical intervention markedly. For example, in the presence of an altered collagen metabolism the impact of the surgical procedure may become negligible. Or in the case of poor surgery the relevance of an adequate design of the device may be ignorable, as even the best design will end up in failure.

Most of our outcome parameters are considered as results of a multifactorial pathophysiology and caused by a mixture of influences. Correspondingly when calculating the cohort size for a study the number of patients to test the specific impact of a sole factor needs to be much higher than estimated, as maybe only a small percentage of the outcome if affected by this confounder. As outlined before, with respect to the incidences of the targeted outcome, if you want to test the impact of a procedure and to show an improvement in pain or recurrence rate with a decrease from 5% to 3%, you need more than 1000 patients per group to reach a sufficient statistical power of >0.9. But if a third of these complications develop independently from the intervention to be tested, the number of patients has to be far higher! Similar calculations evaluating the advantage of a laparoscopic resection for colorectal cancer showed that even in a multi-center setting with dozens of participants it would take several decades of unchanged treatment for recruitment of enough patients [22].

11. Limitations as Consequence of a Crucial Problem: Bias-Variance Dilemma or Principle of Uncertainty in Medicine

The criticism on the selection, coding and evaluation of variables may be regarded as not unavoidable and cor-
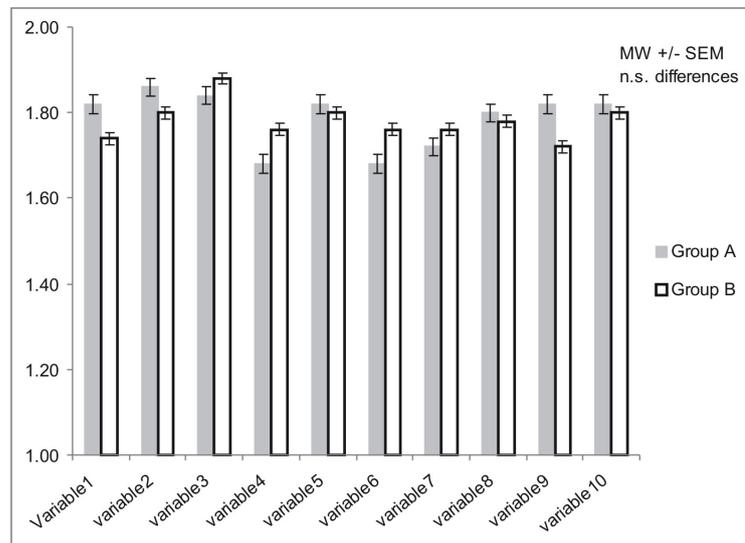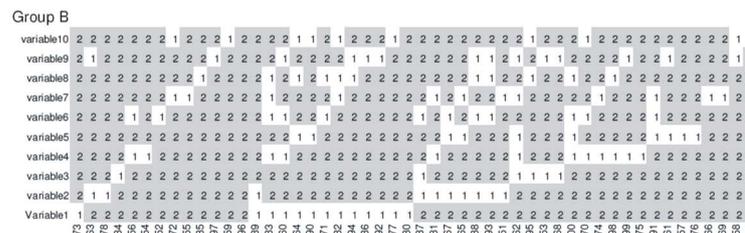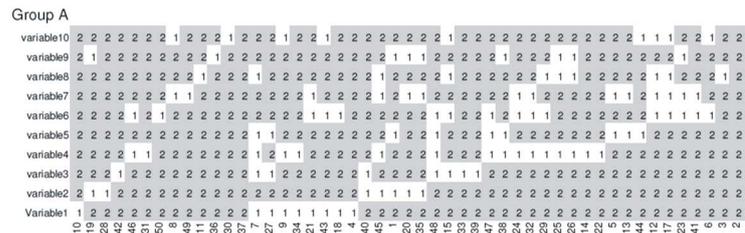
Figure 5. (a) Two groups A and B of hypothetical cases were formed each with 50 cases. Every case was characterised by 10 variables, which each have been coded as either "1" or "2" based on random figures with an estimated 80% share for coding with "2" and correspondingly a mean expression value of 1.8. Given is the mean expression value with the standard deviation of all 50 cases of a group for each variable. There was not any significant difference between the two groups for either variable (Chi^2 > 0.05); (b) Pattern of 50 hypothetical cases (x-axis) with 10 variables (lines of the y-axis) coded either as "1" (yellow) or "2" (blue) based on random figures, two groups A (top) and B (bottom), sorted for similarity (placed left to the black line all paired cases with equal pattern for the 10 variables. 10/73; 19/63; 28/78; 42/84; 46/56; 31/54; 50/52; 8/72; 49/55; 11/85; 36/97; 37/96; 30/59). Only 13 out of 50 cases (26%) of group A have a corresponding partner with an identical pattern for the ten variables in the second group B. 37 of the 50 cases showed at least one different value in the list of 10 variables.

