Modern Mathematics for Modern Cancer Medicine

Prof ACC Coolen
Institute for Mathematical and Molecular Biomedicine, King's College London
& London Institute for Mathematical Sciences

ABSTRACT

The genomic revolution holds the promise to improve the life expectancy and quality of life for cancer patients via *personalised* cancer therapies. However, we need more advanced mathematical methods than those routinely used in cancer research. 'Next generation' data are still being analysed with 'previous generation' methods, which are insufficient to the task. We aim to use the extensive medical data that are now available for personalised cancer medicine by developing advanced quantitative techniques, and applying them to cancer data analysis, outcome and treatment response prediction, therapy innovation, and the design of clinical trials.

CONTENTS

1. INTRODUCTION

Modern cancer research. The nature of cancer research is changing profoundly. The quality and quantity of the biomedical data now available, or which soon will be available, has made personalised cancer medicine, i.e. the systematic use of information about an individual patient to select or optimize that patient's therapeutic care, a realistic ambition. The question now is how this ambition can be realised as quickly and effectively as possible. Unfortunately, mathematical methods for biomedicine have not kept pace with recent experimental advances in the biomedical sciences. Expensive 'next generation' genomic data are still analysed with 'previous generation' methods, causing non-reproducible therapeutic claims, modest ability to predict individual treatment response, and low success rates of medical trials.

To extend and save lives of cancer patients, improve their quality of life by preventing ineffective treatments, and speed up the generation of new intelligent and personalised therapies, we need more advanced mathematical tools that are designed and able to fully exploit the complex biomedical data that are now available.

False leads and spurious patterns. Medical trials are organised in phases. In a Phase 1 trial a new drug is tested in a small group of people (20-80) to evaluate safety, dosage ranges, and side effects. In a Phase 2 trial one evaluates its effectiveness in a larger group (100-300). Finally, in a Phase 3 trial it is tested in large groups (1,000-3,000) to evaluate its performance compared to the present standard. Many new cancer drugs are entered into this trial process. Only about 15% make it to Phase 3, and of these only some 40% show a statistically significant result. Hence just 6% lead to a useful drug. The decision to launch a trial is always prompted by prior evidence for the clinical value of the proposed drug, so the problem must lie with the quality or the analysis of this evidence.

Similarly, with the increased availability of affordable tests for genes and their expression levels (i.e. the extent to which they are switched on) many research groups have tried to identify gene expression patterns that are predictive of good or poor cancer outcomes, of response to specific treatments, or of metastasis sites. Unfortunately, many studies reported successful tests of novel signatures that (often years later) cannot be reproduced when tested on new patient groups and in different medical centres. Also here the problem must lie with the quality or the analysis of prior data.

Biomedical complexity. Cells are extremely complex. They can change their functionality at different levels and on different timescales. They can modify which parts of their genetically stored 'program' are executed, and they can even modify the program itself.

A greater understanding of this complex biochemical hardware is vital for mapping the cascade of events that leads cells to become cancerous. It will allow for the identification of new therapeutic targets in cancer, and for identifying the mechanisms that cause acquired resistance to initially effective treatments. Similarly, an improved understanding of the human immune system is needed to understand, and subsequently undermine, a cancer cell's ability to avoid elimination by the host's immune cells.

The limited progress that has been made so far in tackling these problems is quantitative in origin: intervening intelligently in complex heterogeneous biological systems with millions of interacting variables requires advanced mathematical modelling tools.

Example projects and potential benefit. The problems with false leads in cancer research, which waste resources and stifle progress, are caused by data analysis and data quality issues. Research aimed at improving data analysis methods in cancer medicine is relatively inexpensive. It imposes no extra burden on cancer patients, and we know already

quite well which are the main hurdles: we need tools for handling disease and cohort heterogeneity, dimension mismatch, and overfitting. Only commitment stops us from the development of novel mathematical methodology with which to model complex heterogeneous biological systems with many interacting variables.

2. CURRENT RESEARCH PROJECTS

Below we describe three recently initiated research projects, which aim to develop new mathematical techniques for cancer research. This is followed by rough estimates of their expected benefits to cancer patients.

2A. Survival Analysis for Heterogeneous Cohorts

Prediction and population heterogeneity are at the heart of modern cancer medicine. Personalised therapy requires that we map and exploit heterogeneity: we aim to predict accurately from an individual's characteristics whether, when and where their disease is most likely to recur, and to which therapies they will respond.

