Clinical research of kidney diseases 1: researchable questions and valid answers

Pietro Ravani\textsuperscript{1,2}, Patrick S. Parfrey\textsuperscript{1}, Bryan Curtis\textsuperscript{1} and Brendan J. Barrett\textsuperscript{1}

\textsuperscript{1}Clinical Epidemiology Unit, Faculty of Medicine, Memorial University of Newfoundland, Canada and \textsuperscript{2}Divisione di Nefrologia, Azienda Istituti Ospitalieri di Cremona, Italy

Introduction

Clinical epidemiology is the science of human disease investigation. It has a broad scope, ranging from the study of disease occurrence and prognosis to diagnostic and management options. This series of articles will discuss general and specific issues related to the design, analysis and interpretation of studies using nephrology examples. This first paper introduces general issues of ‘definition’ and ‘measurement’ in clinical studies, along with the related concepts of ‘reliability’ and ‘validity’, or ‘random’ error and ‘systematic’ error (bias).

Research objectives

Much medical research can be simplified as the study of the frequency of an illness and the assessment of some ‘input–output relationship’ in a defined patient population. The purpose is to characterize specific diseases and establish whether an explanatory input variable or exposure (test, predictor or intervention) is related to an output or outcome (gold standard result, response variable).

The goal of research is to estimate these phenomena in the population (characteristics or parameters) by making measurements on samples from the target population. The hope is that the study estimates be as close as possible to the true values in the population (accuracy) with little uncertainty (imprecision) around them (Table 1). However, an error component exists in any study. This is the difference between the value observed in the sample and the true value or phenomenon of interest in the parent population.

Error types

There are two main types of errors: random or accidental error; and systematic error (bias). Random errors are due to chance and compensate in large enough population samples, since their average effect is zero. Systematic errors are non-compensating distortions in measurement (Figure 1). Mistakes caused by carelessness, or human fallibility (e.g. incorrect use of an instrument, error in recording or in calculations), may contribute to both random and systematic error. Both errors arise from many sources and both can be minimized using different strategies. However, their control can be costly and complete elimination is impossible. Systematic error, as opposed to random error, is not limited by increasing the study size and replicates if the study is repeated.

Confounding is a special error, since it is due to chance in experimental designs but is a bias in non-experimental studies. Confounding occurs when the effect of the exposure is mixed with that of another variable (confounder) related to both exposure and outcome without lying in the causal pathway between them. For example if late referral is found to be associated with higher dialysis mortality, while age varies by referral status, it is important to consider the confounding effect of age in the assessment of the relationship between referral and death [1].

For any given study, the aim of design is to minimize error. In some cases, pilot studies are helpful in identifying the main potential sources of error (known sources of variability and bias—Table 1) such that the design of the main study can control them [2,3]. Some errors are specific to some designs.
and will be discussed in a subsequent paper of this series. Both random and systematic errors can occur during all stages of a study, from conceptualization of the idea to sampling (participant selection) and actual measurements (information collection).

Ideas and structure of clinical research

Most clinical research projects share the same general structure, irrespective of whether they address diagnostic, prognostic or intervention questions (Figure 2). The research process usually starts with a general idea (construct) or initial problem. For example, a randomized controlled trial to compare models of care delivery might derive from the idea that structured multidisciplinary pre-dialysis care may be of greater benefit for patients than traditional follow-up [4]. Research ideas may originate from practical problems, request for proposals by funding agencies or private companies, or reading the literature and thinking of ways to extend or refine previous research. Literature review is always required, to identify related research, define the knowledge gap, avoid redundancy

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1. The effect of the error type on study results. Each panel compares the distribution of a parameter observed in a study (continuous lines) and the corresponding true distribution (dashed lines). Random error lowers the precision of the estimates increasing the dispersion of the observed values around the average (studies #3 and #4). Systematic error (bias) causes incorrect estimates or ‘deviations from the truth’: the estimated averages correspond to rings distant from the target centre (studies #1 and #3) even if results are precise (study #1).
when the answer is already clear, and set the research within a proper conceptual and theoretical context based on what is already known. The next step is to generate a researchable question from the general idea.

