Nephrol Dial Transplant (2007) 22: 2785–2794
doi:10.1093/ndt/gfm433

NDT
Nephrology Dialysis Transplantation

Special Feature

See http://www.oxfordjournals.org/our_journals/ndtplus/

Clinical research of kidney diseases II: problems of study design

Pietro Ravani[1,2], Patrick S. Parfrey[1], Elizabeth Dicks[1] and Brendan J. Barrett[1]

[1]Clinical Epidemiology Unit, Faculty of Medicine, Memorial University of Newfoundland, Canada and
[2]Divisione di Nefrologia, Azienda Istituti Ospitalieri di Cremona, Italy

Introduction

The aim of study design in any field of clinical inquiry is to limit bias and maximize reliability [1]. The present article introduces the types of study design currently recommended for assessing prognosis, therapy and diagnostic tests with nephrology examples. The concept of clinical relevance as opposed to statistical significance of study results is also briefly discussed.

Study design

Hierarchy of evidence

Fundamental to evidence-based health care is the concept of 'hierarchy of evidence', deriving from different study designs addressing a given research question (Figure 1). Evidence grading is based on the idea that different designs vary in their susceptibility to bias and, therefore, in their ability to predict the true effectiveness of health care practices. For assessment of interventions, randomized controlled trials (RCTs) or systematic review of good quality, RCTs are at the top of the evidence pyramid, followed by longitudinal cohort, case-control and cross-sectional studies [2,3]. However, the choice of the study design depends on the question at hand, the nature of the exposure and the frequency of the disease.

Intervention questions are ideally addressed with *experiments* (RCTs), since observational data are prone to unpredictable bias and confounding that only the randomization process will control [1]. Appropriately designed RCTs allow also stronger causal inference for disease mechanisms. However, ideal RCTs cannot be implemented in the same way to answer all intervention questions. Some therapies can even not be masked or randomly assigned (e.g. dialysis modalities). In circumstances where the 'intervention' is clearly identified and easily applied, such as the use of a new oral medication to reduce proteinuria, both internal and external validity can be reasonably maximized using standard approaches (limited exclusion criteria, multiple blinding, minimization of missing data and dropouts). In contrast, when the intervention is aimed at achieving a clinical target, such as haemoglobin or blood pressure levels, many treatment adjustment decisions are often left to the discretion of the treatment team during the trial, blinding may be difficult to maintain and patients are often exposed to multiple strategies (e.g. iron supplementation, erythropoietic agents, antihypertensive medications in studies of haemoglobin targets). In such cases, practitioners may be left with uncertainty as to what aspect of the intervention led to the observed trial results. For example, if higher cardiovascular event rates were associated with aiming for higher haemoglobin targets, it might be unclear whether this was due to the dose of erythropoietic agents employed, the amount of iron given or indeed the interaction between these factors and characteristics of the trial subjects. Those at higher baseline cardiovascular risk might be more difficult to get to target and particularly susceptible to the adverse effects associated with higher doses of iron and erythropoietic agents given in an effort to achieve those targets. However, understanding these relationships as a result of a trial, particularly if confirmed in further research helps inform practitioners on how to best individualize the application of therapy.

Prognostic and aetiologic questions are best addressed with *longitudinal cohort studies*, in which exposure is measured first and participants are followed forward in time. At least two (and possibly more) waves of measurements over time are undertaken. Initial assessment of an input–output relationship may derive from *case-control studies*, where the direction of the study is reversed. Participants are identified by the presence or absence of disease and exposure is assessed retrospectively. *Cross-sectional studies* may be appropriate for an initial evaluation of the accuracy of new diagnostic tests as compared to a gold standard.

Correspondence and offprint requests to: Pietro Ravani, MD, Divisione di Nefrologia, Azienda Istituti Ospitalieri di Cremona, Italy, Largo priori 1, Cremona, 26100, Italy.
Email: p.ravani@ospedale.cremona.it