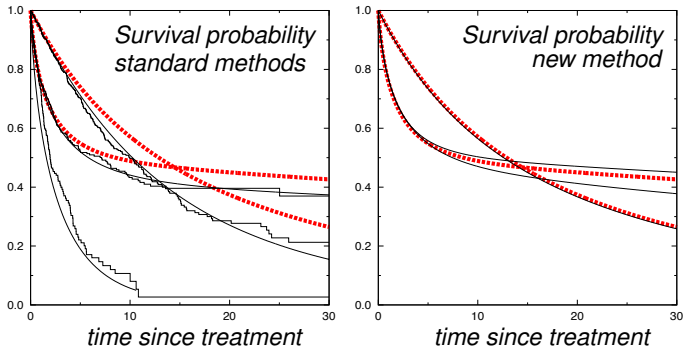
Survival analysis. Survival analysis is the branch of medical statistics that aims to find and quantify patterns that relate quantities that we can measure in a patient (the 'covariates') to the patient's clinical outcome, where outcome is measured as time to a specified event. In cancer studies this event is usually disease recurrence. A typical data set would describe a cohort of patients who were followed for several years, where for each we know e.g. tumour grade, size and histological type, lymph node status, age, hormone levels, and other markers obtained from blood samples; all collected at the time of first treatment. In addition we are given for each patient the time between their initial cancer treatment and either the point of relapse, or the point when they were last seen and found disease-free. Given these retrospective data, can we predict for *future* patients the time to relapse from their covariates?

Standard methods. Most survival data in medical research are still being analysed with just two (old) methods: Kaplan-Meier estimators (1958) and Cox regression (1972). Both were designed for a time when computing power was minimal, and they consequently involve various simplifying assumptions. They cannot probe heterogeneity of cohorts beyond that in the patient characteristics that we happen to measure. More recent methods, the 'frailty' or 'random effects' models, make assumptions similar to Cox but include some heterogeneity by allowing for the effects of informative variables that were not measured. In addition, oncologists have automated decision support tools, such as Adjuvant! Online (www.adjuvantonline.com), that predict disease outcome from a patient's covariates. It is not always easy to obtain full details on their methods, but judged from one of the more recent tools, they are still based on Cox, frailty or random effects models.

Heterogeneity and competing risks. Only a subset of patients will generally benefit from any cancer therapy. In some cases we know what distinguishes this subset (e.g. with Herceptin), but in most cases we do not. If systematically only 10% of a group of cancer patients with similar covariates benefit from a given chemotherapy regime, then this group is heterogeneous. Unless and until we can understand this heterogeneity and identify the 10% *a priori*, we condemn 90% to treatment that is not only without benefit to them, but does further damage.

A further problem is that in heterogeneous cohorts one has to worry about the impact of other diseases (the 'competing risks'), whose incidence may correlate with the cancer under study. Especially in cohorts with many older patients, as with prostate cancer, individuals will be lost to observation because of competing risks and this distorts patterns in survival data. Virtually all existing methods assume for simplicity that different risks are statisticaly independent. If this assumption is violated, their predictions can be completely wrong, see e.g. the figure below.

The rationale for this project. Cohort heterogeneity is the norm, not the exception. No two patients are identical, even if their tumour sizes/grades and blood serum levels are. We know also that cancer is a heterogeneous disease, and even individual tumours are histologically heterogeneous. Personalised cancer medicine requires that these differences between patients and tumours can be identified and used to tailor treatment protocols. We want to develop a new generation of survival analysis tools, that (i) generate precise survival predictions even for heterogeneous cohorts with competing risks, so that we can give patients reliable information and target aggressive treatments at those who need it most, and (ii) are able to detect, quantify and map the heterogeneity of a cohort from survival data alone, in order to guide the discovery of new covariates that may allow us to tailor individualised treatments more accurately to patients.



Left: survival predictions of standard methods (Kaplan-Meier estimators and Cox regression, black curves), for three patient groups. The red curves give the true probabilities. Right: predictions of our new method (black curves).

Preliminary results. In collaboration with the cancer epidemiology team of Prof Holmberg at King's College, we have started developing novel mathematical methodology for handling disease or patient heterogeneity that is not captured by covariates. It is found to give significantly

more accurate survival curves on synthetic data, compared to conventional methods, see the figure above. Preliminary application of our new methods to prostate cancer data led to plausible new explanations for previous counter-intuitive inferences of standard methods. The results were recently presented at an international biometric conference in Stockholm (June 2013), and well received. A manuscript is under review for publication in the Journal of the Royal Statistical Society B. We are now applying the method to breast cancer data, with the objective to predict heterogeneous metastasis sites from patient characteristics.