rectable in a better trial setting. Though this may compensate some of the limitations, unfortunately, it is more difficult—it is based on a fundamental problem, which has already been named: "the principle of uncertainty" by Grenander in 1951 [23] or the bias-variance dilemma by Gemann in 1992 [24]. It gives, in fact, principal limitations to what we can predict by analyzing variables and to its use for predictive modeling. Briefly it means that we need many variables for full description of the patients' individual condition with little bias. Because any variable adds a possible variation, any lowering of the bias by an expanding list of variables is linked to an increase

(a)
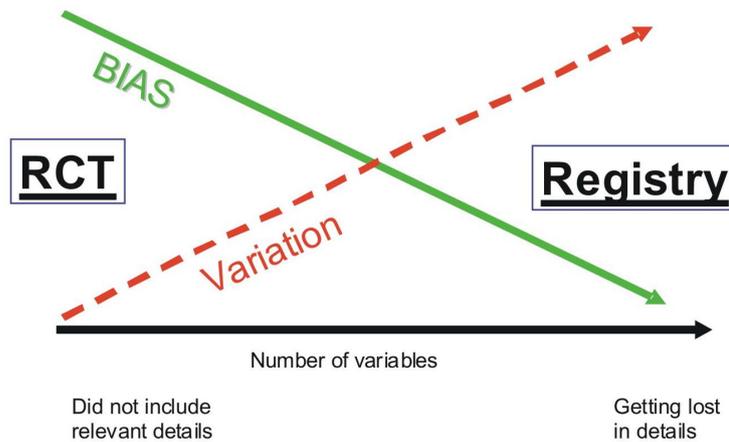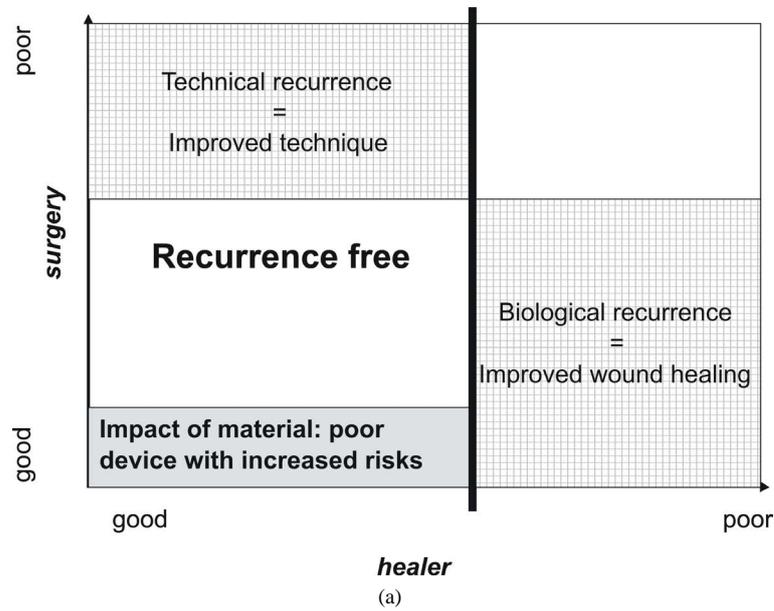
bias–variance dilemma (Geman et al., 1992)
uncertainty principle (Grenander et al 1951)

(b)

Figure 6. (a) Recurrence may be caused by poor surgery, poor healing or by poorly designed devices, whereas the contribution of each often is not clear. However every trial testing an improvement intervention has to consider that some of the recurrences developed independently from the surgical procedure, and thus will reduce the power of the test; (b) Inverse relationship between bias and variation in dependency of the number of variables to define a setting (as outlined in [23] [24]).

of variance, and vice versa, lowered variance means increased bias (**Figure 6(b)**). The consequences are significant. Modeling the patients' outcome with many variables ends up in a best fit of reality, but makes the cohort heterogeneous and is accompanied by poor prediction. Use of only few variables provides a poor fit of reality, but provides a better prediction as the cohort remains rather homogenous. It is essentially unavoidable that a full description of each individual patient requires so many variables that their exponentially increasing variance prevents the formulation of any standard model, and will result in a collection of individual courses. It is not an insignificant reason for which we have to consider a "growing list of null trials" [25].