Definitions

Once a problem is identified and its possible solutions hypothesized (conceptualization), the research idea is translated into a specific study question through three main steps (operationalization). First, the nature of the exposure-outcome relationship is specified. For example, a prognostic study might assess the association between exposure to volatile solvents and glomerular disease [5]; an intervention study might assess the impact of ACE inhibitors on the degree of proteinuria in those with diabetic nephropathy [6]; a diagnostic study might examine the association of Anti-Neutrophilic Cytoplasmic Antibodies level and the presence or activity of vasculitis as determined by biopsy [7]. Second, the broad initial idea is translated into a feasible research project. For example, a broad idea might be to implement screening strategies to improve vascular access outcomes. However, studying all possible screening strategies in all possible vascular accesses is an impossible task for any single study. Narrowing down the area of research is necessary to formulate an answerable question. For example, physical examination of the fistula might be compared with some specific access flow measurement technique in preventing fistula loss. Third, the target population of the study—the set of individuals to whom the results of the trial might be applied (e.g. haemodialysis patients receiving a new fistula) is determined along with a meaningful disease measure—pre-specified study outcome (e.g. fistula survival).

In stating researchable questions, it is crucial to define the Patients, Interventions, Controls and Outcomes (PICO) of relevance. The study question should define (i) who are the patients to be studied (e.g. prevalent or incident), through clearly defined eligibility criteria. These should specify the problem (e.g. thrombosis of newly created fistula), the comorbid conditions to include because the answer(s) to the research question may vary by their levels (e.g. diabetes, cardiovascular diseases); and those not to include because for them, the question may be of less interest or hardly answerable (e.g. those with short expected survival). Secondly (ii), the type of the exposure (intervention or prognostic factor or test is defined) and its specifics (e.g. what is meant by access monitoring, which technology is used) need to be addressed. Next (iii), the comparison group is defined (e.g. what is meant by standard surveillance, which signs are part of the examination). Finally (iv), the outcome of interest is declared (e.g. what will constitute fistula failure, or how will survival be measured). Following these considerations a researchable question can be posed, e.g.: “Does the addition of monitoring fistulae by access flow using a dilution method to surveillance by pulse assessments and bleeding time after needle withdrawal reduce the risk of fistula loss?”

The operationalization of the study must be consistent with its purpose. If the question is of one efficacy (Does it work in an ideal world?), then the measurement tools identified should be very accurate,
The likelihood of CKD may be positively or negatively related to factors associated with receiving care or working in a particular setting. On the other hand, if a study aimed at understanding the characteristics of patients with CKD referred to a nephrologist, a study of consecutive patients referred for CKD would probably provide a reasonably generalizable result [4].

**Random assignment to exposure and analysis of trials**

If the purpose of the study is to estimate a measure of effect due to some intervention, then the sampling problem is not finished. Here the comparability of study groups, other than with regard to the exposure of interest, must be ensured. Indeed to measure the effect of a therapy, we need to contrast the experience of people given the therapy to those not so treated. However, people differ from one another in myriad ways, some of which might affect the outcome of interest. For example, in a retrospective study using health administrative data, Suissa et al. [10] found an association between long-term exposure to ACE inhibitors and the likelihood of end-stage kidney disease (ESRD). The investigators went to some lengths to establish the comparability of the groups treated with ACE inhibitors to those receiving other antihypertensive medications. Nevertheless, the study database lacked information on the actual level of blood pressure either before or during therapy. As a result, there is potential that level of blood pressure might represent an unmeasured confounder.

To avoid such concerns in studies of therapy, random assignment of study participants to therapy is usually recommended, to ensure comparability of study groups in the long run. Groups of sufficient size help reduce the possibility that some measurable or unmeasurable prognostic factors remain unbalanced across them (random confounding).

The randomization process consists of three interrelated manoeuvres: generation of random allocation sequences; strategies to promote allocation concealment and intention-to-treat analysis.