2B. Dimension Reduction for Genetic Signals

To use genomic information effectively in cancer medicine we must remedy the imbalance between the small number of patients in typical data sets and the enormous number of genetic variables that we can measure for each. The genetic variables are too numerous to make regression and survival analysis computationally feasible, and the imbalance causes 'overfitting', i.e. detection of spurious patterns. We need rigorous quantitative methods that can reduce the dimensionality of genetic signals without loss of information.

Overfitting in genome data analysis. Imagine we have data on four cancer patients, two of whom respond to treatment (group A) and two of whom do not (group B). We investigate 50 genes in their tumour samples, and measure for each patient whether these are switched on (1) or off (0):

A 10010100101001010101001000101011100100100100100100
A 01000100001010100101010101001010100011110010100100
B 00101000111010110110010010011100111001010010101010
B 10101100101011001010010011110010010110011101011101

We then look for patterns in this information that would have enabled us to predict whether a patient will be in group A or B. These are called 'gene signatures'. A signature gene takes the same values for all patients in group A and a different value for those in group B. Here, for instance, we find the signature genes shown in red:

A 10010100101001010101001000101011100100100100100100
A 01000100001010100101010101001010100011110010100100
B 00101000111010110110010010011100111001010010101010
B 10101100101011001010010011110010010110011101011101

However, the above data are generated randomly; the 'patterns' we see are just random accidents. If we would mix up the patient's group labels, we would find a new set of 'signature genes' that appear to predict outcome:

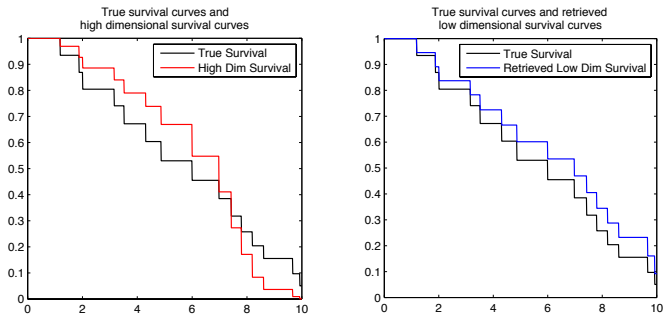
A 10010100101001010101001000101011100100100100100100
B 01000100001010100101010101001010100011110010100100
A 00101000111010110110010010011100111001010010101010
B 10101100101011001010010011110010010110011101011101

The more variables we measure, the more spurious patterns we will find. How do we distinguish between real and spurious patterns? This is the overfitting problem.

Consequences of overfitting. The numbers of genes and patients in the above example may not be realistic, but their ratio is. In genomic data sets we tend to have in the order of 100-1000 patients, and for each patient we can measure more than 10,000 gene expression values, increasing to more then 1,000,000 in next generation sequencing (NGS). Overfitting is a major contributing factor to the modest reproducibility of gene signatures and the low success rates of medical trials. There is no theory yet even for the simple Cox method to tell us how many patients we need for a given signal dimension to avoid overfitting, let alone for more advanced methods. Measuring vastly more patients is not an option. Apart from the impossibility to recruit a million patients with the same cancer, NGS analysis costs around 1 K£ per patient. Most researchers resort to simpe gene-by-gene testing for correlations with outcome, and correct results for multiple testing.

Routes for dimension reduction. All methods for dimension reduction of genetic signals must avoid using clinical outcome information, as this could lead to implicit overfitting. This leaves two approaches. The first is to use knowledge from cellular biology. One may try to identify small groups of genes that are part of the same functional modules of a cell, and represent the activity of each module by a single 'representative' gene. The problem is that the biological data on gene interactions and modules are unreliable, and affected by experimental bias. The second approach is to use regularities in the gene signals themselves (e.g. correlations) to achieve dimension reduction. However, there is no guarantee that the main regularities actually carry the information on cancer outcome that we seek to extract, and the commonly used methods are limited to linear relations among genes. There are no accepted standards yet, and most methods are still ad hoc.

Rationale for this project. We must ensure that genetic information can be used optimally in prediction and survival analysis, by developing improved mathematical and computational methodology for dimension reduction. Within the approach based on using biological information we must develop systematic methods to decontaminate interaction data for experimental bias, and precise methods for the identification of modules and the optimal characterisation of module activities by a single 'meta-gene'. Within the approach based on using signal regularities we want to develop a more systematic and probabilistically rigorous protocol based on 'latent variable' representations of the data. Ultimately, these two routes should be integrated.