Most RCT try to control many different impact factors in a multifactorial setting, but often get lost in the bias-variance dilemma, which favors the production of non-significant results.

12. Future Perspective: Postoperative Outcome Surveillance by Use of Registries Offers a Promising Alternative

Without any doubt the introduction of RCT in the late 1940s has been a giant step forward in improving the quality of clinical studies. A RCT like the one from Janes *et al.* clearly confirmed the benefit of a prophylactic mesh for the prevention of a parastomal hernia [26]. However we rarely have the chance to test a new procedure that is so powerful that it reduces the number of complications from 44% to zero. Instead we usually try to realize far smaller improvements. Correspondingly, to reach sufficient significance and power we have to expand the number of patients—and variables. Interestingly the trial from Janes was stopped for ethical reasons because the local committee found that it was no longer justifiable to randomize patients while offering a clearly inferior treatment.

Despite these outstanding examples of successfully performed RCTs the majority of RCTs lack sufficient power and external validity. It may reflect the desperate attempts to save a RCT that Hannink *et al.* found a PostHoc violation of the primary outcome parameter in 32% of RCT published in peer reviewed journals [27]. In 1996 Black identified four main reasons why RCT are limited: experimentation may be unnecessary, inappropriate, impossible, or inadequate [5]. He concluded that "when trials cannot be conducted, well designed observational methods offer an alternative to doing nothing". Observational studies nowadays can best be done with the help of registries, which provide a structure and a set of variables that are known to reflect all major influences on the patients' outcome. In this regard it uses the same variables as the RCT but did not restrict its data acquisition to a small group of study patients [28]. The formulation of the variables is not limited to a comparison for confirmation of similarity, but is open for further sub-grouping. In general, registries include most of the data that are gathered by RCTs, as well. In contrast to RCT/meta-analysis, which tries to test a specific hypothesis or prediction, registries primarily are descriptive. Every surgeon can participate and use the tools of the registry for documentation, instead of having to establish one's own follow up protocol. The list of variables is open to changes, and can be considered in new datasets. PostHoc analysis of subgroups is feasible though these results should be verified prospectively. The value of the dataset improves over time with every follow up and with increasing numbers of participants. However, the main aim of a registry cannot be to prove the superiority of any intervention, as it is the goal of the RCT/meta-analysis, but to disclose any poor performance. Comparing results of subgroups with the outcome of the entire database will help to identify procedures with serious deviation from the mean of the entire dataset. It usually cannot prove the superiority of any procedure, but in fact the RCT or meta-analysis cannot, either.

The need for adding registries to our current tool of RCT/meta-analysis is not least a consequence of the marked improvement in surgery during the past decades. Whereas at the time of Billroth when mortality was 50% and reflected a personal series, any clinical documentation helped to improve understanding and attempts to achieve further improvements, today's results require more sophisticated tools as the RCT represents. Meanwhile, we are struggling to reduce failure rates of 1% - 3% or even less. Correspondingly we should accept that we need to change our research tools again. And it is time to give a higher level of evidence to registries as EuraHS, the new database for incisional hernia from the European hernia society [28].

With their set of variables and open access to everyone, it not only serves as quality control but offers the possibility of developing a personal approach with tailored surgery in a heterogeneous cohort of patients—by identifying failures and sorting out high risk strategies.

13. Conclusions

A RCT can provide an answer as to whether one specific aspect influences the outcome in a highly selected group of patients, but for the majority of our patients and for most of our current questions these answers may not be reliable. Considering the bias-variance dilemma, any attempt to "control" the study conditions requires an extreme selection of patients in lowering any external validity. Unfortunately, we usually do not know for certain whether our next patient fits in the study conditions of a RCT, thus making it difficult to apply any of the results to the patients' treatment.

In our daily practice we want to reduce even rare complications. Considering the enormous heterogeneity of our patients, it would be helpful to have comprehensive scores to define preoperatively the individual risks. By definition of subgroups e.g. with high risk for recurrence, high risk for infection or high risk for chronic pain, it may be possible to select the surgical procedure, specifically that tailored to our individual patients providing the

best risk-benefit balance. This can be achieved by the use of registries but unlikely by the use of RCT.