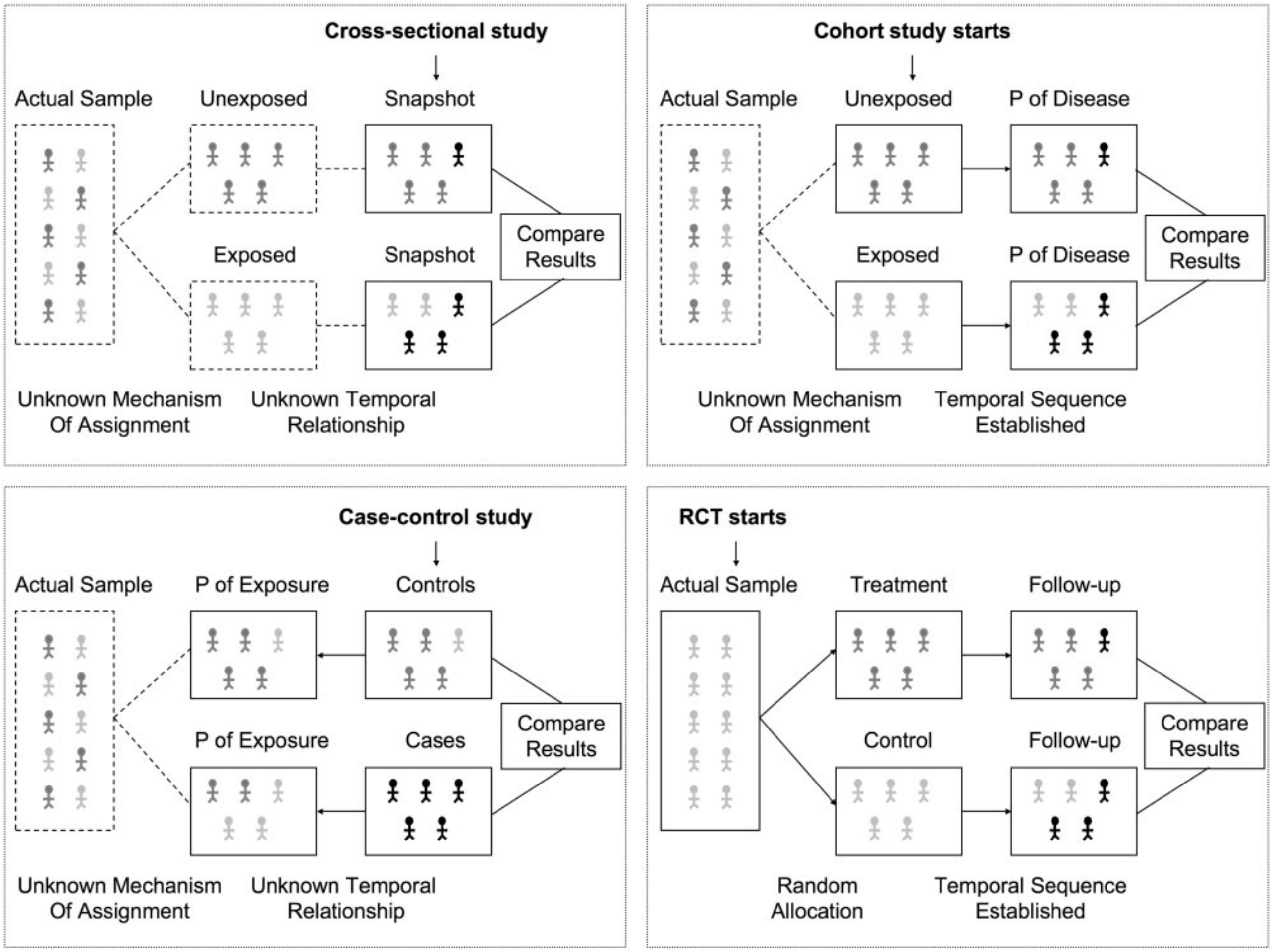

Fig. 1. Examples of study designs. In cross-sectional studies inputs and output are measured simultaneously and their relationship is assessed at a particular point in time. In case-control studies participants are identified based on presence or absence of the disease and the temporal direction of the inquiry is reversed (retrospective). Temporal sequences are better assessed in longitudinal cohort studies where exposure levels are measured first and participants are followed forward in time. The same occurs in randomized controlled trials (RCTs) where the assignment of the exposure is under the control of the researcher. P: Probability (or risk).

Further assessments of diagnostic programmes are performed with longitudinal studies (observational and experimental). Common biases afflicting observational designs are summarized in Table 1.

Additional biases in longitudinal designs

In prognostic studies, as well as in most RCTs, the outcome measure is usually *time to an event of interest* that can be death, a better or worse disease stage, or a complication or recovery from an illness condition. Among the possible threats to internal validity of a study [1], loss to follow-up, drop-outs and attrition bias can induce important errors in the measurements of this outcome variable and, consequently, in the derived *risk estimates* (Figure 2).

The risk of any event is a probability (thus with no dimension and with possible values ranging from 0 to 1), and cannot be directly measured in any single person, since an individual either does or does not develop that event. Rather, the risk is estimated as the proportion of subjects developing the event of interest (D) among a larger group of people (N) who are disease-free at the beginning of the study, and thus at risk over a certain period of time. The resulting incidence proportion (D/N) estimates the individual risk of getting the disease in that period. For example, an observed risk of End-Stage Renal Disease (ESRD) of 0.1 in 10 years in a group of subjects means that each subject of that group has a probability of 10% of developing ESRD in 10 years. It is clear that the definition of the time interval over which the risk applies is fundamental to the interpretation of risk and to proper planning of a prognostic study. In fact, a risk can be thought of as the speed with which the phenomenon can occur in the population. If the risk of ESRD is 0.1 in 10 years in one group and 0.1 in 20 years in another, the speed is twice as high in the first group.

The speed of the disease process has implications for the study design. In fact the faster the evolution of the disease, the shorter the study can be, and the likelihood

Table 1. Bias categories in observational (non-experimental) designs

Design	Sampling bias	Measurement bias
Longitudinal cohort	Patient selection related to their exposure status where exposed and unexposed come from different populations	Wrong outcome classification can be non-differential i.e. unrelated to the exposure level (associations are underestimated); differential misclassification occurs when information on outcome is measured with different accuracy by exposure level (the effect can be over or under-estimated)
Case-control	Patient selection related to their case/control status where cases and controls come from different populations	Wrong exposure classification can be non-differential i.e. unrelated to the disease status (associations are underestimated); differential misclassification occurs when information on exposure is measured with different accuracy by disease status (the effect can be over or underestimated)
Cross-sectional	Preferential patient enrolment based on some characteristics; e.g. volunteerism; non-responsiveness	Both misclassification of exposure and outcome can occur

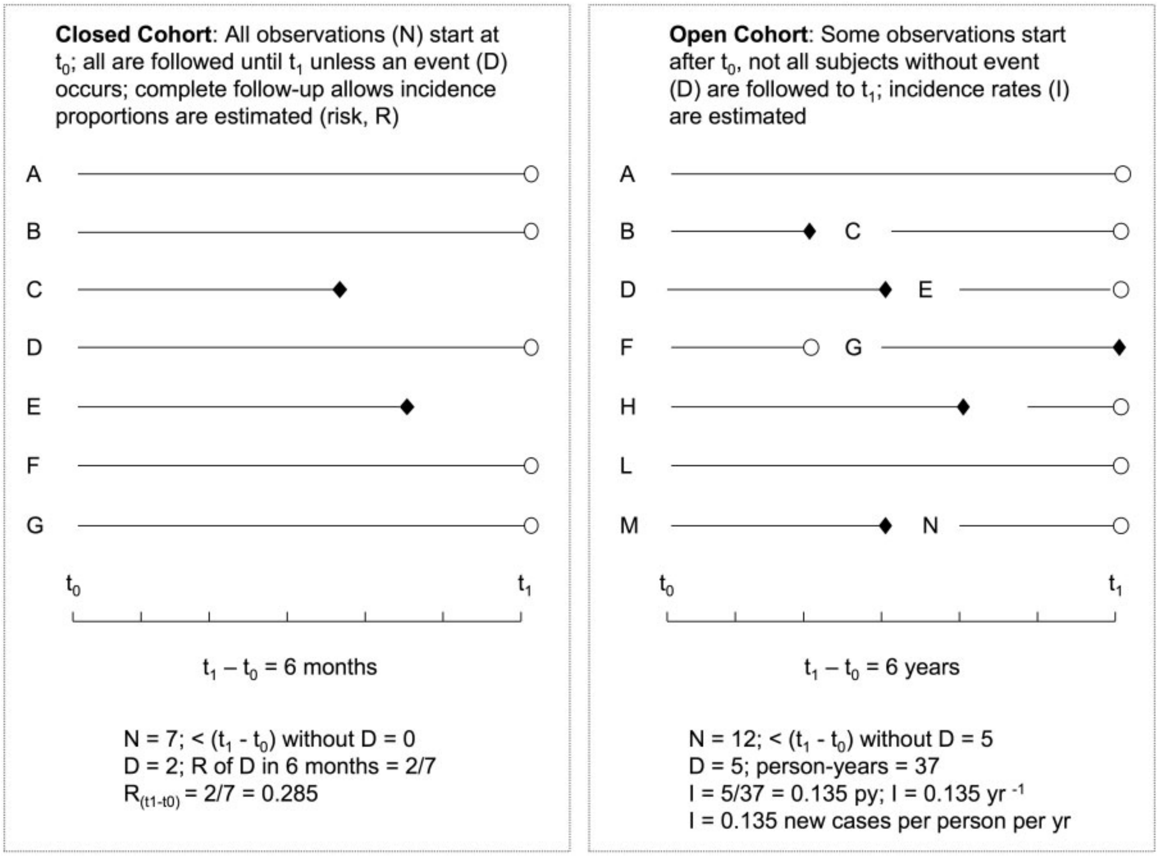
Closed Cohort: All observations (N) start at t_0; all are followed until t_1 unless an event (D) occurs; complete follow-up allows incidence proportions are estimated (risk, R)
A
B
C
D
E
F
G
t_0
t_1
$t_1 - t_0$ = 6 months
N = 7; < (t_1 - t_0) without D = 0
D = 2; R of D in 6 months = 2/7
$R_{(t1-t0)}$ = 2/7 = 0.285
Open Cohort: Some observations start after t_0, not all subjects without event (D) are followed to t_1; incidence rates (I) are estimated
A
B
C
D
E
F
G
H
L
M
N
t_0
t_1
$t_1 - t_0$ = 6 years
N = 12; < (t_1 - t_0) without D = 5
D = 5; person-years = 37
I = 5/37 = 0.135 py; I = 0.135 yr^{-1}
I = 0.135 new cases per person per yr

