The bumpy road to evidence: why many research findings are lost in translation

Thomas F. Lüscher*

European Heart Journal, Zurich Heart House, Zürich, Switzerland

Online publish-ahead-of-print 2 October 2013

We have come a long way

When, more than 2500 years ago, Thales of Milet (624–547 BC) claimed that nature was ruled by laws and not by gods, he changed the world.¹ Such a concept allowed for the discovery and mathematical proof of impersonal causation of what, until then, had been a mystery ruled by unpredictable gods. Indeed, ever since then the understanding of nature—today we would call it the natural sciences—has become a major activity of mankind. With this strategy, we have come a long way. Initially, it allowed us to use the position of stars and planetary motion for navigation and the discovery of new continents; next, it set the basis for the development of engines and technologies; and finally, it led to the discovery of the human body. Eventually, this allowed for the rise of modern medicine, among many other achievements.

As a consequence, theology and philosophy, the dominant disciplines of ancient times, were increasingly challenged and, to a great extent, replaced by scientifically based knowledge about the world, about nature, mankind, and disease. Such knowledge produced practically useful consequences, such as pumps, steam, and later, petrol-driven engines, trains, cars, aeroplanes, and rockets, allowing us to fly to New York or to the moon. In life sciences, it brought about hygiene, anaesthesia, and in turn, aseptic surgery, antibiotics, vaccination, resuscitation, cardiac surgery, and interventional cardiology;²,³ an impressive and unanticipated achievement indeed.

What is evidence?

What is evidence? At the beginning of any discovery stands an individual with curiosity; an initial observation is the first step in the process. When Columbus sat at the beach—as the famous movie by Ridley Scott leads us to believe—watching ships leaving the harbour with his son Diego, he asked him, ‘Look!’

‘Half of the ship has gone’, replied Diego.
‘And now?’
‘It’s gone.’
‘What does it tell you?’
Diego was not sure.

‘It is round’, replied his father, ‘like this’, presenting an orange he was about to eat.

Thus, the interpretation of a finding is as important as the observation itself. But that is not enough; it needs proof, i.e. the journey to the West and the persistence to pursue it. Columbus did it all and discovered a new continent, now known as the Americas.

What is a scientific fact? Above all, it should be provable, i.e. confirmed or falsified by one’s own data and those of other scientists. And indeed, most scientists spend their day providing data to confirm or disprove it. Karl Popper (1902–94) taught us that this is not it.⁴ In fact, observations have to survive the proof of time; the scientific process develops along conjectures and refutations. The statement that all swans are white was falsified with the discovery of black swans in Australia.⁶,⁷ Another important aspect of discovery is the fact that it always evolves within a paradigm,⁸ i.e. in medicine, a basic concept of what the major causes and mechanisms of a disease are. Importantly for what is discussed below, by shear probability new claims are more likely to be falsified; thus, the road to evidence is, by design, a bumpy one.

And indeed, innovations are difficult to predict. Simon Newcomb (1835–1909), a physicist of the 19th century, said: ‘Flying with machines that are heavier than air is without practical importance and senseless, if not completely impossible.’ Yet today, about 3 billion passengers travel by air each year. Around the same time, the famous surgeon Theodor Billroth said: ‘That surgeon who ever would attempt to sew a wound of the heart can be sure of losing any respect of his colleagues for ever.’ Yet today, more than a million cardiac operations are performed every year, not to mention catheter interventions.

Particularly in an applied science such as medicine, the practical consequences of a theory are as important as the fact that it has stood the test of time. Indeed, Anitschkow’s seminal observation in rabbits fed a high-fat diet led to the concept that fat or its components will lead to atherosclerosis and its complications, such as myocardial infarction and stroke.⁹ The Framingham study prospectively studied this relationship and found an association also in humans.¹⁰ However, only the discovery of 3-hydroxy-3-methyl-glutaryl-coenzyme A inhibitors, the statins, enabled studies that eventually proved a causal
relationship. Ever since then, the difference between association and causality has been stressed.

In medicine, different levels of evidence have been distinguished (Figure 1), as follows: clinical intuition, unsystematic clinical experience, case reports and patient cohorts, pathophysiological concepts (the basis of most paradigms in medicine), small trials and meta-analyses or systematic reviews thereof, and finally, large randomized trials.

