In the previous chapter I have analyzed the theoretical assumptions of the hierarchical model of evidence and proposed an alternative circular model. Now I want to discuss the consequences resulting from the two models in even greater detail. I consider the hierarchical model unfit. I make no secret of it. In the next chapter I will give a few examples showing that the present model functions poorly and is in the long term too expensive and scientifically unsatisfactory.

The practical consequence of the hierarchical model

The advantage of the classic strategy: The Experiment

According to the requirements of the hierarchical model, in order to find the "true" effect of an intervention we need to experiment as soon as possible in the research process. I use the quotation marks for "true" because I think that in this context this "truth" is only a fiction. (This does not mean that there is no truth as was shown in a simple argument by St. Augustine that truth must exists as a guiding principle: Even he who says there is no truth, claimed truth for his statement. Therefore, truth must exist as a limit/marginal idea.) However, it is a fiction to believe in a single truth in the medical context, a single truth valid for all people in all contexts and under all circumstances in all cultures and at all times, which is equally effective when used by any therapist. This, however, seems to be the standard opinion, or at least suggestion, when one reads statements like "xyz therapy improves the relapse rate in the chronically depressed by 38%." The hierarchical model, as described briefly earlier, uses experimental results whenever possible to produce such statements, because they allow clear conclusions.

Why? Imagine you had two therapies for the treatment of depression: "Muckel-therapy" and psychotropic drugs. Imagine you also have two large groups of patients, those who opt for Muckel-therapy, and those who prefer to take psychotropic drugs. After a certain period of observation you may notice that the patients in the Muckel-therapy group are faring better than those in the psychotropic group. Can this difference be attributed to the therapy? Not necessarily. Because it is possible, for example, that all or many of the patients opting for Muckel-therapy have a certain unknown genetic predisposition resulting in a better metabolism of omega-3 fatty acids, and that a deficiency of such fatty acids is a contributing factor in depression. Thus implicitly our Muckel-group contains people who have a better ability to recover from depression on their own, and we incorrectly attributed their spontaneous improvement to the therapy. Another aspect could be that our Muckel-group patients are a little more educated. It is known that more educated people can more quickly

utilize their own resources for recovery. Then the observation that the Muckel-group is doing better is actually based on the social differences, but because we didn't control for that, we falsely assumed that the differences between the groups were due to the therapy.

There are a myriad of possible factors influencing diseases. Some we know, such as genetic factors, metabolism, education, social status, smoking or alcohol consumption. Many factors we don't know. It might be that someday it turns out that being born during the winter months is a risk factor associated with a particular genetic make-up for a particular kind of disease, for example.

Randomization

Scientists like to use a little trick to control for such well-known and unknown factors: they randomly assign the patients to the conditions using a computer program. Thus, all possible factors are distributed between both groups equally. Then when the intervention is introduced to one group only and careful measurements are taken, it is assumed with some certainty that the differences between the groups are based on the effect of the intervention and not on the implicit prior existing differences between the groups. This theory is valid when the studies are large enough (starting with about 300 patients), and when the randomization is not restricted. The latter is rarely done, because if one just rolls the dice to decide, it may happen that the resulting randomized groups are of unequal size, which one tries to avoid because statistically the smallest group always determines how powerful the test is. So if there is a difference of 50 people between two groups, (for example 150 people in one group and 200 in the other group), then the extra 50 people were recruited in vain. Since the inclusion of patients in trials is expensive, one tries to avoid such differences in group sizes by randomizing in blocks of people. This means that the randomization is done in blocks of 4 or 8 or 10, etc., so that the resulting groups can maximally differ by this many patients. But then the effectiveness of random allocation decreases. This is why randomization only works well with a minimum of approximately150 patients per group. Alternatives exist, such as the so-called minimization strategy in which computer programs allocate participants by calculating the differences between patient groups, but these have unfortunately not persisted in research because they are little more complicated.