Fig. 2. Closed *vs* open cohorts and risk measures. In closed cohort studies risk estimates are assessed in a short time interval as the ratio of cases over those at risk at the beginning of the study (incident proportion or risk in a given interval). In open cohorts persons per unit time are the denominator of the ratio estimating the risk. The dimension of the resulting incidence rate is 1/time (unit time^{-1}).

that any individual leaves the study before the end of the observation period without experiencing the event of interest is lower. Studies of acute illnesses such as pyelonephritis or complications such as contrast media nephropathy, are usually of short duration. In these studies, strategies to reduce the risk of losing patients during follow-up are likely to be successful. When the probability of leaving the study earlier without event is low, the outcome measure is a valid estimate of the true risk, because the denominator of the ratio is not substantially affected (Figure 2, left panel). When the study is longer (e.g. time to ESRD or cardiovascular complications), incidence rates are estimated rather than incidence proportions, because more people can be lost to follow-up for several reasons (unknown, competing risks, moving), and new people are often enrolled to maintain the size of the cohort. These incidence rates have as numerator the number of events (D) and person-time as denominator (Figure 2, right panel). Incidence rates have a range of values from 0 to ∞ (depending on the unit of time chosen) and the dimension of 1/time.

The risk can be estimated from the incidence rate using special techniques called survival analyses. However, these techniques do not provide valid estimates of the true risk if the reasons for leaving the study prematurely are related to the exposure (side effects of treatments for example), or the event (earlier manifestations of the final outcome, such as mild symptoms of cardiovascular events). This phenomenon is called 'informative censoring' in survival analysis terminology. Attrition bias may result not only from differential drop-out rates, but also from differential distribution of the reasons for withdrawal. Strategies should be considered for limiting loss to follow-up during the study implementation and careful data reporting once the study is completed [4]. This is problematic in prognostic studies, but may occur also in RCTs. For example, the CHOIR trial compared normalization of haemoglobin with erythropoietin in patients with chronic kidney disease with partial correction of anaemia. Limitations of this study were the extremely high overall drop-out rate and failure to report the reasons for participant withdrawal by exposure level [5]. Biased estimates may also occur if the characteristics of the participants entering the study or the study conditions change over time. For example, a recent study of factors impacting outcomes in atheroembolic renal disease analysed data collected over 20 years [6]. It is possible that milder forms of the disease were more likely to be recognized late in the study as a result of the awareness and experience of the investigators (Will Rogers phenomenon).

Lead-time bias and length-time bias are errors related to the natural history of the disease and timing of diagnosis (Figure 3). Lead-time bias occurs when diagnosis is made earlier than usual in a group of patients, independently of disease progression, such as in early referrals [7]. Measuring survival from dialysis initiation makes prognosis appear better in those who started dialysis with better renal function [8]. Length-time bias occurs when there is a differential distribution of subgroups by level of exposure to a risk factor, where the subgroups have the same disease, but different rates of progression (from biologic onset to death). A higher speed of progression may reduce the likelihood of timely diagnosis with consequent under-representation of faster progressors and overestimation of the survival times depending on the study design. For example, those with persistent heavy proteinuria would be expected to have a shorter length of time between disease onset and ESRD than those with lesser degrees of proteinuria. In a prognostic study of a proteinuric disease, length time bias might occur if prevalent cases were recruited. Prognosis would appear more benign than in reality, since such prevalent case samples contain a smaller proportion of subjects with heavy proteinuria than samples of incident patients. Similarly, screening programmes for chronic diseases tend to detect more subjects with slowly progressive forms and longer pre-clinical phases. Length time bias may partly explain the apparent survival advantage observed in non-experimental studies comparing screening programmes to routine clinical care [9].

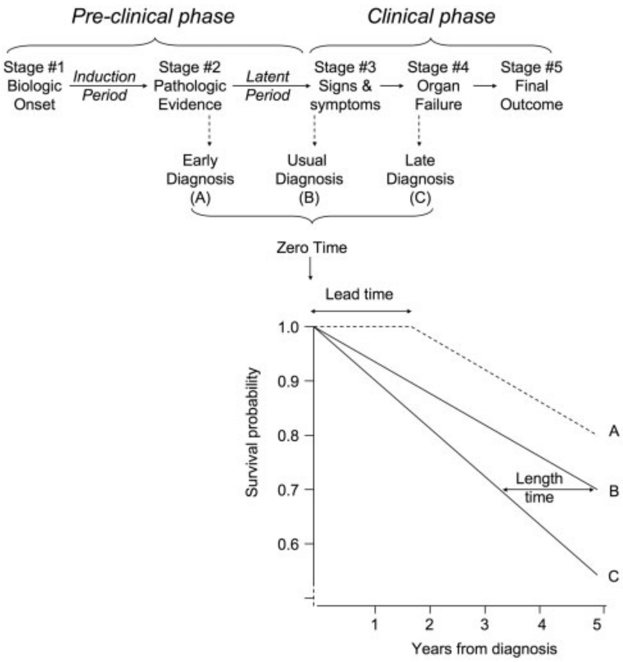

Fig. 3. Natural course of a disease and possible biases related to the timing of diagnosis. The course of a disease is represented as a sequence of stages, from biologic onset to a final outcome such as death. Disease diagnosis can be made as soon as pathologic lesions are detectable (stage #2); when initial signs and symptoms occur (stage #3); or later on (stage #4). *Lead-time bias* occurs when subjects are diagnosed earlier (A) than usual (B) independent of the speed of progression of the disease. If group A, for example, contains more subjects diagnosed in stage #2 (e.g. early nephrology referrals who start dialysis with higher renal function) the apparent observed benefit (e.g. 10% higher 5-year survival probability) is due to a zero-time shift backward from the time of usual diagnosis leading to a longer observed duration of illness. *Length-time bias* occurs when more severe forms of the disease (C), characterized by shorter induction and/or latent periods and lower likelihood of early or usual diagnosis, are unbalanced by group. The apparent difference in prognosis (e.g. 15% 5-year survival probability) is due not only to differences in disease progression (slope) but also to differences in timing of diagnosis.