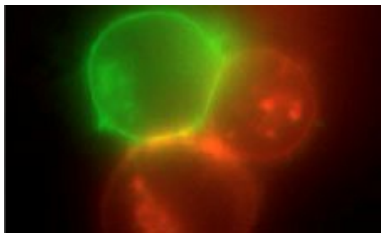
Left: ordinary survival curves (red), for synthetic data. Right: improved survival curves calculated following dimension reduction (blue). Both are compared to the true values (black).

Preliminary results. We are developing methodology for Bayes-optimal dimension reduction of genetic signals, using nonlinear latent variable mappings; see figure below. A paper has been submitted to the Journal of Machine Learning Research. Another study aims to predict analytically for proportional hazards regression the number of genetic variables relative to the number of patients at which overfitting sets in, for large data sets, building on mathematical methodology from theoretical physics.

2C. Cancer-Specific Immune System Reprogramming

The adaptive immune system is a powerful natural defense mechanism, which we cannot yet exploit therapeutically in cancer medicine because we do not understand properly how it works. We wish to explore, building on new developments in mathematical modelling, whether we can achieve a level of quantitative understanding that allows us to reprogram the immune system into hunting down and killing a patient's presently tolerated cancer cells.

Immune system and cancer. There is overwhelming evidence for the major role played by the immune system in deciding cancer outcome. It is normally effective in removing malignant cells early; the vast majority do not develop into cancer. For instance, when the immune systems of organ transplant recipients are suppressed to prevent rejection, this causes an up to 80-fold risk increase for certain skin cancers. In ER-negative breast cancer one finds that genes related to immune response provide important prognostic information. The modern 'immuno-editing' picture of oncogenesis is a three-stage process: elimination (the healthy state where immune cells detect and eliminate malignant cells), followed by equilibrium (coexistence of small tumours held in check by immune cells), and escape. In the escape stage the tumour has developed a successful strategy to evade the host's immune system, like viruses, by 'hiding' from immune cells, by disabling them, or by interfering with and even hijacking their control signals.

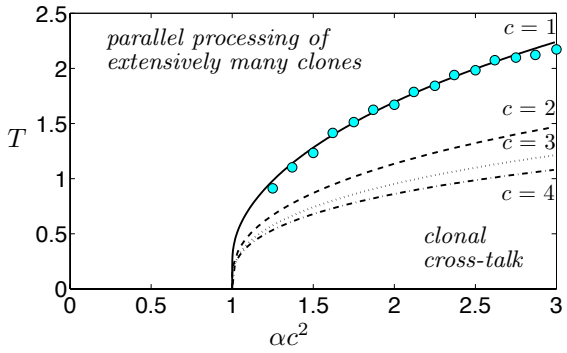


An immune cell (Natural Killer cell, in green) in the process of attacking two breast cancer cells (in red). Image recorded in Prof Tony Ng's Lab (King's College London).

Present approaches in cancer medicine. We know that the immune system can kill unwanted cell types very selectively, in principle with devastating effectiveness, but that it falters in the case of cancer. Since moreover it can reach nearly all parts of the body, it is an obvious vehicle for cancer therapies. At present, most immune-related strategies in cancer medicine are focusing on the development of cancer vaccines (which will also help trigger responses in existing patients), and on the monitoring of immune-related genomic information for tumour classification and clinical outcome prediction. There are some cell-based immunostimulation approaches that seek to exploit a patient's immune system to tackle cancer.

Controlling the immune system. The immune system has been studied for decades, yet our understanding of how its components orchestrate its function is still limited, partly because in the past not all its parts were known and partly because we did not have the necessary mathematical and computational tools. This has prevented us from exploiting the power of the immune system therapeutically. Most quantitative studies of tumour-immune interaction were limited to simple 'predator-prey' models for cell concentrations. This situation is changing. Recent years witnessed a surge of interest in mathematical models of the immune system that build on the theory of heterogeneous many-variable systems in physics. These models are mathematically similar to models of neural information processing systems, which have been studied extensively in the past, which suggests that we may be able to achieve a quantitative understanding of the immune system's complexities.