**Random sampling**

Recruitment of a random sample of the target population is necessary to ensure generalizability of study results. For example, if we wish to estimate the prevalence of Chronic Kidney Disease (CKD) in the general population, the best approach would be to use random sampling, possibly over-sampling some subgroup of particular interest (e.g. members of a racial group) in order to have sufficiently precise estimates for that subgroup [8,9]. In this instance, a sample of subjects drawn from a nephrology or a diabetic clinic, any hospital department, school, or workplace would not be representative of the general population. The likelihood of CKD may be positively or negatively related to factors associated with receiving care or working in a particular setting. On the other hand, if a study aimed at understanding the characteristics of patients with CKD referred to a nephrologist, a study of consecutive patients referred for CKD would probably provide a reasonably generalizable result [4].
recruitment proceeds (i.e. using sequences of 4–10 balanced assignments in random order and of variable size). Alternatively, in adaptive methods, probabilities change as the study progresses to balance exposure and/or potential confounders or outcome (Figure 4).

Allocation concealment is meant to prevent those recruiting trial subjects from the knowledge of upcoming assignment. Such concealed group allocation protects against selection biases that tend to be associated with inflated estimates of therapy effect. Useful ways to implement concealed allocation include the use of central randomization (performed remote from trial location and communicated by an independent person or an interactive voice system), or the use of sequentially numbered sealed opaque envelopes.

Intention-to-treat analysis consists of keeping all randomized patients in their original assigned groups during analysis, regardless of adherence or any protocol deviations. This is necessary to maintain group comparability and is standard for trials of drug effectiveness (i.e. those aimed at showing what a therapy does in practice). Alternatively, per protocol analyses, where only those who conform to study protocol and take therapy as prescribed are analysed, can be used in earlier phase efficacy trials, the aim of which is to establish what a therapy can do under more ideal circumstances. However, phenomena such as loss to follow-up, drop-out (earlier termination), protocol violations (e.g. wrong assignment or dosage) and drop-in (group crossover) can be related to the exposure. For example, if a treatment has side effects, drop-outs can be higher among the least well participants. Patient exclusion after allocation may make treatment appear effective when it is not. Patients not adhering to treatments and drop-ins may have different prognoses and are unlikely to be representative of all study subjects. In these cases, if the amount of drop-outs and drop-ins is not excessive, it may still be possible to interpret analysis comparing randomized groups. For example, drop-out has been common in trials of lifestyle therapies for obesity [12]. In such circumstances, sensitivity analyses can be included as one way to examine the robustness of trial conclusions: each participant is analysed as randomized and those dropping-out earlier in one arm are assumed to have the worst possible outcome, whilst those dropping-out in the parallel arm are assumed to have the best possible outcome, and vice versa. Drop-in also affects the ability to interpret results. This was a particular problem in a trial comparing medical therapy to angioplasty for reno-vascular disease [13]. A large fraction of those assigned to medical therapy...
subsequently underwent angioplasty, probably because of inadequate response to medical therapy. In such a case, intention to treat analysis would tend to over-estimate the effectiveness of medical therapy, if in fact angioplasty was superior. Drop-in tends to dilute the hypothesized treatment effect. This phenomenon is usually taken into account, hypothesizing an effect smaller than expected when estimating the study size.

**Sample size estimation**

When planning a comparative study, two possible random errors (type I and II) are considered. Type I error occurs if the study provides evidence for a hypothesized effect that in fact does not exist (false positive). The maximum risk of this error is commonly set at 5% (significant overall $P$-value of 0.05). A type II error occurs if the study fails to support an effect that truly exists (false negative). The type II error rate is usually set at 10% or 20%. Consider an experiment where a new drug is compared with placebo in lowering blood pressure (Figure 5). In the absence of blood pressure change, the average values in those receiving placebo ($\mu_0$) vs active drug ($\mu_1$) will not be different ($\mu_0 = \mu_1$). The top curve represents the theoretical distribution of blood pressure changes under this ‘null hypothesis’ ($H_0$). The distribution of the mean observed difference under the null hypothesis is centred at zero ($\mu_1 - \mu_0 = 0$) and has some variability, or dispersion, around the central value (i.e. values $>0$ or $<0$ are possible even if there is no change). The study or ‘alternative hypothesis’ ($H_a$) is that the expected blood pressure change with the study drug will not be compatible with this distribution ($\mu_0 \neq \mu_1$). However, prior to the study, it is unknown whether the alternative distribution lies to the left or right of the null (the drug increases or reduces blood pressure). This is acknowledged representing the type I error at both tails of the null hypothesis distribution ($\alpha/2$ for two-sided alternative hypothesis). Conversely, the type II error lies in the non-rejection region of the alternative hypothesis. Since the $H_0$ and $H_a$ curves overlap and tend to infinity at both sides, study results may be compatible with either hypothesis, although