There is no doubt that the successful treatment of patients, proven at follow-up investigations, is the highest evidence possible. Again it is Billroth who must be considered as the founder of outcome-based surgical research, and who at the end of the 19th century claimed that every patient's records should be kept scrupulously [17]. He pointed out that it was not sufficient to simply rely on personal impressions but that it is essential to add numerical expressions! He thus presented a vision, which converted empirical medicine to quality management and evidence-based medicine [29].

The many present day registries can help provide a platform for documentation and allow every surgeon to participate and perform his own postoperative outcome quality surveillance study. Give the patients the benefit of what they expected from you to have done already, the serious study of comprehensive statistics as it will apply to them individually, not as a part of a heterogeneous cohort.

References

[1] Gabriel, S.E. and Normand, S.L. (2012) Getting the Methods Right—The Foundation of Patient-Centered Outcomes Research. *New England Journal of Medicine*, **367**, 787-790. http://dx.doi.org/10.1056/NEJMp1207437

[2] Phillips, B., Ball, C., Sackett, D., *et al.* (2009) Oxford Centre for Evidence-Based Medicine—Levels of Evidence (March).

[3] Miserez, M. and Fitzgibbons Jr., R.J. (2012) Schumpelick V: Meta-Analyses of Lightweight versus Conventional (Heavy-Weight) Mesh in Inguinal Hernia Surgery. *Hernia*, **16**, 503. http://dx.doi.org/10.1007/s10029-012-0962-x

[4] Sanson-Fisher, R.W., Bonevski, B., Green, L.W., *et al.* (2007) Limitations of the Randomized Controlled Trial in Evaluating Population-Based Health Interventions. *American Journal of Preventive Medicine*, **33**, 155-161. http://dx.doi.org/10.1016/j.amepre.2007.04.007

[5] Black, N. (1996) Why We Need Observational Studies to Evaluate the Effectiveness of Health Care. *BMJ*, **312**, 1215-1218. http://dx.doi.org/10.1136/bmj.312.7040.1215

[6] Memon, M.A., Khan, S. and Osland, E. (2012) Meta-Analyses of Lightweight versus Conventional (Heavy Weight) Mesh in Inguinal Hernia Surgery. *Hernia*, **16**, 497-502. http://dx.doi.org/10.1007/s10029-012-0987-1

[7] Rothwell, P.M. (2005) External Validity of Randomised Controlled Trials: "To Whom Do the Results of This Trial Apply?". *Lancet*, **365**, 82-93. http://dx.doi.org/10.1016/S0140-6736(04)17670-8

[8] Juurlink, D.N., Mamdani, M.M., Lee, D.S., *et al.* (2004) Rates of Hyperkalemia after Publication of the Randomized Aldactone Evaluation Study. *New England Journal of Medicine*, **351**, 543-551. http://dx.doi.org/10.1056/NEJMoa040135

[9] Brody, B.A., Ashton, C.M., Liu, D., *et al.* (2013) Are Surgical Trials with Negative Results Being Interpreted Correctly? *Journal of the American College of Surgeons*, **216**, 158-166. http://dx.doi.org/10.1016/j.jamcollsurg.2012.09.015

[10] Schmedt, C.G., Sauerland, S. and Bittner, R. (2005) Comparison of Endoscopic Procedures vs Lichtenstein and Other Open Mesh Techniques for Inguinal Hernia Repair: A Meta-Analysis of Randomized Controlled Trials. *Surgical Endoscopy*, **19**, 188-199. http://dx.doi.org/10.1007/s00464-004-9126-0

[11] Forbes, S.S., Eskicioglu, C., McLeod, R.S., *et al.* (2009) Meta-Analysis of Randomized Controlled Trials Comparing Open and Laparoscopic Ventral and Incisional Hernia Repair with Mesh. *British Journal of Surgery*, **96**, 851-858. http://dx.doi.org/10.1002/bjs.6668

[12] Goodney, P.P., Birkmeyer, C.M. and Birkmeyer, J.D. (2002) Short-Term Outcomes of Laparoscopic and Open Ventral Hernia Repair: A Meta-Analysis. *Archives of Surgery*, **137**, 1161-1165. http://dx.doi.org/10.1001/archsurg.137.10.1161

[13] Yin, Y., Song, T., Liao, B., *et al.* (2012) Antibiotic Prophylaxis in Patients Undergoing Open Mesh Repair of Inguinal Hernia: A Meta-Analysis. *American Surgeon*, **78**, 359-365.

