Tidsskriftet
DEN NORSKE LEGEFORENING

Knowing the numbers or knowing why?

MEDICINE AND NUMBERS

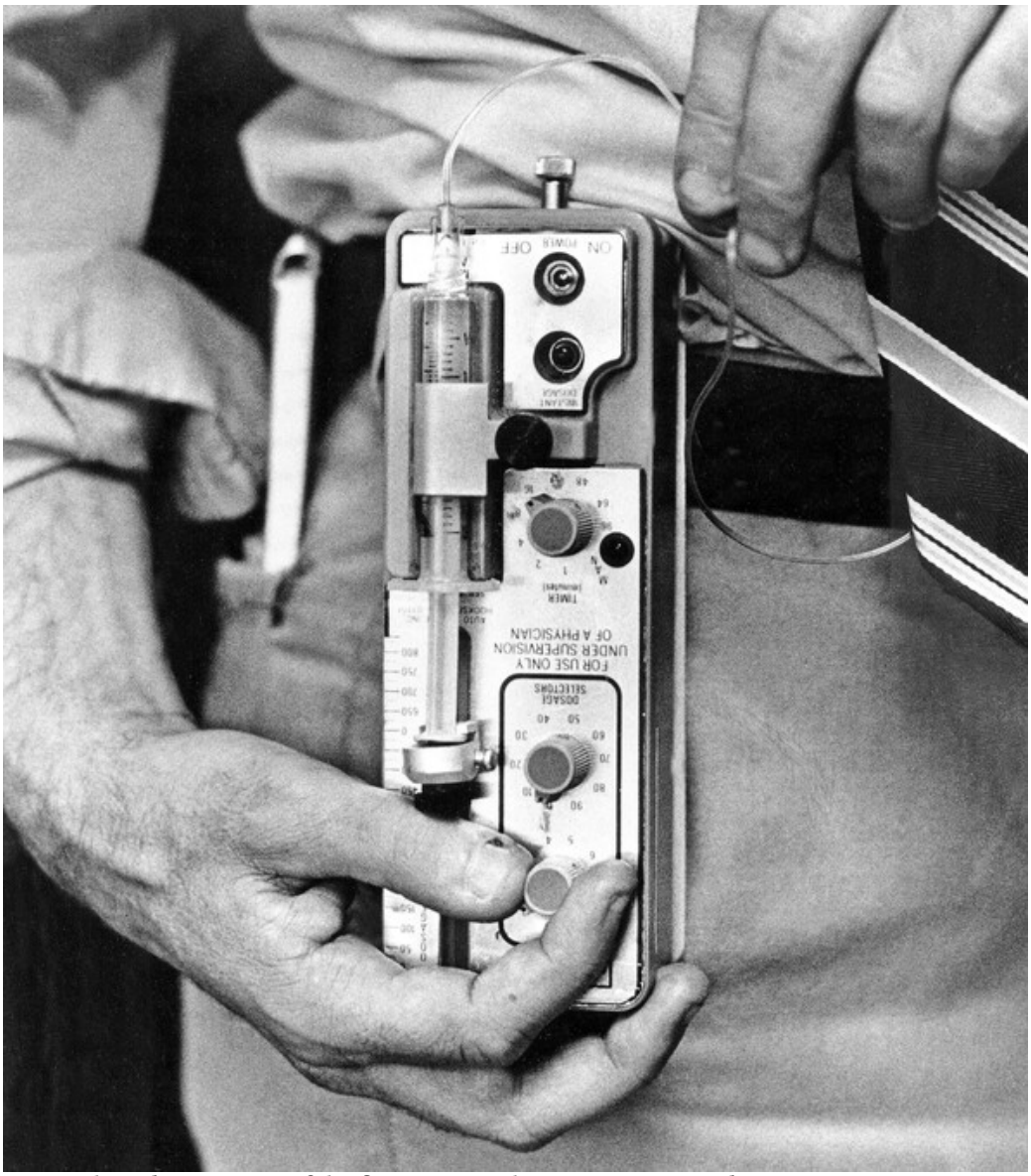
KATHRINE FREY FRØSLIE

Kathrine Frey Frøslie (born 1971) is a statistician at the Norwegian National Advisory Unit on Women's Health, Oslo University Hospital, Rikshospitalet. She has a PhD in biostatistics, is an experienced lecturer and research advisor, and runs the popular science knitting blog statistrikk.no. The author has completed the ICMJE form and reports no conflicts of interest.

JO RØISLIEN

E-mail: jo@joroislien.no
Jo Røislien (born 1972) is professor of medical statistics at the Faculty of Health Sciences, University of Stavanger, and adjunct associate professor at the Department of Mathematical Sciences, Norwegian University of Science and Technology. He is a renowned science communicator and science TV presenter.
The author has completed the ICMJE form and reports no conflicts of interest.