---

**Fig. 5.** Study power. A new intervention assumed to reduce blood pressure might not change or even increase blood pressure. In all cases, its effects will show some variability. The first curve (#1) represents the distribution of blood pressure changes under the null hypothesis ($H_0$) of no effect. The second (#2) and the third curve (#3) similarly represent the alternative hypothesis ($H_a$) that the drug either increases (#2) or decreases blood pressure (#3) as compared with placebo. The two tails of $H_0$ in dark grey represent the probability of making a type I error (overall 0.05). This error occurs if the study results exceed the corresponding critical value of $Z$ (defining the rejection regions) in either direction, but in fact data belong to the top distribution and the $H_0$ is true (false positive). The probability of making a type II error ($\beta$, in light grey) occurs when the study results are in the non-rejection region (and are thought to belong to $H_0$), but in fact the data belong to either of the two alternative distributions $H_a$ (#2 or #3). In other words, the new drug has an effect but a wrong decision is taken failing to reject the null hypothesis (false negative). The power of a study ($1 - \beta$ white area under the $H_a$; true positive) is the probability that the correct action is taken rejecting a false $H_0$. The study power will be higher if the critical $Z$-value is lower (greater $\alpha$ and smaller non-rejection region, #4); if the curve spread is smaller (less variance in a larger study, #5); or, finally, if the degree of curve separation is larger (greater effect to be detected, #6).
with different probabilities. For this reason, the critical value of a statistic (such as $Z$) is declared (e.g., $z = 1.96$) to set the level of statistical significance (type I error probability) and define the regions of rejection and of non-rejection of the $H_0$.

The study size is estimated from the declared maximum acceptable levels of these errors (the smaller the errors the larger the study), the effect size to be detected (the larger the effect the smaller the study) and the expected outcome variability (the larger the variability the larger the study). Different formulae exist to estimate the sample size depending on the type of response variable and the analytical tool used to assess the input–output relationship [14]. But in all cases the relationship among sample size, variability in the data, effect size, level of significance and study power ($1 – \text{type II error}$) has the same structure.

**Measurement**

**Variable types**

As in all sciences, measurement is a central feature of clinical epidemiology. Both input and output variables are measured on the sample according to the chosen definitions. Inputs can be measured once at baseline, if their value is fixed (e.g. gender), or more then once if their value can change during the study (such as blood pressure or type of therapy). Outputs can also be measured once (e.g. average blood pressure values after 6 months of treatment) or multiple times (repeated measures of continuous variables such as blood pressure or events such as hospital admissions). The information gained from input and output variables depends on the type of observed data (Table 2).

In clinical epidemiology, the type of outcomes influences study design and determines the analytical tool used to study the relationship of interest. For example, suppose the question is whether derangements in divalent ion metabolism increase the risk of cardiovascular events through worsening atherosclerosis and vascular calcification. This question would be addressed by a longitudinal prognostic study in incident haemodialysis patients, with time to event as primary outcome. Baseline plus repeated measures of calcium/phosphate/PTH values [15] as well as information about any vascular diagnostic data can be collected over time. Time-to-event analysis can be used to study the hypothesized associations and test whether vascular damage mediates the association between mineral data and cardiac events. However, lack of multiple assessments of both mineral data and vascular information would weaken any inference drawn from the data [15,16] and would not clarify whether vascular lesions are intermediate variables [17].

Intermediate variables are often considered as surrogate outcome candidates and used as an outcome instead of the final end-point, to reduce sample size and study cost (Table 3). Candidate surrogate outcomes are many and include measures of the underlying pathological process (e.g. vascular calcifications), or of preclinical disease (e.g. left ventricular hypertrophy). However, well validated surrogate variables highly predictive of adverse clinical event, such as systolic blood pressure and LDL cholesterol, are very few and only occasionally persuasive (Table 4).