This may also be problematic in RCTs of prevalent rather than incident patients, because the prevalent group would have lower overall basal risk of the event of interest, and consequently lower study power, increasing the risk of false negative results [1]. For example, in the CREATE trial of different haemoglobin targets in chronic kidney disease, the annual event rate was lower than expected (6% *vs* 15%). Volunteer bias and Hawthorne effect (whereby the control group performs better than expected) may have played a role. However, the study enrolled also prevalent subjects, whereas the sample size was estimated from event rates in incident studies [10].

Experimental designs for intervention questions

The RCT design is appropriate for assessment of clinical effects of drugs, procedures, or care processes, definition of target levels in risk factor modification (e.g. blood pressure, lipid levels and proteinuria), and assessment of the impact of screening programmes [1]. Comparison to a placebo may be appropriate if no current standard therapy exists. When accepted therapies exist (e.g. statins as lipid lowering agents, ACE-I for chronic kidney disease progression, etc), the comparison is an 'active' control group that receives usual or recommended therapy.

Figure 1 shows an example of the most common type of RCT, the two group *parallel-arm* trial. However, trials can compare any number of groups. In *factorial trials* at least two active therapies (A; B) and their combination (AB) are compared with a control (C). Factorial designs can be efficient since more therapies are simultaneously tested in the same study. However, the efficiency and the appropriate sample size are affected by the impact of multiple testing on both type I and type II error, and whether there is an interaction between the effects of the therapies. In the absence of interaction, the effect of A, for example, can be determined by comparing A + AB to B + C. Interactions where use of A enhances the effectiveness of B, for example, do not reduce the power of the study. However, if there is antagonism between treatments, the sample size can be inadequate [1].

The HEMO study used a two-by-two factorial design, and tested two interventions, with no interaction assumption [11]. The trial failed to show the existence of a 25% reduction in the risk of death for either intervention: higher *vs* standard dialysis dose or use of high *vs* low flux membranes [11]. The AASK trial had a two-by-three factorial design (six groups) testing the effect on renal function decline (primary outcome) and on composite end-points (time to renal function halving, ESRD, or death) of two blood pressure levels by three anti-hypertensive treatments (Ramipril, Metoprolol, Amlodipine) with no interaction assumption [12]. Since there were multiple possible comparisons, three primary treatment comparisons were pre-specified: lower *vs* usual blood pressure goals, Ramipril *vs* Metoprolol and Amlodipine *vs* Metoprolol. The only significant findings reported in this study should be considered with caution since (i) they were effects on the secondary outcome and the study power is estimated on the primary outcome measure; (ii) the level of significance of one of these effect (Ramipril *vs* Metoprolol) was only $P = 0.04$ (non-significant after considering multiple testing) and (iii) Ramipril *vs* Amplodipine had not been pre-specified [12].

The *cross-over design* is an alternative solution when the outcome is reversible. In this design, each participant serves as their own control by receiving each treatment in a randomly specified sequence. A washout period is used between treatments, to prevent carryover of the effect of the first treatment to the subsequent periods. The design is efficient in that treatments are compared within individuals, reducing the variation or noise due to subject differences. However, limitations include possible differential carryover (one of the treatments tends to have a longer effect once stopped); period effects (different response of disease to early versus later therapy), and a greater impact of missing data because they compromise within subject comparison and therefore variance reduction [3]. For example, Schjoedt *et al.* [13] used a cross-over design, to test whether spironolactone reduces proteinuria in diabetic subjects with nephrotic syndrome. Patients were treated in random order with spironolactone 25 mg once daily and matched placebo for 2 months, in addition to ongoing antihypertensive treatment, including an angiotensin-converting enzyme inhibitor or an angiotensin II receptor blocker. No washout period was planned between the two treatment periods, although the hypothesis of no carryover does not seem to be biologically tenable, considering the mechanism of action of the drug. Instead, the investigators searched for evidence of carryover. This was excluded based on statistical testing. However, the assumption underlying this approach (no carryover in absence of statistical support) is questionable, since such tests have limited power [1,14].

Finally, RCTs may attempt to show that one treatment is not-inferior (sometimes incorrectly called equivalence) rather than to establish its superiority to a comparable intervention [15]. These studies are often done when new agents are being added to a class (e.g. another ACE inhibitor), or when a new therapy is already known to be cheaper or safer than an existing standard. In such RCTs, the study size is estimated based on a pre-specified maximum difference that would still be considered irrelevant. For example, the claim might be made that a new ACE inhibitor is non-inferior to Enalapril, if the mean 24 h blood pressure difference between them was no more than 3 mmHg. Non-inferiority trials have been criticized, as imperfections in study execution, which tend to prevent detection of a difference between treatments, actually work in favour of a conclusion of non-inferiority. Thus, in distinction to the usual superiority trial, poorly done studies may lead to the desired outcome for the study sponsor.

Designs for diagnostic questions

When assessing a diagnostic test the reference or 'gold standard' tests for the suspected target disorders are often either inaccessible to clinicians or avoided for reasons of cost or risk. Therefore the relationship between more easily measured phenomena (patient history, physical and instrumental examination, and levels of constituents of body fluids and tissues) and the final diagnosis is an important subject of clinical research. Unfortunately, even the most promising diagnostic tests are never completely accurate.

Table 2. Level of evidence in diagnostic studies using troponin as test (T) and acute myocardial infarction (AMI) as target disorder (D)

Diagnostic question	Direction	Design	Problems	Example	Ref
Do D^+ patients have different levels of T?	From D back to T	Cross-sectional	Reverse association Sampling bias	Difference in Troponin levels by AMI +/−	[16]
Are patients T^+ more likely to be D^+?	From T to D	Cross-sectional	Effectiveness not assessed Sampling bias	Troponin performance in distinguishing AMI +/−	[16,17]
Does the level of T predict $D^{+/-}$?	From T to D	Longitudinal	Missing data Sampling bias	Outcome study in subject at risk for AMI	[17]
Do tested patients have better final outcomes than similar patients who do not?	From T to D	Experiment	Missing data	Outcome (randomized) comparison in subject at risk for AMI	[19]

Positive (+); Negative (−). Missing data are possible in longitudinal or experimental designs: e.g. subjects lost before assessment or with data not interpretable. Strategies should be set up to (i) minimize the likelihood of missing information and (ii) plan how subjects with missing information can be treated avoiding their exclusion (e.g. sensitivity analysis, propensity analysis, etc.).