James Lind and scurvy
Scurvy is a disease which leads to open sores and loss of movement, a condition which, until the 19th century, was particularly prevalent among sailors and soldiers. The ship’s surgeon of the British Royal Navy, James Lind, was the first to find a cure for the disease. While at sea in May 1747, Lind treated some of his sailors who were suffering from scurvy with oranges and lemons, while others received cider, vinegar, sulfuric acid or seawater, along with their usual food. Historically, this has to be considered the first randomized (although not blinded), controlled trial. In spite of the very small number of patients involved, the results conclusively showed that citrus fruits prevented the disease. Lind published his observation in 1753 in his ‘Treatise on the Scurvy’.

Lind’s approach was not adopted by medical researchers until the 20th century, when Austin Bradford Hill (1897–1991), an English epidemiologist and statistician, set out to test the effects of the recently discovered streptomycin in patients with tuberculosis. At this point, it had been recognized that biases of both the patient and the treating physician may influence the perception of the effectiveness of medical interventions. Tuberculosis was an endemic disease commonly treated by bed rest on the ‘Magic Mountain’ and other institutions, mostly at high altitude. Hill wanted to prove the advantages of streptomycin compared with that standard treatment and developed the principle of randomization to exclude as many biases as possible. The results were a breakthrough, both for his scientific approach and for the treatment of tuberculosis. The trial lasted 6 months and involved 52 controls treated by bed rest and 55 patients receiving 2 g of streptomycin four times daily. As he wrote in his seminal article published in the British Medical Journal in 1948: ‘The difference between the two series is statistically significant; the probability of it occurring by chance is less than one in a hundred.’ Ever since then, randomization (today also with blinding of doctors and patients, if possible) and statistical analysis have become the cornerstones of clinical research.

From bench to bedside
It is obvious that any clinical trial rests on the results of often decades of basic research. Indeed, there are several levels of science involving genes, proteins, organelles and cells, tissues, organs, and finally, patients and populations (Figure 2). Evidence has to evolve over many steps in order to graduate from bench to bedside. This process is not unidirectional; indeed, a seminal observation may, for instance, start in a tissue, such as a blood vessel of an animal, where a novel phenomenon, e.g. endothelium-dependent relaxation, is observed, then may progress to organs and organisms, only to move down to the molecular level, where the responsible protein, i.e. endothelial nitric oxide synthase, is discovered. Clinical trials, Mendelian randomization studies and others, may follow later.

In drug discovery, the process is more unidirectional, starting with basic research in cells, tissues, and animals and then moving to phase I studies focusing on safety, pharmacokinetics, and haemodynamics. The dose–effect relationship is then investigated in phase II, while clinical endpoints are the focus of phase III studies (Figure 3). Dose is a difficult issue, particularly in the absence of reliable surrogate endpoints. Indeed, dosages used in vitro or in animal models are often several magnitudes higher than those effective and tolerated in humans.

![Figure 1](https://example.com/figure1.png)

**Figure 1** Levels of evidence in clinical medicine.

![Figure 2](https://example.com/figure2.png)

**Figure 2** Levels of research from bench to bedside.
Lost in translation

When moving through all these levels of research and development, concepts, drugs, and devices may be lost in translation. Why are things lost in translation? First, the hypothesis may be wrong, then the animal models used may not reflect human disease, the data may not be solid or may even be fraudulent, endpoints may not have been well chosen, unrecognized off-target effects may suddenly appear that outweigh the benefits, and finally, there are miscellaneous reasons.

Let us look at a few examples. In the 1980s and 1990s, restenosis after angioplasty or stenting was a real issue. At the same time, the vascular effects of angiotensin II were discovered, and researchers at Roche published an article in Science demonstrating that the angiotensin-converting enzyme (ACE) inhibitor cilazapril prevented intimal hyperplasia induced by vascular injury in the carotid artery of the rat. Swiftly, several clinical trials, MERCATOR and MARCA-TOR, were set up to prove these findings in patients undergoing angioplasty. As it turned out, cilazapril at either a high or a low dose did not prevent restenosis or improve outcomes. Obviously, the rat carotid artery was an inappropriate model, because most interventions worked that later proved ineffective at the clinical level. Once a new molecule, rapamycin, proved effective in the pig model of stent restenosis, the results could be confirmed at the clinical level.