[14] Sanabria, A., Dominguez, L.C., Valdivieso, E., *et al.* (2007) Prophylactic Antibiotics for Mesh Inguinal Hernioplasty: A Meta-Analysis. *Annals of Surgery*, **245**, 392-396. http://dx.doi.org/10.1097/01.sla.0000250412.08210.8e

[15] Sanchez-Manuel, F.J., Lozano-Garcia, J. and Seco-Gil, J.L. (2012) Antibiotic Prophylaxis for Hernia Repair. *Cochrane Database of Systematic Reviews*, **2**, CD003769.

[16] Shankar, V.G., Srinivasan, K., Sistla, S.C. and Jagdish, S. (2010) Prophylactic Antibiotics in Open Mesh Repair of Inguinal Hernia—A Randomized Controlled Trial. *International Journal of Surgery*, **8**, 444-447. http://dx.doi.org/10.1016/j.ijsu.2010.05.011

[17] Billroth, T. (1869) Chirurgische Erfahrungen, Zürich 1860-1867. *Langenbecks Archiv fü Chirurgie*, **10**, 1-113

[18] Rothwell, P.M. (2006) Factors That Can Affect the External Validity of Randomised Controlled Trials. *PLoS Clinical*

Trials, **1**, e9. http://dx.doi.org/10.1371/journal.pctr.0010009

[19] Hoer, J., Lawong, G., Klinge, U. and Schumpelick, V. (2002) Factors Influencing the Development of Incisional Hernia. A Retrospective Study of 2983 Laparotomy Patients over a Period of 10 Years. *Der Chirurg*, **73**, 474-480. http://dx.doi.org/10.1007/s00104-002-0425-5

[20] Rosemar, A., Angerås, U., Rosengren, A. and Nordin, P. (2010) Effect of Body Mass Index on Groin Hernia Surgery. *Annals of Surgery*, **252**, 397-401. http://dx.doi.org/10.1097/SLA.0b013e3181e985a1

[21] Klosterhalfen, B. and Klinge, U. (2013) Retrieval Study at 623 Human Mesh Explants Made of Polypropylene—Impact of Mesh Class and Indication for Mesh Removal on Tissue Reaction. *Journal of Biomedical Materials Research Part B*: *Applied Biomaterials*, **101**, 1393-1399. http://dx.doi.org/10.1002/jbm.b.32958

[22] Gastinger, I., Koch, A., Marusch, F., Schmidt, U., Köckerling, F. and Lippert, H. (2002) Significance of Prospective Multicenter Observational Studies for Gaining Knowledge in Surgery. *Der Chirurg*, **73**, 161-166. http://dx.doi.org/10.1007/s00104-001-0383-3

[23] Grenander, U. (1951) On Empirical Spectral Analysis of Stochastic Processes. *Arkiv för Matematik*, **1**, 503-531. http://dx.doi.org/10.1007/BF02591360

[24] Geman, S., Bienenstock, E. and Doursat, R. (1992) Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, **4**, 1-58. http://dx.doi.org/10.1162/neco.1992.4.1.1

[25] Agarwal, R. (2013) What Can We Learn from Null Randomized Controlled Trials? *Journal of the American Society of Nephrology*, **24**, 691-693. http://dx.doi.org/10.1681/ASN.2013030295

[26] Janes, A., Cengiz, Y. and Israelsson, L.A. (2004) Preventing Parastomal Hernia with a Prosthetic Mesh. *Archives of Surgery*, **139**, 1356-1358. http://dx.doi.org/10.1001/archsurg.139.12.1356

[27] Hannink, G., Gooszen, H.G. and Rovers, M.M. (2013) Comparison of Registered and Published Primary Outcomes in Randomized Clinical Trials of Surgical Interventions. *Annals of Surgery*, **257**, 818-823. http://dx.doi.org/10.1097/SLA.0b013e3182864fa3

[28] Muysoms, F., Campanelli, G., Champault, G.G., DeBeaux, A.C., Dietz, U.A., Jeekel, J., *et al*. (2012) EuraHS: The Development of an International Online Platform for Registration and Outcome Measurement of Ventral Abdominal Wall Hernia Repair. *Hernia*, **16**, 239-250. http://dx.doi.org/10.1007/s10029-012-0912-7

[29] Ayer, G. (2010) Die Zeit Billroths in Zürich 1860-1867. In: Peiper, H.J. and Hartel, W., Eds., *Das Theodor-Billroth-Geburtshaus in Bergen auf Rügen*, Wallstein-Verlag, Göttingen, 151.

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.