Randomization thus results, at least theoretically and, in large studies, practically, in an equal distribution of baseline values between the two groups. But is randomization sufficient to ensure equal baselines? In most cases, no.

Homogenization

Most researchers utilize a number of different ways to hedge their studies. They try above

all to create homogeneous groups. Why? Because then they can show an effect of the intervention with smaller numbers of patients. Considering that including more patients in a trial is expensive (it is estimated that each patient in a longer study costs up to $28,000; the money is spent on the doctor who gets a bonus, on the scientific and medical personnel who collect, evaluate and monitor the data, etc.), investigators try to use only as many patients as needed to show a significant effect. That's ethically necessary, because after all each experiment is always associated with its own burdens, possible drawbacks or side effects, and ethics committees try to ensure that no unnecessary experimentation is conducted. But how can the subtle signal of the intervention be separated from the background noise of the control group? Simply, investigators work with homogeneous groups. This is accomplished by formulating criteria under which the therapy is assumed to work best. Exclusion criteria tell which patients were not treated in the study. Standard exclusion criteria are age limit, language skills proficient for communication, and pregnancy and lactation (because of possible side effects to mother and child). Often these criteria also exclude patients with a certain severity of the diagnosis (for example slightly or severely depressed patients) or patients with multiple diagnoses (for example patients with depression and anxiety, addiction or personality disorders). As a result, it is usually easier to generate an effect which is greater than- or equally as good as- the control condition (depending on which control condition is selected) and easier to show the effect you want.

Other formal and content specific conditions of the experiment

Experiments on humans can and should only be carried out for good reasons. One of the main prerequisites for human experimentation is that it is not exactly known what works well, therefore, our knowledge is in the balance ("equipoise"), as is the case when testing new interventions with unknown effect. Consequently, neither the clinician nor the patient should have a real preference for any particular treatment. Experiments on humans can also only be conducted if the patients know exactly what will be done to them and agree to it. Practically this is done by explaining, verbally and in writing, the intervention design, everything that will be done to them, how often they have to come, what forms they have to fill out, the advantages and disadvantages they should expect, what measurements will be taken, what conditions will be tested (e.g. therapy and placebo, or two different treatments) – and what side effects to expect. Furthermore, such studies necessitate the appropriate logistics found typically only in large hospitals, universities or specialized companies. It is estimated that only 1-5% of all patients are recruited in general practitioners' offices for clinical studies, the rest are recruited in hospitals i.e. specialized treatment centers.

Thus only certain types of patients enroll in clinical studies: those who do not have a specific treatment preference, those who trust the clinic/ the study center/the doctor completely, and those with conditions no longer treatable in the established practice who land in the

hospitals.

The disadvantage of the classic strategy – a lack of generalizability

This shows the main disadvantage of this experimental strategy: the results are applicable only to a very small number of patients. It is unclear if the results are applicable to the remaining 95% of the patients. This is the problem of generalizability, or the so-called "external validity". The worst part is that it is unknown how internal validity, i.e. the characteristics of a methodical study, and external validity, i.e. the generalizability to other patients, are linked, and therefore, this shortcoming can't be corrected with mathematical models or by further considerations. We know only one thing: the higher the internal validity, the greater the likelihood that the external validity decreases. With every excluded patient (whether based on any exclusion criteria, on his or her own resistance to be allocated randomly to a treatment, or simply because his or her condition is not treated in a specialized center), the generalizability of the study decreases. This is less of a problem for densely researched areas, such as acute oncology, because the patients are recruited where they are actually treated and it is easier to discern what works. This not the case for rather "vague" illnesses or diseases that often are associated with various other diagnoses. These constitute the vast majority of all diseases.