Table 3. Measures of association in diagnostic studies

		Diagnosis (D)		
		Positive (D^+)	Negative (D^-)	
Test (T)	Positive (T^+)	$True^+$	$False^-$	$PPV = [D^+]/[T^+]$
	Negative (T^-)	$False^+$	$True^-$	$NPV = [D^-]/[T^-]$
		$SN = [T^+]/[D^+]$	$SP = [T^-]/[D^-]$	$Pr = D^-/Totals$

Test (T) sensitivity (SN) and specificity (SP) are the probabilities of T^+ among D^+ ($T^+|D^+$) and T^- among D^- ($T^-|D^-$), where '|' means 'given' or 'conditional on'. Positive predictive value (PPV; $D^+|T^+$) and negative predictive value (NPV; $D^-|T^-$) are posterior or post-test probabilities. Sensitivity and specificity are relatively stable test characteristics since they depend on the mechanism of detection/action and the population characteristics. Conversely, PPV and NPV vary depending on disease prevalence. The likelihood ratio of a positive test (LR^+) is the ratio of true positive and false negative rates, SN/[1–SP]. The likelihood ratio of a negative test (LR^-) is the ratio of true negative and false positive rates, SP/[1–SN]. Likelihood ratios estimate how much more likely the presence and absence of the disease are when the results of the test are positive and negative respectively. Of note, the $False^+$ rate and $False^-$ rate correspond to the type I (alpha) and type II (beta) error rates of an outcome study.

For tests with continuous outcome values, such as serum sodium concentration, clinicians need to know reference (normal) values to identify disease. From an epidemiological perspective, these reference values are best defined based on the diagnostic relevance rather than distributional assumptions (Gaussian for example). In other words, by chance, a fraction of a population without disease will have a test result that differs from the mean by some amount. However, the test becomes useful to clinicians when unusually high or low values are generally associated with some clinical condition of importance. For example, reference values of troponin T and I have been established in outcome studies of subject with suspected myocardial infarction, rather than assessing their distribution in the general population.

Clinical implications of test results should ideally be assessed in four types of diagnostic studies. Table 2 shows examples from troponin studies in coronary syndromes. As a first step, one might compare test results among those known to have established disease, to results from those disease-free [16]. Cross-sectional studies can address this question (Figure 1). However, since the direction of interpretation is from diagnosis back to the test, the results do not assess test performance. To examine test performance (Table 3) requires data on whether those with positive test results are more likely to have the disease than those with normal results [17]. When the test variable is not binary (i.e. when it can assume more than two values) it is possible to assess the trade-off between sensitivity and specificity at different test result cut-off points. In such instances, classification into just two groups is wasteful of information. Distinction of at least three classes is more useful. For example, a Dutch study identified three levels of serum creatinine in hypertensive subjects (≤70, 71–110, >110 μmol/l) associated with likelihood ratios of renal artery stenosis of 0.31, 0.77 and 4, respectively [18]. This means that the third category gives reasonable evidence for stenosis, the first against stenosis and the intermediate is uninformative, as likelihood ratios between 0.5 and 2 are considered uninformative. The Receiver Operating Characteristics (ROC) plot is one way to investigate to what extent the test results differ among people who do or do not have the diagnosis of interest without requiring any data grouping [19]. The ROC curve is a plot obtained computing sensitivity and specificity for every distinct observed test value and plotting sensitivity against 1—specificity. Diagnostic test accuracy is assessed estimating the area under the ROC curve (AUC), which corresponds to the probability that a random person with the disease has a higher test value than a random person without disease. In other words, if the test has an AUC of 0.8 and results are used to distinguish which of the two persons has the disease, the test will be right 80% of the times. The area is 1 for perfect tests and 0.5 for uninformative tests.

In all these diagnostic studies, it is crucial to ensure independent blind assessment of results of the test being assessed and the gold standard to which it is compared, without the completion of either being contingent on results of the other.

Longitudinal studies are required to assess diagnostic tests aimed at predicting future prognosis or development of established disease [17]. The most stringent evaluation of a diagnostic test is to determine whether those tested have more rapid and accurate diagnosis, and as a result better health outcomes, than those not tested. The RCT design is the proper tool to answer this type of question [10,20].

A final issue of great interest for nephrologists is the applicability of findings from different settings to the renal population. The performance of cardiac markers such as troponin, for diagnosis of acute coronary syndromes, is less accurate in patients with kidney disease than in those with more normal kidney function [21], although their prognostic value is generally maintained [22].

Maximizing the validity of non-experimental studies

When randomization is not feasible, the knowledge of the most important sources of bias is important, to increase the validity of any study. This may happen for a variety of reasons: when study participants cannot be assigned to intervention groups by chance either for ethical reasons (e.g. in a study of smoking), or participant willingness (e.g. comparing haemo- to peritoneal dialysis), the exposure is fixed (e.g. gender), or the disease is rare and participants cannot be enrolled in a timely manner. When strategies are in place to prevent bias, non-experimental studies have been shown to yield similar results to rigorous RCTs [23]. These strategies are summarized in Table 4. However, also in non-experimental studies, strategies

Table 4. Strategies to maximize validity in non-experimental studies

Phase	Error type	Problem/contamination	Strategy	Example HD *vs* PD[a]
Definition	Question not reflecting the idea	Problem of construct validity: ideal sample ≠ target population	Specify the question explicitly and verify consistency with the idea	Define Patients, HD, PD, and outcomes
Sampling	Wrong sample	Generalizability issue: actual ≠ ideal sample	Referral patterns definition; no late referral	Patient suitability for both PD and HD options; training
	Survivor bias	Generalizability issue: actual sample ≠ ideal sample	Definition of inception cohort, common entry stage; no prevalent pts	Chronic patients commencing dialysis for the first time
	Selection, confounding, length-time bias	Differences in known and unknown prognostic factors, including disease severity	Inclusion/exclusion criteria definition; good quality of care & available resources for both options in all centres	Define renal function level when access is inserted; exclude severe comorbidity (cancer, severe cardiac disease) & centres where PD/HD not equally offered
Measurement	Lead-time bias	Outcome measured differently by group or inconsistently with the definitions	Pre-specified outcome definition; define zero time for survival analysis accordingly in both groups	Time from enrolment to death for any (primary) and cardiac/infectious (secondary) causes
	Confounding bias	Factors known to impact outcomes not included in the data collection	Pre-specified potential prognostic factors (baseline and updated) for adjustment	Social status, familial support, diabetes, age, gender, compliance, etc)
Follow-up	Attrition bias	Drop out rates differ by group	Maximize completeness of follow-up; intention to treat analysis	Plan contacts; decide whether transplant is a reason for censoring or a time varying covariate
	Dilution bias	Drop-in (treatment change)	Consider dilution in the sample size estimation	Increase sample size to detect a smaller effect
Assessment of the data	Detection bias	Differential outcome assessment	Assessor blinding	Necessary for cause specific not for overall mortality
Analysis	Increase type I error rate (multiple testing)	Primary model not specified; use of unplanned analyses	Pre-specified hypothesis & primary model (which stratifying &/or time-varying variables to use); secondary analyses	Specify the role of previous cardiac events, diabetes, transplant, death before *vs* after dialysis starts, etc
	Selection and confounding	Reasons for drop-in -out related to the exposure	Intention to treat analysis	Maintain subjects in the original exposure group

[a] Example question: 'In patients receiving pre-dialysis care, which dialysis option is associated with longer survival, haemodialysis (HD) or peritoneal dialysis (PD)?'

that maximize internal validity tend to reduce generalizability of the results, and vice versa [1]. For example, among the most common confounding factors in hard outcome studies, age, gender and race can be more easily defined and more consistently and accurately measured than other cardiovascular risk factors (hypertension, dyslipidaemia, smoking, physical exercise, body mass index) and important comorbid conditions (diabetes, cardiovascular disease, malignancies). This has implications for the cost and complexity of efforts to increase internal validity by controlling for confounders (e.g. collection of detailed information about smoking and multiple measurements of cholesterol levels over time might be required). Furthermore reducing confounding, by participant selection based on strict eligibility criteria, limits applicability of the results (e.g. exclusion of some ethnic groups, patients with systemic diseases or worse prognosis).