The rationale for this project. The immune system is a remarkably powerful and versatile defense weapon. We know that it is able to attack selectively specific cell types of the host, which occurs in organ-specific autoimmune diseases, and that it can be triggered by environmental or genetic factors into switching from selective tolerance to selective intolerance. We now want to answer this question: is it possible, using mathematical models of the immune system's signalling processes and the interaction between tumours and the immune system, and of the latter's ability to switch between selective tolerance and selective intolerance, to design *feasible and safe* interventions that (for a subset of cancers) reprogram the immune system into targeting a presently tolerated tumour? Or, for breast cancer patients who had a double mastectomy, can we trick the immune system into developing *breast-cell-specific* autoimmune disease, i.e. into hunting down and killing all residual breast cells of the type that had developed the tumour?



Clonal interference transition lines, obtained via statistical mechanical modelling. The relevant system parameters are the overall immune repertoire (measured by α), the average level c of B-cell promiscuity, and the T-cell stochasticity T.

Preliminary results. In collaboration with a team in Rome, we are developing solvable mathematical models of immune signalling between B- and T-lymphocytes. Two papers are under review. We seek to predict the performance of the immune system under different conditions, understand how and where it may fail, and how it can be reprogrammed.

3. ESTIMATES OF PATIENT BENEFIT

Population statistics and cancer incidence data. The present population of the UK is about 61M. The world population is nearly 7000M, divided roughly as 4230M (Asia), 1050M (Africa), 940M (Americas), 740M (Europe), and 40M (Oceania). The number of these with (limited) access to medical cancer care in five years from now, on which cancer treatment innovation can therefore in principle have an impact, would be somewhere in the range 1500M–2500M.

The estimates below are calculated for breast cancer, as an explicit example. One expects similar benefits for other prevalent cancer types. The most recent reliable publicly available figures for breast cancer (BC) incidence and survival in the UK are (source: Cancer Research UK):

- About 49,000 new BC cases per annum
- Overall BC survival rates: 85% beyond 5 years, 77% beyond 10 years
- Breast cancer incidence in Europe, USA, Australia and New Zealand: about 80 per 100,000

Globally there are about 1.4M new breast cancer cases per annum.

Estimating patient benefit of quantitative methods. Estimates of the possible benefits of quantitative innovation will at best represent orders of magnitude, due to the absence of existing precise statistical methodology and survival projections, the intransparancy of commonly used decision support software (e.g. Adjuvant Online publishes neither the details of its regression method, nor the characteristics of the data to which the regression was applied), and the variability of oncologists' treatment decision protocols at the various cancer centres (which are mostly based on local expert judgement).

In our extrapolations below we used the above UK breast cancer incidence rates, and used the value 60M for the UK's population size, and 1700M for the size of the world's population with access to cancer treatment. Although the breast cancer incidence rate in underdeveloped countries is significantly lower than that of the UK (and other developed countries), in our judgement the UK incidence rate should be used since (i) we extrapolate only to the fraction of the population with access to cancer treatment (i.e. the relatively wealthy fraction) and (ii) we expect most international BC incidence rate differences to be competing risk effects, rather than true risk differences, and (iii) the figure 1700M together with an annual incidence rate of 80/100,000 is consistent with the recorded number of about 1.4M global BC cases per annum.

Improved quantitative methods in cancer medicine can benefit cancer patients via increased effectiveness of personalised treatment, improved quality of life, generation of new therapies, and reduction of the time between therapy discovery and roll-out to clinical practice.

Treatment effectiveness and increased quality of life. Cancer treatment decisions are based on the likelihood of a patient with certain characteristics benefiting from a proposed treatment. Oncologists estimate the probability of response (possibly with software support), usually by using as a proxy the fraction of all such patients that have responded in the past. Treatment is recommended if the response probability exceeds a given threshold. The value of this threshold is usually in the range $[0.05, 0.10]$ (i.e. treatment is recommended if the likelihood of benefit is 5%-10% or more). Treatments with more severe the side-effects would be given larger cut-off values, i.e. the evidence for benefit should be stronger. There are two possible decision errors: patients can be denied treatment which would have benefited them (type I errors), and patients can be given treatment without benefit (type II errors). Current methods for breast cancer outcome prediction (e.g. Mammaprint, based on gene profiles) give values of around 40% for the overall sum of type I and type II error rates.

Improved statistical methods will lead to more precise formulae for the response probabilities, via more informative and individualised patient characterisations and more accurately extracted relationships between patient characteristics and probability of response. Those predicted to respond will thereby have a significantly smaller probability of failing to benefit, and those predicted to be non-responders will thereby have a significantly smaller probability of being denied benefit.

