Gut microbiome, big data and machine learning to promote precision medicine for cancer

Giovanni Cammarota, Gianluca Ianiro, Anna Ahern, Carmine Carbone, Andriy Temko, Marcus J. Claesson, Antonio Gasbarrini and Giampaolo Tortora

Abstract | The gut microbiome has been implicated in cancer in several ways, as specific microbial signatures are known to promote cancer development and influence safety, tolerability and efficacy of therapies. The 'omics' technologies used for microbiome analysis continuously evolve and, although much of the research is still at an early stage, large-scale datasets of ever increasing size and complexity are being produced. However, there are varying levels of difficulty in realizing the full potential of these new tools, which limit our ability to critically analyse much of the available data. In this Perspective, we provide a brief overview on the role of gut microbiome in cancer and focus on the need, role and limitations of a machine learning-driven approach to analyse large amounts of complex health-care information in the era of big data. We also discuss the potential application of microbiome-based big data aimed at promoting precision medicine in cancer.

Over the past decade, the progress achieved in understanding microbial ecology has enabled us to focus on the crucial role of gut microbiota in maintaining human health by fulfilling important functions (including digestion, control of pathogens, immune regulation and production of beneficial metabolites)[1]. Similarly, imbalances in intestinal microbial composition and the consequent alterations of their functions have been variably associated with several clinical conditions, such as inflammatory bowel disease (IBD)[2], metabolic syndrome and obesity[3,4], antibiotic-associated diarrhoea[5], pregnancy[6,7], neurological disorders[8], cardiovascular disease[9] and cancer[10].

Microbiome analysis is evolving (BOX 1). To provide adequate robustness to data analysis, large-scale datasets of ever increasing size and complexity have been produced, but the complexity of managing and integrating multidimensional big data is a challenge, and these datasets currently do not enable any application that could influence the clinical course of diseases. Less than 5 or 6 years ago, bioinformatics software was principally useful to analyse and interpret the data generated from experiments carried out for a specific purpose. The early development of biological databases, such as the National Center for Biotechnology Information's GenBank and Gene Expression Omnibus, introduced some early predecessors of big data bioinformatics, but these databases were primarily designed to be repositories of completed experiments and their data. In the past few years, the concept of big data has revolutionized the application of bioinformatics to clinical and translational research. Currently, many active consortium projects are aimed at collecting data to generate a wide range of genome-related information (principally associated with cell lines, tissues and tumour samples)[11–13]. The purpose of these projects is to generate general reference datasets available for all scientists, with the chance to use these open reference data and explore specific hypotheses without conducting new experiments. This approach has the potential to reduce substantially the number of new experiments and to investigate complex working hypotheses on a large scale. The sharing of data on a large scale naturally requires new analytical tools, in addition to the formulation of specific experimental questions, the annotation and cleaning of open data and the performance of appropriate retrospective analysis by powerful software (such as artificial intelligence (AI)-based models and advanced machine learning (ML), which are particularly useful in translational medicine).

ML-driven analysis of gut microbiota could be particularly useful in oncology owing to the plethora of evidence relating the microbiome to cancer and the immense size of multiple integrated large-scale datasets generated from the individual omic systems (such as the genome, proteome, metabolome and transcriptome)[14], with the objective of developing and remodelling strategies aimed to prevent, diagnose and treat cancer. Data suggest that the gut microbiota can affect the natural history of malignancies, as specific microbial signatures are known to promote cancer development and influence the safety, tolerability or even the efficacy of cancer therapies[10]. The management of patients with cancer could, therefore, particularly benefit from addressing specific issues affected by the gut microbiota, including the prediction of drug response or toxicity and the discovery of regulatory microbial networks. There is, consequently, an increasing interest in profiling gut microbiota composition in these patients, as the resilience and stability of the gut microbiota and its responsiveness to physiological, environmental and pathological changes offer the opportunity to use the intestinal microbiota as a biomarker and/or therapeutic target. Nevertheless, this process would require the curation of extensive human databases and a careful correlation of different microbial signatures with multiple, complex laboratory and clinical parameters. The achievement of this goal depends largely on the availability of adequate computational resources that integrate large volumes of patient data (including demographic, clinical, family and lifestyle histories; comorbidity; radiological, histopathological and laboratory analyses) with 'omics' features and translate them into effective predictive models of clinical response[15]. ML and the use of data-driven sciences (BOX 2), dedicated to

Box 1 | **Microbiome analysis: technologies and related challenges**

Technologies

High-throughput technologies currently enable the sequencing and classification of the vast array of microorganisms that inhabit the human body down to the genus and species level (genomic sequencing of bacterial 16S ribosomal RNA (rRNA), shotgun sequencing, internal transcribed spacer and the IS-pro technique)[58,222,223], the determination of which microorganisms are metabolically active and which microbial genes are being actively expressed (metatranscriptomics)[224], as well as the analysis of proteins (metaproteomics)[225] and metabolites (metabolomics)[226].