Research questions in genetic epidemiology

Genetic disorders often present additional challenges to those who design clinical studies and may require adaptation of methods or specific solutions. Definition of the start time in longitudinal studies, and identification of patient and controls to compare outcomes are three key issues. The first is usually addressed using the birth date as time zero for survival analysis. The second can be solved by enrolling incident patients to prevent survivor bias. For example, diagnostic and prognostic questions were addressed in a study comparing time to ESRD in the two main genetic forms of adult (autosomal dominant) polycystic kidney disease, ADPKD1 and ADPKD2 [24]. The two main challenges of that study were the definition of the families representative of the population at risk and the identification of carriers. Probands were identified by all nephrologists in the community and pedigrees were constructed to identify all individuals at 50% risk of having autosomal-dominant polycystic kidney disease (ADPKD). Genetic testing was considered the gold standard to identify cases, and when genetic testing was not possible, renal ultrasound using Ravine's criteria was adopted [25]. This test was a reliable indicator of inherited ADPKD in adults who were 30 years or older [24]. Depending on pedigree position, obligate carrier status was demonstrated in some individuals. Thus it was possible to identify most families with ADPKD in the community and to enrol incident family members who carried the ADPKD mutation, and make a reliable prediction of outcome.

Recruitment of participants when the disorder is rare is a problem, because the low frequency of genetic diseases often requires the use of case control designs (retrospective) or longitudinal historical cohort studies [26]. An obvious limitation of this type of studies is that changes in diagnostic criteria and health care over time can influence apparent prognosis.

A final issue is the identification of the appropriate controls to compare outcomes. For some genetic diseases that are not immediately lethal, such as ADPKD, outcomes can be assessed by randomized trials or prospective studies [24]. In rare disorders, matching techniques are often used in choosing appropriate comparison groups when patients cannot be randomly assigned to therapy. For example, a cohort study was conducted to assess the benefits of an implantable cardioverter defibrillator (ICD) in Arrhythmogenic Right Ventricular Cardiomyopathy (ARVC), an autosomal dominant condition that causes sudden cardiac death [27]. The survival of patients with the disorder who received the ICD (cases) was compared with a non-randomly assigned control group, both for practical (low frequency of the disease) and ethical reasons (absence of alternative treatments to prolong survival of affected individuals). To prove that the intervention improved survival, a control group was assembled from family members who carried the ARVD mutation, who did not have an ICD implant matched for age, gender and family. To increase comparability, the controls had to be first or second degree relatives of the cases receiving ICD implantation to reduce genetic variation; had to be of the same gender because survival was worse in males and had to have survived up to the age that the ICD was implanted in the cases. This strategy demonstrated that the survival benefit of ICD was such as to make it a dominant strategy, despite the bias associated with the enrolment of some historical controls.

Clinical relevance *vs* statistical significance

The concepts of clinical relevance and statistical significance are often confused. Clinical relevance refers to the amount of benefit or harm resulting from an exposure or intervention sufficient to change clinical practice or health policy. In planning study sample size, the researcher has to determine the minimum level of effect that would have clinical relevance [1]. The level of statistical significance chosen is the probability that the observed results are due to chance alone. This will correspond to the probability of making a type I error, i.e. claiming an effect when in fact there is none. By convention, this probability is usually 0.05 (but can be as low as 0.01). The *P*-value or the limits of the appropriate confidence interval (a 95% interval is equivalent to a significance level of 0.05 for example) is examined, to see if the results of the study might be explained by chance. If $P < 0.05$, the null hypothesis of no effect is rejected in favour of the study hypothesis, despite it still being possible that the observed results are simply due to chance. However, since statistical significance depends on both the magnitude of effect and the sample size, trials with very large sample sizes can theoretically detect statistically significant but very small effects, that are of no clinical relevance.

Figure 4 summarizes the two problems related to the confusion surrounding clinical relevance and statistical significance. Two aspects must be