I want to illustrate my thoughts with an example: the wealth of psychopharmacologic treatments for depression. They are all licensed, meaning that at some point they were shown to be more effective than a sham treatment, in this case placebo, in one or more studies. For almost all of these depression treatments there is also a plethora of studies showing that they were no better than placebo. More precisely, this is so in more than half of the cases, but broadly speaking they do work. The effects are not overwhelming in size, but all together, the placebo and pharmacologic effects in these studies are large enough so that one gets the impression that the drugs work. (The placebo effect will be discussed later). Now, these data were all obtained in specific experiments: with patients who suffered from depression only, who were not too depressed but were "depressed enough", had no alcohol dependence, where the depression occurred not as a consequence of other diseases, where there was no additional anxiety disorder, and so forth.

In practice, however, most depressed people have many other problems. Thus a huge study that examined the effects of depression treatment, as it occurs in real practice, the STAR*D study, was carried out. In this study psychiatrists were allowed to change the prescription of one medication to another when the first didn't work, were allowed to prescribe psychotherapy as a complementary treatment, until finally even completely new and strong medications with a lot of side effects were used, just as in real practice. The results were sobering: less than 50% of patients became permanently (in this case, at least one year) free

of their depression. A critical analysis shows that actually only 38% benefited from this pharmacological therapy. This example shows that results gained from randomized clinical experiments aren't necessarily applicable to practice – precisely because the generalizability of the results is limited by the experiment itself.

So we must always sail between Scylla and Charybdis: on the one hand we want valid results, on the other hand we want also applicable results. Is it possible to combine the two in a really good study? Yes and no. One could conduct so-called "mega-trials" with 100,000 people divided randomly into two conditions and treatment with no exclusion criteria other than the diagnosis. Then one would have maximally generalizable experimental data. The problem is that such studies are extremely expensive and not really feasible in Europe. Therefore, proponents of such studies suggest Russia, China or elsewhere to carry them out. It is unclear, however, that these results would still be applicable to Europe. Moreover, it could be that a valuable treatment method works only in a particular group of patients. Such differentiated effects are overlooked in such giant studies. Therefore, one cannot invent a jack of all trades which will provide both valid and generalizable results. Rather, one must resort to a strategy that generates data in different studies, and then merges them. The circular model suggests precisely that.

Literature:

- Aickin, M. (1983). Some large trial properties of minimum likelihood allocation. *Journal of Statistical Planning and Inference*, 8, 11-20.
- Aickin, M. (2001). Randomization, balance, and the validity and efficiency of design-adaptive allocation methods. *Journal of Statistical Planning and Inference,* 94, 97-119.
- Aickin, M. (2002). Beyond randomization. *Journal of Alternative and Complementary Medicine*, 8, 765-772.
- Fava GA, Tomba, E., & Grandi, S. (2007). The road to recovery from depression – do not drive today with yesterday's map. *Psychotherapy and Psychosomatics*, 76, 260-265.
- Khan, A. Khan, S., & Brown, WA (2002). Are placebo controls necessary to test new antidepressants and anxiolytics? *International Journal of Neuropsychopharmacology*, 5, 193-197.
- Pigott, HE, Leventhal, AM, Alter, GS, & Boren, JJ (2010). Efficacy and Effectiveness of antidepressants: current status of research. *Psychotherapy and Psychosomatics*, 79, 267-279.
- Rush, J.A., Trivedi, M.H., Wisniewski, S.R., Nierenberg, A.A., Stewart, J.W., Warden, D., et al. (2006). Acute and longer-term outcomes in depressed outpatients Requiring one or several treatment steps. A STAR * D report. *American Journal of Psychiatry,*

163, 1905-1917
- Stewart, D.J., Whitney, S.N., & Kurzrock, R. (2010). Equipoise lost: ethics, costs, and the regulation of cancer clinical research. *Journal of Clinical Oncology*, 28, 2925-2935.
- Walach H, Falkenberg, T., Fonnebo, V., Lewith, G., & Jonas, W. (2006). Circular instead of hierarchical – Methodological principles for the evaluation of complex interventions. *BMC Medical Research Methodology,* 6 (29).