Furthermore, although these surrogates may be useful in studies of the general population, their relationship with clinical outcomes is not linear in some conditions such as ESRD, making them less useful in those settings [18,19]. Finally, several examples also exist in the nephrology literature concerning the risk of relying upon surrogate markers to advocate treatments or explain cause-effect relationships. For example, divergence between primary outcome (renal function decline) and proteinuria reduction resulting from lower blood pressure goal in

---

**Table 2. Types of variables measured in clinical studies**

<table>
<thead>
<tr>
<th>Type of measurement</th>
<th>Variable characteristics</th>
<th>Example</th>
<th>Appropriate statistics, comparison</th>
<th>Information content; power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualitative nominal</td>
<td>Unordered categories</td>
<td>Sex, blood type, vital status, race, genotypes</td>
<td>Proportions, mode; averages do not make sense</td>
<td>+</td>
</tr>
<tr>
<td>Qualitative ordinal</td>
<td>Ordered categories (not quantifiable intervals)</td>
<td>Degree of pain, CKD stages</td>
<td>Proportions, mode, median (range); no calculation makes sense</td>
<td>++</td>
</tr>
<tr>
<td>Quantitative interval; no meaningful zero</td>
<td>Can take integer (counts; how many) or any values on a scale (continuous; how much)</td>
<td>C and F scales of temperature; pH: ‘zero’ does not mean ‘nothing’</td>
<td>Mode, median (range); mean (SD); only adding, subtracting possible: 100–90 degrees $F$ = 90–80; a pH of 3 is not twice as acidic as a pH of 6</td>
<td>+++</td>
</tr>
<tr>
<td>Quantitative ratio; zero is meaningful</td>
<td>Can take integer (counts; how many) or any values on a scale (continuous; how much)</td>
<td>Number of days; age, weight; K scale of temperature; ‘zero’ actually means ‘nothing’</td>
<td>Mode, median (range); mean (SD); all arithmetic functions possible; 4 grams of a substance is twice a weight of 2 grams</td>
<td>++++</td>
</tr>
</tbody>
</table>
the AASK trial shed uncertainties about the validity of proteinuria as a surrogate outcome of kidney disease [20]. For these reasons, trials generally change practice if they use **hard** clinical events as outcomes. These are final clinical end-points such as ESRD, cardiovascular events (myocardial infarction and stroke, for example) and death. Although doubling of serum creatinine is not a hard clinical endpoint, it is a marker of progression of renal disease used by nephrologists as an endpoint [18].

**Measures of disease frequency and effect**

Hard outcomes are used to measure disease occurrence, as well as to estimate the effects of an exposure (Table 5). Prevalence and incidence of a disease are measures of disease occurrence in a population. Prevalence studies estimate the risk of **having** the disease. They assess the burden of disease in a population and are useful for health care policy. For example, if 25 of 100 subjects have a history of myocardial infarction, the prevalence of the disease in that group of individuals is 0.25. The risk of **getting** the disease is estimated in incidence studies. The observation of new cases over a certain period of time within a population of people at risk (i.e. disease free at the beginning of the study) allows estimation of the incidence of the disease in that population, over the specified time period. If five new cases occur in a year among 100 initially disease free individuals, the incidence of the disease is estimated at 0.05 per year. Measures of effects include comparisons of two or more measurements of disease occurrence by level of a factor. For example, prevalence or incidence rates can be compared by presence or absence of diabetes, using either a rate difference or ratio.

**Measurement errors**

Some systematic and random errors may occur during measurement (Table 1, Figure 2). Of interest to clinical trials are the strategies to reduce performance bias (additional therapeutic interventions preferentially provided to one of the groups) and to limit information...
Table 5. Measures of disease frequency and measures of effect

<table>
<thead>
<tr>
<th>Two by two table</th>
<th>Disease status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure status</td>
<td>$D_1$</td>
</tr>
<tr>
<td></td>
<td>a</td>
</tr>
</tbody>
</table>

**Frequency measures**

- **Prevalence proportion**
  - Quantity of disease ($D/N$) attributable to each unit of observation ($N$) in a defined moment ($t$; point prevalence) or interval ($t_1$ - $t_0$; period prevalence).