If we estimate conservatively that a fraction 1/10 of those patients that are wrongly denied treatment would not survive as a consequence of this error, and if our quantitative innovations were to be accepted as the new standard in treatment response prediction, this would for breast cancer lead approximately to

target:	reduction of sum of type I and type II response prediction errors by 10%
	e.g. Type I error ↓8%, Type II error ↓2%:
UK impact: (per annum)	prevents 4,000 ineffective treatments prevents 1,000 wrongly denied treatments prevents 100 avoidable deaths
globally: (per anum)	prevents 110,000 ineffective treatments prevents 28,000 wrongly denied treatments prevents 2,800 avoidable deaths

New targeted treatment regimes. The efficacy of new cancer treatments can obviously not be judged prior to their discovery. However, one can make order of magnitude estimates. We assume the new treatment not to generate significant indirect excess mortality. If the treatment is based on properly quantified cellular or genomic characteristics, then we will be able to predict with reasonable accuracy *which* patients are in the group of responders. A conservative estimate can then be obtained by making the following (reasonable) assumptions for breast cancer:

(i) 25% of all individuals presently diagnosed are not cured by current treatments.
(ii) 10% of all patients are predicted by us to possibly benefit from the new treatment.
(iii) half of those predicted to benefit indeed do so.
(iv) in the group of those who benefit, 75% would have also survived with standard treatment.
(v) in the group of those who benefit, of those 25% that would not have survived:
90% extend their disease-free survival (DFS) on average by one year,
10% survive as a consequence of the new treatment

This implies, in combination, that as a consequence of the new treatment: 1.1% of all breast cancer patients (from the group that wil not be cured) extend their DFS by one year, and 0.13% of all patients now survive but would not have done so with current treatments. Hence

target:	1.1% of patients get 1-year DFS extension 0.13% of patients prevent BC death
UK impact: (per annum)	560 added years of disease-free survival, 63 prevented breast cancer deaths
globally: (per annum)	15,000 added years of disease-free survival, 1,800 prevented breast cancer deaths

New systemic treatment regimes. New systemic treatment regimes that do not target specific molecular pathways, such as generic immune therapies, can in principle benefit a larger group of patients. Again we assume the new treatment not to generate significant indirect excess mortality. A conservative estimate can then be obtained by making the following (reasonable) assumptions for breast cancer:

(i) 25% of all individuals presently diagnosed are not cured by current treatments.
(ii) 50% of all patients are judged to possibly benefit from the new treatment.
(iii) half of those predicted to benefit indeed do so.
(iv) in the group of those who benefit, 75% would have survived also with standard treatment.
(v) in the group of those who benefit, of those 25% that would not have survived:
90% extend their disease-free survival (DFS) on average by one year,
10% survive as a consequence of the new treatment,

This implies, in combination, that as a consequence of the new treatment: 5.5% of all breast cancer patients (from the group that wil not be cured) extend their DFS by one year, and 0.65% of all breast cancer patients now survive but would not have done so with current treatments. Hence

target:	5.5% of patients get 1-year DFS extension 0.65% of BC prevent BC death
UK impact: (per annum)	2700 added years of disease-free survival, 320 prevented BC deaths
globally: (per annum)	75,000 added years of disease-free survival, 9,000 prevented BC deaths

4. ABOUT THE AUTHOR

Ton Coolen obtained his PhD in theoretical physics from the University of Utrecht, came to Oxford in 1991, and joined King's College London in 1995, where in 2000 he was appointed to a chair in applied mathematics. He is a member of the UK's Engineering and Physical Sciences Research Council's peer review council, the Biotechnology and Biological Sciences Research Council's systems biology board, and a Fellow of the London Institute for Mathematical Sciences. He has published two books and 130+ refereed academic papers, and supervised 19 PhD students.

In recent years Coolen redirected his work fully towards mathematical innovation in support of cancer research, including biomarker analysis and cancer outcome prediction, application of graph theory and statistical methods to cellular signalling networks and immune networks, and survival analysis for heterogeneous cohorts. He has active research collaborations with oncologists, cancer epidemiologists and immunologists at King's College London.

Coolen served for several years as an elected member in The Royal Marsden NHS Foundation Trust Council, representing the constituency of cancer partients and their carers. He is now on a mission to raise the mathematical and statistical standards in cancer medicine, and promote the increased involvement of applied mathematicians in quantitative cancer research. He recently created and leads a new Institute for Mathematical and Molecular Biomedicine at King's College London, see www.qbio.kcl.ac.uk/IMMB, he gives regular seminars to cancer researchers on the hazards of sub-standard statistical methods in cancer data analysis, and is involved in the preparation of a further research Institute devoted to quantitative cancer research.