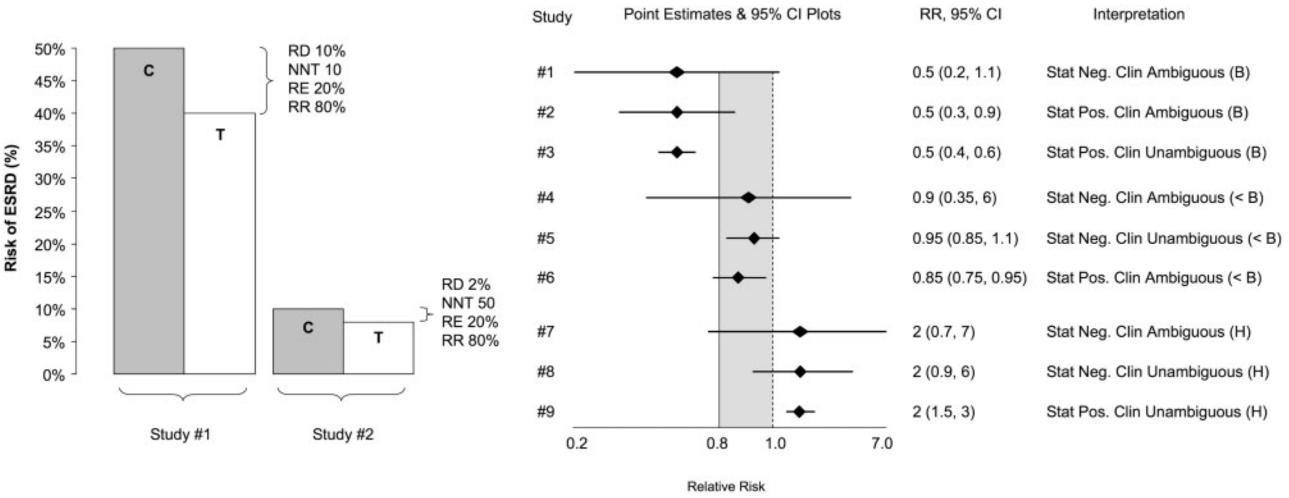

Fig. 4. Left panel: Clinical relevance *vs* statistical significance: risk difference *vs* measures of relative effect. The bar chart shows the results of two studies of an intervention (T) to prevent end-stage renal disease (ESRD); in populations where controls (C) are at higher risk (study #1) and lower risk (study #2). The risk difference (RD = risk among C – risk among T) informs about the size of the treatment effect, whereas the relative effect (RE = risk difference/risk in controls) or the relative risk (RR = risk ratio) does not. The Number Needed to Treat (NNT) is the reciprocal of the risk difference and is much higher in study #2. The NNT is the number of subjects necessary to treat to avoid one outcome event (NNT to benefit) or develop a complication (NNT to harm) and can be derived from relative measures [28]. Of note, the RE and the RR may be the same in both studies (and be statistically significant) but only the absolute RD informs about the clinical relevance of the effect. Right panel: Clinical relevance *vs* statistical significance: positive and negative study results. Study results may support conclusions about clinical relevance whether or not they reach statistical significance and vice versa. This occurs because statistical significance requires that a zero effect be excluded, while clinical relevance is declared based on some non-zero effect. For example, suppose the minimum clinically meaningful effect of an intervention be a 20% reduction in the risk of ESRD relative to controls (RR), shown as a cutoff of 0.8. The cutoff for declaring statistical significance is 1 (null effect for RR). The upper bound of the 95% Confidence Interval (CI) of studies where RR <1 (the intervention has more benefit—B than harm—H) can overlap all (# 1), one (#2) or neither cutoffs (#3). The same may occur for the lower bound of the 95% CI of studies where RR >1 (studies #7 to #9). Result interpretation depends on whether the RR estimate is less than or greater than 1 and which, if any, cutoff boundary is included in the 95% CI around the point estimate. Greater uncertainty occurs when the point estimates are between the two cutoffs (grey zone, studies #4 to #6).

considered: the effect measure chosen to demonstrate the importance of the effect (Figure 4, left panel) and the distinction between the chosen level of clinical relevance and statistical significance (Figure 4, right panel). This is important, since results may be statistically positive (do not support the null hypothesis) but clinically ambiguous (do not support the clinical hypothesis).

Reporting

Adequate reporting is critical to the proper interpretation and evaluation of any study results. Guidelines for reporting primary (CONSORT, STROBE and STARD for example) and secondary studies (QUORUM) are in place to help both investigators and consumers of clinical research [29–32]. Scientific reports may not fully reflect how the investigators conducted their studies, but the quality of the scientific report is a reasonable marker for how the overall project was conducted. The interested reader is referred to the above-referenced citations, for more details of what to look for in reports from prognostic, diagnostic and intervention studies.

Acknowledgements. P.R. held a young investigator award from the Italian Society of Nephrology for the year 2005-2006 and received funding from the EU (Marie Curie Actions-OIF, proposal #021676) for the year 2006-2007.

Conflict of interest statement. None declared.

References

1. Ravani P, Curtis B, Parfrey PS, Barrett BJ. Clinical research of kidney diseases I: researchable questions and valid answers. *Nephrol Dial Transplant* 2007; XX: zzz–zzz
2. http://www.cebm.net/levels_of_evidence.asp (last accessed March 23, 2007)
3. http://www.cebm.utoronto.ca/index.htm (last accessed March 23, 2007)
4. Keough-Ryan T, Hutchinson T, MacGibbon B, Senecal M. Studies of prognostic factors in end-stage renal disease: an epidemiological and statistical critique. *Am J Kidney Dis* 2002; 39: 1196–1205
5. Singh AK, Szczech L, Tang KL *et al*. CHOIR investigators: correction of anemia with epoetin alfa in chronic kidney disease. *N Engl J Med* 2006; 16: 2085–2098
6. Scolari F, Ravani P, Gaggi R *et al*. The challenge of diagnosing atheroembolic renal disease: clinical features and prognostic factors. *Circulation* (in press)
7. Lameire N, Wauters JP, Teruel JL, Van Biesen W, Vanholder R. An update on the referral pattern of patients with end-stage renal disease. *Kidney Int* 2002; 80: 27–34
8. Lameire N, Biesen WV, Vanholder R. Initiation of dialysis–is the problem solved by NECOSAD? *Nephrol Dial Transplant* 2002; 17: 1550–1552

9. Hewitson P, Glasziou P, Irwig L, Towler B, Watson E. Screening for colorectal cancer using the faecal occult blood test, Hemoccult. *Cochrane Database Syst Rev* 2007; 24: CD001216
10. Drueke T, Locatelli F, Clyne N *et al*. Normalization of hemoglobin level in patients with chronic kidney disease and anemia. *New Engl J Med* 2006; 355: 2071–2208
11. Eknoyan G, Beck GJ, Cheung AK *et al*. Hemodialysis (HEMO) Study Group: effect of dialysis dose and membrane flux in maintenance hemodialysis. *N Engl J Med* 2002; 347: 2010–2019
12. Wright JT Jr, Bakris G, Greene T *et al*. African American Study of Kidney Disease and Hypertension Study Group: effect of blood pressure lowering and antihypertensive drug class on progression of hypertensive kidney disease: results from the AASK trial. *JAMA* 2002; 288: 2421–2431
13. Schjoedt KJ, Rossing K, Juhl TR *et al*. Beneficial impact of spironolactone on nephrotic range albuminuria in diabetic nephropathy. *Kidney Int* 2006; 70: 536–542
14. Sibbald B, Roberts C. Understanding controlled trials: crossover trials. *Br Med J* 1998; 316: 1719
15. Salvadori M, Holzer H, de Mattos A *et al*. The ERL B301 Study Groups: enteric-coated mycophenolate sodium is therapeutically equivalent to mycophenolate mofetil in de novo renal transplant patients. *Am J Transplant* 2004; 4: 231–236
16. Majeed R, Jaleel A, Siddiqui IA, Sandila P, Baseer A. Comparison of troponin T and enzyme levels in acute myocardial infarction and skeletal muscle injury. *J Ayub Med Coll Abbottabad* 2002; 14: 5–7
17. Antman EM, Grudzien C, Sacks DB. Evaluation of a rapid bedside assay for detection of serum cardiac troponin T. *JAMA* 1995; 273: 1279–1282
18. Habbema JDF, Eijkemans R, Krijnen, Knottnerus JA. Analysis of data on the accuracy of diagnostic tests. In: *The Evidence Base of Clinical Diagnosis*. BMJ Books, London: 2002; 117–144
19. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. The interpretation of diagnostic data. In: Sackett DL, Haynes RB, Guyatt GH, Tugwell P, eds. *Clinical Epidemiology, a Basic Science for Clinical Medicine*. Little, Brown and Company, Toronto, CA: 1991; 117–119
20. Alp NJ, Bell JA, Shahi M. A rapid troponin-I-based protocol for assessing acute chest pain. *Q J Med* 2001; 94: 687–694
21. Mockel M, Schindler R, Knorr L *et al*. Prognostic value of cardiac troponin T and I elevations in renal disease patients without acute coronary syndromes: a 9-month outcome analysis. *Nephrol Dial Transplant* 1999; 14: 1489–1495
22. Aviles RJ, Askari AT, Lindahl B *et al*. Troponin T levels in patients with acute coronary syndromes, with or without renal dysfunction. *N Engl J Med* 2002; 346: 2047–2052
23. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000; 342: 1887–1892
24. Dicks E, Ravani P, Langman D, Davidson WS, Pei Y, Parfrey PS. Incident renal events and risk factors in autosomal dominant polycystic kidney disease: a population and family-based cohort followed for 22 years. *Clin J Am Soc Nephrol* 2006; 1: 710–717
25. Ravine D, Gibson RN, Walker RG, Sheffield LJ, Kincaid-Smith P, Danks DM. Evaluation of ultrasonographic diagnostic criteria for autosomal dominant polycystic kidney disease. *Lancet* 1994; 343: 824–826
26. Moore SJ, Green JS, Fan Y *et al*. Clinical and genetic epidemiology of Bardet-Biedl syndrome in Newfoundland: a 22-year prospective, population-based, cohort study. *Am J Med Genet* 2005; 132: 352–360
27. Hodgkinson KA, Parfrey PS, Bassett AS *et al*. The impact of implantable cardioverter-defibrillator therapy on survival in autosomal-dominant arrhythmogenic right ventricular cardiomyopathy (ARVD5). *J Am Coll Cardiol* 2005; 45: 400–408
28. Altman DG, Andersen PK. Calculating the number needed to treat for trials where the outcome is time to an event. *Br Med J* 1999; 319: 1492–1495
29. Moher D, Schulz KF, Altman DG, for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Lancet* 2001; 357: 1191–1194
30. http://www.strobe-statement.org/ (last accessed March 23, 2007)
31. http://www.consort-statement.org/stardstatement.htm (last accessed March 23, 2007)
32. David Moher, Deborah J Cook, Susan Eastwood, Ingram Olkin, Drummond Rennie, Donna F Stroup, for the QUOROM Group.Improving the quality of reports of meta-analyses of randomized controlled trials: the QUOROM statement. *Lancet* 1999; 354: 1896–1900