- **Incidence proportion (Risk)**
  - Speed with which a certain event ($D_1$) propagates within a population at risk ($N$; subject event free) in a certain time interval ($t_1$ - $t_0$).

- **Incidence rate (Rate)**
  - Number of new cases ($D_1$) over the subjects at risk measured as person-time ($T$).

- **Odds**
  - Probability over the complement of probability (or, number of subjects affected/unaffected; $a/b$ for $E_1$; $c/d$ for $E_0$).

**Effect measures**

- **Risk difference**
  - Difference in a measure of risk estimated among exposed ($E_1$) and unexposed ($E_0$).

- **Risk ratio**
  - Ratio of two risk estimates among $E_1$ and $E_0$.

- **Relative effect**
  - Risk difference relative to the basal risk (which is the risk among $E_0$).

- **Odds ratio**
  - Ratio of two odds estimates (among $E_1$ and $E_0$); cross product ($a*d$)/($c*b$); output of logistic regression.

- **Incidence rate ratio**
  - Ratio of two incidence rates (among $E_1$ and $E_0$); output of Poisson regression (for count outcomes); when the unit of time is very small rates are called hazards (survival analysis).

and detection bias (ascertainment or measurement bias) by masking (blinding) [22]. Masking is a process whereby people are kept unaware of which interventions have been used throughout the study, including when outcome is being assessed. Patient/clinician blinding is not always practical or feasible such as in trials comparing surgery with non-surgical therapy.

Finally, measurement error can occur in the statistical analysis of the data. Important elements to specify in the protocol include: definition of primary and secondary outcome measures; treatment of missing data; subgroup (secondary) analyses of interest; consideration of the inflation of type I error rate with multiple comparisons; potential confounders to control for and possible effect modifiers (interaction). The latter issue has implications for modelling techniques, as discussed in subsequent papers.

**Concluding remarks: external and internal validity**

The operational criteria applied in the design influence the external and internal validity of the study (Figure 2). Both construct validity and external validity relate to generalization. However, construct validity involves generalizing from the study programme/measure to the underlying concept of the programme/measure. It reflects how well the variables in the study (and their relationships) represent the phenomena of interest. For example, how well does the level of proteinuria represent the presence of kidney disease? Construct validity becomes important when a complex process, such as care for CKD, is being described. Maintaining consistency between the idea or concept of a certain care programme and the operational details of its specific components in the study may be challenging. External validity involves generalizing conclusions from the study context to other people, places or times. External validity is reduced if study eligibility criteria are strict, or the exposure or intervention is hard to reproduce in practice. The closer the intended sample is to the target population, the more relevant the study to this wider, but defined, group of people, and the greater its external validity.

For example, the applicability of results of a study comparing angioplasty and surgical revision to correct fistula stenosis is limited to the type of lesion, the portion of the affected conduit and to centres where the relevant treatments can be applied as it was in the study. The internal validity of a study depends primarily on the degree to which bias is minimized. Selection, measurement and confounding biases can all affect the internal validity.

In any study there is a balance between external and internal validity, as it is difficult and costly to maximize both. Designs that have strict inclusion and exclusion criteria tend to maximize internal validity, while compromising external validity. For example, if participants are only selected for a trial of blood pressure lowering if they demonstrate adequate adherence to therapy during a run-in phase, the internal validity will be enhanced through high adherence, and low drop-in and drop-out rates. However, the results of such a trial are not likely to be replicated in usual
practice. Internal validity is especially important in efficacy trials to understand the maximum likely benefit that might be achieved with an intervention, whereas external validity becomes more important in effectiveness studies. Involvement of multiple sites is an important way to enhance both internal validity (faster recruitment, quality control and standardized procedures for data collection, management, and analysis) and external validity (generalizability is enhanced because the study involves patients from several regions).

Acknowledgments. 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

Received for publication: 20.4.07
Accepted in revised form: 7.6.07