Things can get worse; for example, TGN1412 (or CD28-SuperMAB) was the working name of a supposedly immunomodulatory drug originally intended for the treatment of B cell chronic lymphocytic leukaemia or rheumatoid arthritis. In March 2006, six volunteers were entering a phase 1 trial at Northwick Park Hospital as the first humans to receive the drug. Unexpectedly, the drug caused catastrophic systemic organ failure due to a massive cytokine storm in the volunteers enrolled, despite it being administered at a dose 500 times lower than that found safe in animals. Obviously, the problems resulted from biological actions in humans not foreseen from the experiments in rats and mice.

Although mice share about 80% of their working DNA with humans, they are obviously at best a model; they are not humans. They may differ substantially in some respects, while they may be similar in others. Crossing the species border is always a risk in translational research.

Off-target effects are another problem. Often, novel therapeutic targets are identified in very distinct experimental settings. For instance, cyclo-oxygenase-2 inhibitors were developed to reduce bleeding and increases in blood pressure associated with the use of non-steroidal anti-inflammatory drugs. Although those aims were partly achieved, rofecoxib was associated with an increased incidence of myocardial infarction, possibly due to prothrombotic effects, while celecoxib is still being tested in the large PRECISION trial. With an ever older population being treated in cardiovascular medicine, comorbidities and drugs used to treat those, cardiovascular safety becomes an increasing issue that needs to be considered in any drug-development programme.

Another example is provided by drugs that raise high-density lipoprotein cholesterol (HDL-C) (Lüscher et al. in press). Epidemiologically, it appeared obvious that raising HDL-C would reduce cardiovascular events in patients at risk. The cholesterol ester transport protein inhibitors therefore raised big hopes and indeed, the first in its class, torcetrapib, more than doubled HDL-C levels, but unexpectedly increased mortality. Later, basic research showed that torcetrapib increased aldosterone release from the adrenal glands and endothelin release from the vasculature, while suppressing endothelial nitric oxide synthase expression and endothelial function. These effects were considered off-target, because other molecules of the same class, such as dalcetrapib, did not share these properties. In phase II studies, such as Dol-Vessel, dalcetrapib increased HDL-C by 30% in patients with hyperlipidaemia and low HDL-C, while leaving blood pressure unchanged. However, dalcetrapib did not improve endothelial dysfunction or suppress markers of inflammation. In line with this, dalcetrapib was ineffective in patients after acute coronary syndromes in the large Dol-Outcomes trial. Thus, it is likely that HDL-C dysfunction in patients with acute coronary syndromes may explain the neutral results, in spite of a marked rise in the lipoprotein. Indeed, another HDL-C-raising drug, i.e. niacin, proved equally ineffective in two large outcome trials. Thus, HDL-C may be a marker rather than a therapeutic target, unless a drug also improves HDL-C dysfunction and protein composition in patients with acute coronary syndromes or coronary artery disease, or the HDL-C paradigm is wrong altogether.

Therefore, surrogate endpoints are an important need (Table 1). While blood pressure and low-density lipoprotein cholesterol are accepted surrogate endpoints predicting risk, and when lowered pharmacologically lead to a reduced risk, the record of other surrogates is less convincing. For instance, in the 1970s, ventricular ectopic beats were considered ideal surrogates for patients at risk of sudden cardiac death, until the CAST trial and the SWORD trial, using antiarrhythmic drugs, such as encaïnide, flecaïnide, moricizïn, or d-sotalol, respectively found an increased rather than decreased mortality, in spite of effective suppression of ventricular ectopic beats.

Likewise, in heart failure, exercise performance, haemodynamic improvements, and left ventricular ejection fraction have been used with disappointing results. Indeed, many drugs, such as inotropes or phosphodiesterase inhibitors, improved haemodynamics and exercise preformance, but increased mortality. Thus, the paradigm of stimulating the heart was falsified, while the concept of unloading it proved effective. It appears that only a reduction of left ventricular remodelling and of brain natriuretic peptide have some predictive value for clinical outcome in this patient population. Indeed, the lack of remodelling with endothelin antagonists predicted negative outcome studies.30

Figure 3 Levels of drug-development programmes.
Why many research findings prove to be false