Nephrol Dial Transplant (2008) 23: 2439
doi: 10.1093/ndt/gfn199
Advance Access publication 9 April 2008

Gfm433—Erratum

Clinical research of kidney diseases II: problems of study design

Pietro Ravani[1,2], Patrick S. Parfrey[1], Elizabeth Dicks[1] and Brendan J. Barrett[1]

[1]Clinical Epidemiology Unit, Faculty of Medicine, Memorial University of Newfoundland, Canada and
[2]Divisione di Nefrologia, Azienda Istituti Ospitalieri di Cremona, Italy

Nephrol. Dial. Transplant., 2007; doi:10.1093/ndt/gfm433

Table 3. Measures of association in diagnostic studies

		Diagnosis (D)		
		Positive (D$^+$)	Negative (D$^-$)	
Test (T)	Positive (T$^+$)	True$^+$	False$^+$	PPV = [D$^+$]/[T$^+$]
	Negative (T$^-$)	False$^-$	True$^-$	NPV = [D$^-$]/[T$^-$]
		SN = [T$^+$]/[D$^+$]	SP = [T$^-$]/[D$^-$]	Pr = D$^+$/Totals

Test (T) sensitivity (SN) and specificity (SP) are the probabilities of T$^+$ among D$^+$ (T$^+$|D$^+$) and T$^-$ among D$^-$ (T$^-$|D$^-$), where '|' means 'given' or 'conditional on'. Positive predictive value (PPV; D$^+$|T$^+$) and negative predictive value (NPV; D$^-$|T$^-$) are posterior or post-test probabilities. Sensitivity and specificity are relatively stable test characteristics since they depend on the mechanism of detection/action and the population characteristics. Conversely, PPV and NPV vary depending on disease prevalence (Pr = D$^+$/Totals). The likelihood ratio of a positive test (LR$^+$) is the ratio of true positive and false positive rates, SN/[1−SP]. The likelihood ratio of a negative test (LR$^-$) is the ratio of false negative and true negative rates, (1−SN)/SP. Likelihood ratios estimate how much more likely the presence and absence of the disease are when the results of the test are positive and negative respectively. Of note, the False$^+$ rate and False$^-$ rate correspond to the type I (alpha) and type II (beta) error rates of an outcome study.

This table was previously published with errors. The authors would like to apologize for this mistake and any inconvenience.

NDT Advance Access published April 9, 2008

Nephrol Dial Transplant (2008) 1
doi: 10.1093/ndt/gfn199

Gfm433—Erratum

Clinical research of kidney diseases II: problems of study design

Pietro Ravani[1,2], Patrick S. Parfrey[1], Elizabeth Dicks[1] and Brendan J. Barrett[1]

[1]Clinical Epidemiology Unit, Faculty of Medicine, Memorial University of Newfoundland, Canada and
[2]Divisione di Nefrologia, Azienda Istituti Ospitalieri di Cremona, Italy

Nephrol. Dial. Transplant., 2007; doi:10.1093/ndt/gfm433

Table 3. Measures of association in diagnostic studies

		Diagnosis (D)		
		Positive (D^+)	Negative (D^-)	
Test (T)	Positive (T^+)	True$^+$	False$^+$	PPV = $[D^+]/[T^+]$
	Negative (T^-)	False$^-$	True$^-$	NPV = $[D^-]/[T^-]$
		SN = $[T^+]/[D^+]$	SP = $[T^-]/[D^-]$	Pr = D^+/Totals

Test (T) sensitivity (SN) and specificity (SP) are the probabilities of T^+ among D^+ ($T^+|D^+$) and T^- among D^- ($T^-|D^-$), where '|' means 'given' or 'conditional on'. Positive predictive value (PPV; $D^+|T^+$) and negative predictive value (NPV; $D^-|T^-$) are posterior or post-test probabilities. Sensitivity and specificity are relatively stable test characteristics since they depend on the mechanism of detection/action and the population characteristics. Conversely, PPV and NPV vary depending on disease prevalence (Pr = D^+/Totals). The likelihood ratio of a positive test (LR^+) is the ratio of true positive and false positive rates, SN/[1−SP]. The likelihood ratio of a negative test (LR^-) is the ratio of false negative and true negative rates, (1−SN)/SP. Likelihood ratios estimate how much more likely the presence and absence of the disease are when the results of the test are positive and negative respectively. Of note, the False$^+$ rate and False$^-$ rate correspond to the type I (alpha) and type II (beta) error rates of an outcome study.

This table was previously published with errors. The authors would like to apologize for this mistake and any inconvenience.