Only a century ago, an early and painful death was predicted for those experiencing excessive urination, and urine with a sweet odour or taste. This was before any mechanistic understanding of diabetes was established. Diabetes was a predictor for death, but little could be done to prevent either. Not enough was known about the disease.



Practical prediction: One of the first ever insulin pumps in everyday use in 1986. A mechanical pump knows nothing about the pathophysiology of diabetes, but based on measurements of the subcutaneous intracellular fluid, it can still inject correct doses of insulin,. Photo: Associated Press/NTB scanpix.

A main aim of statistical analysis is to uncover associations between variables, and one of the most versatile tools in the statistical toolbox is regression analysis. Regression analysis can help us establish whether one variable can be used to predict another. Sugar in the urine predicts death, the weather today predicts the weather tomorrow, and postal codes predict school test performances.

Sweet urine: A predictor

A statistical prediction model is democratic and inclusive. Any variable may be a predictor: urine odour, weather, postal codes – anything. And the accompanying mathematical equation may have any level of complexity. The only thing that matters is whether we are able to predict the future based on the information at hand.

Big data, data mining and machine learning are all about trawling large amounts of data for statistical associations, and then using these associations to provide tailored ads on your Facebook page, or to determine which genetic subgroup will benefit from a new, but costly, cancer medicine. It has an aura of science fiction about it, but often it is merely ordinary regression analysis.

Prediction models are useful, even without knowledge of the underlying mechanisms. A woman giving birth to a baby weighing more than 4500 g will be classified as having high risk of subsequent type 2 diabetes, simply because the prevalence of type 2 diabetes is high among women who give birth to large babies. This is useful knowledge, as it tells the woman she should check her blood glucose levels more often than she might normally do.

But giving birth to a large baby does not *cause* subsequent type 2 diabetes. There is an association, not a causal relation. Without causal knowledge we usually cannot intervene, and we cannot inform the woman on how to actually avoid developing type 2 diabetes. If this is our aim, we need to know something about the physiological mechanisms of diabetes.

An explanatory variable must be explicable

The turning point for diabetes was when the role of insulin and the pancreas was discovered in the early 1900s. Armed with the new knowledge of the importance of insulin regulation, doctors could now intervene and thereby influence survival. "Good or poor insulin regulation" is more than a mere predictor for death, the way "sweet urine" is. It is an *explanatory variable* for death, because it holds information about the physiological mechanisms involved in diabetes.

In 1954, the Danish doctor Jørgen Pedersen (1914–78) suggested that diabetic mothers give birth to large babies due to increased transfer of glucose from the pregnant mother to the foetus (1). In 2008, the Pedersen hypothesis was extended to include non-diabetic mothers, when a regression analysis found an almost linear association between the mother's blood glucose level and the child's birth weight (2). Thanks to meticulous research, we now know that the associations between maternal blood glucose levels and pregnancy outcomes like macrosomia and neonatal hypoglycaemia are not merely correlations; they are causal relations.

When chasing causal relations between the mother's blood glucose levels and adverse birth outcomes, we must use all our physiological and clinical expertise, and evaluate which variables to include in the accompanying regression models.

In contrast to the case of a prediction models, we cannot choose freely what variables to include in regression models used for the estimation of mechanistic effects and causal relations. Whereas a predictor may be any quantitative trait, explanatory variables must be part of the presumed causal chain, and we must specify main exposure, confounders, mediators and colliders. Also, unlike in the case of a prediction model, the mathematical equation should be fairly simple. There is little use in an inexplicable explanatory model.

To predict is not to explain

Regression analysis is a versatile statistical tool: so versatile it can be applied to fundamentally different analytical tasks. Whether the aim of the statistical analysis is to predict an outcome or to estimate presumed causal relations, we apply the same statistical regression models. It is only the words we use that inform the readers – and ourselves – how far up the ladder of knowledge we are.

Medical researchers tend to use the word *predictor* when referring to any independent variable in a statistical regression model. But this choice of words blurs the distinction between two fundamentally different approaches to medicine and health. There is a difference between an association and a causal relation. Knowing the numbers is not the same as knowing why. To predict is not to explain.

LITERATURE

1. Pedersen J. Weight and length at birth of infants of diabetic mothers. Acta Endocrinol (Copenh) 1954; 16: 330 - 42. [PubMed]

2. Metzger BE, Lowe LP, Dyer AR et al. Hyperglycemia and adverse pregnancy outcomes. N Engl J Med 2008; 358: 1991 - 2002. [PubMed][CrossRef]

Publisert: 28 May 2018. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.18.0086
Received 23.1.2018, accepted 19.3.2018.