Thus, as predicted by Karl Popper, most research findings are eventually falsified at different levels of research. While some hypotheses and paradigms survive the entire process, others have to be dismissed at initial or later stages. The reasons are multiple, but may include the following: (i) inappropriate experimental models (i.e. cellular systems, animals); (ii) irreproducible findings (i.e. overstated, selected or fraudulent data, large number of tested relationships, 'hot' scientific field); (iii) study design (i.e. comparator groups, small sample size, wrong study population, extended flexibility in definitions of outcomes); (iv) small effect size; and (v) overwhelming intellectual or financial interests. As outlined above, many findings in cellular systems and animal models cannot be reproduced at the clinical level, because they are limited to specific experimental conditions, due to species differences or inappropriate modelling of human disease. Currently, for convenience, costs, and legal as well as regulatory constraints, mainly rodents are used, although pigs and primates are closer to humans in many respects. Biological systems in animal models should be more carefully evaluated regarding their similarity to human biology. Possibly, humanized mice may be helpful. Furthermore, not all research findings are reproducible, because they may require very specific experimental settings, are seen only in certain, but not other cell lines or animal strains, or have been overstated due to the enthusiasm of the investigators. Not uncommonly, parts of the results are not presented in the published manuscript in...
order to pass the rigid peer review process, or certain experimental data are even suppressed.\textsuperscript{55} Of note, an increasing number of manuscripts have had to be retracted after publication, particularly in high-impact journals.\textsuperscript{56}

In response to that unfortunate trend, C. Glenn Bayley recently published the six red flags to test scientific findings.\textsuperscript{57} First, are experiments performed blinded? Second, were basic experiments appropriately repeated? Third, were all results presented? Fourth, were there positive and negative controls? Fifth, were reagents validated? Appropriately repeated? Third, were all results presented? Fourth, were statistical tests appropriate? It is obvious that there positive and negative controls? Fifth, were reagents validated?

At the clinical level, the study design is particularly important. In general, non-randomized and smaller randomized studies are more commonly refuted by later research.\textsuperscript{51} For instance, an initial case–control study involving 1334 patients and controls suggested that dimethyl sulfoxide, a commonly used solvent, has profound biological effects.\textsuperscript{58} Statistics of all seriously considered papers are currently checked by specialized editors in all high-impact journals, including the European Heart Journal.\textsuperscript{59,60} Finally, in ‘hot’ scientific fields there is a clear danger of publishing too quickly and too enthusiastically, as again stressed by fraud scandals affecting stem cell research.\textsuperscript{61}

The role of scientific journals

It is the aim of the peer review process common to all respected scientific journals to assess research findings critically with regard to their validity, importance, and novelty.\textsuperscript{76,77} Great care has to be taken in evaluating the design, methodology, and data analysis of a given manuscript in order to assure that the data are valid and eventually reproducible. Although less than perfect, the peer review process, particularly when involving three reviewers and knowledgeable editors, may pick up flaws and provide advice on how to improve the manuscript. Nevertheless, the process cannot completely avoid the possibility that some of the published papers are irreproducible or even have to be retracted.\textsuperscript{21,56} Increasingly, journal editorial offices receive allegations from other authors or whistleblowers on the validity or reproducibility of findings. To address this issue, the ESC Journal Family has initiated an independent Ethics Board, where such allegations will be handled.\textsuperscript{78}

There are conflicts of interest for editors also that may endanger a proper peer review and selection of manuscripts. In particular, papers reporting novel data of ‘hot’ areas may be accepted with lesser stringency. For instance, as with gene therapy research in the 1990s, stem cell research currently attracts a lot of attention; hence, even studies with minimal patient numbers are accepted.\textsuperscript{79} Furthermore, pressures from industry, either open or under cover, may affect editorial decisions.\textsuperscript{80} Thus, editors must be aware of these potential biases.

Thus, in summary, the road to evidence is long and winding indeed. As the evidence base of clinical medicine has grown, the process has become even longer and bumpier; as event rates have dropped and the most obvious facts have been discovered, it has become increasingly difficult to demonstrate incremental novelty and/or benefit beyond what has already been achieved. This may explain the increasing number of neutral trials in today’s cardiovascular research. However, the lessons learned from the past may be helpful for discovery programmes of the future.

References

47. McMurray JJV, Adamopoulos S, Anker SD, Aurichio A, Bohm M, Dickstein K, Falk V, Filippatos G, Fonseca C, Gomez-Sanchez MA, Jaarsma T, Kober L, Lip GYH, Magni... 1735–1743.

48. Gardn... 1847.


50. Parker M, on behalf of the ENABLE investigators. Effects of the endothelin receptor-antagonist bosentan on the morbidity and mortality in patients with advanced heart failure. Results of the ENABLE 1 and 2 Trial Program. Presented at the American College of Cardiology 2002.