Shotgun metagenomics versus 16S rRNA sequencing

Although more expensive and more cumbersome, shotgun metagenomics is now preferred over 16S rRNA sequencing for the analysis of gut microbiome[227]. 16S rRNA analysis is able to sequence only a single region of the bacterial genome, whereas shotgun metagenomics can analyse the whole genome, provide a more accurate detection of bacteria both at the species level and bacterial diversity, and study their functional potentials.

'Omics'

Omics technologies are used to characterize the microbial taxonomic bacterial community composition and encoded functions (such as metabolites), as well as to aggregate host–microbiome interactions to identify disease associations, biomarkers and novel diagnostic or therapeutic targets. Thus, studying the microbially derived metabolites is a promising strategy to link the outcomes of the disease to the functional state of the microbiome. This approach has started to provide functional insights into the aetiology of some pathologies[170,188] that go beyond the knowledge of the microbial ecosystem given by the taxonomic survey and also provide a more realistic and direct view of potential therapeutic targets.

Challenges

Although stool microbiota is the most commonly studied, the composition of the mucosa-associated microbiota seems to change markedly along the gastrointestinal tract of some subjects[228] and in turn can be different with the bacterial ecosystems from other sites (such as pancreatic cyst fluids[229]).

extracting knowledge from complex data, are expanding computational fields, the developments in which promise to cover these needs[16–19].

In this Perspective, we first provide a brief description of the role of intestinal bacteria in different types of cancer and then discuss the challenges associated with the need for an ML-driven approach for analysing large amounts of complex health-care information and to combine these into predictions for disease risk, diagnosis, prognosis and appropriate treatment.

Gut microbiota and cancer

The gut microbiota is composed of tens of trillions of microorganisms, including 1,000 different species of known bacteria, fungi, archaea, parasites and viruses with over three million genes[20,21]. Culture-based studies suggested that all healthy adults share most gut bacterial species, constituting a 'core microbiota'[21], but the human microbiota also displays a remarkable degree of variation within and between individuals that is determined by dietary habits, ethnicity, host genetics, age and drug usage[21–23].

Several physiological functions can be modulated by the gut microbiota, such as metabolism, inflammation and immunity[1,24]. In particular, the complex and intriguing relationship between the gut microbiota and cancer has been increasingly recognized as an important and timely topic, including the role of the microbiota as a potential biomarker of neoplasms as well its influence on the safety, tolerability and efficacy of cancer therapies[25–27]. In order to understand the context in which this Perspective focuses on big data and ML, we briefly summarize the link between cancer and the gut microbiome without discussing the mechanistic aspects of the relationship in detail.

Biomarker of neoplasms. Disruption of the inflammatory and immunological interaction between the host and the gut microbiota has been described in several malignancies (such as oral, lung, gastric, pancreatic and colorectal cancer (CRC)), and specific microbial signatures have been associated with different neoplasms, leading to the concept that the gut microbiota could act as both a tumour suppressor and a tumour promoter[28–30].

Helicobacter pylori was the first microorganism classified as a human group 1 carcinogen because of its specific link with gastric cancer[31,32]. This well-understood natural history now provides the clinicopathological rationale for primary and secondary gastric cancer prevention strategies, including the eradication of *H. pylori* infection[31]. Other studies demonstrated that certain bacteria promote tumour growth in both human and animal models of several cancer types such as colorectal (for example, *Fusobacterium* spp., *Bacteroides fragilis*, *Streptococcus gallolyticus*), pancreatic (for example, *Enterococcus faecalis*) and gall bladder (for example, *Salmonella* Typhi) cancer[33–36]. Similarly, *Fusobacterium nucleatum* and *Propionibacterium acnes* potentiate intestinal[37,38] and prostate epithelial[38] carcinogenesis, respectively, in both preclinical and clinical experimental settings. Interestingly, obesity-induced enrichment of the gut microbiota in *Clostridium* species might promote liver cancer in mice models and perhaps also in humans, although further studies are necessary to clarify the exact role in humans[39]. Moreover, the microbial biofilm (which is a higher order microbiota organization structure) can be considered an independent driver in the early stages of CRC carcinogenesis in humans[40]. Biofilm formation enables bacterial pathogens to colonize a wide variety of host niches and to persist in harsh environments, making their elimination by the immune system particularly difficult[40]. From this arises the concept that the biofilm promotes procarcinogenic activities that may underlie a direct mechanistic link with mucinous CRCs[41].

The gut microbiota has the potential to guide carcinogenesis through metabolites and molecules that influence both host epithelial and immune cell responses[42]. Thus, identified microbial species or their secreted molecules could serve as biomarkers of neoplasms. In this context, Yu et al. identified 20 microbial gene biomarkers from faecal samples of patients with CRC that were able to distinguish the presence of cancer[43], whereas Zackular et al. showed a statistically significant difference in the gut microbiota of patients with colon adenomas compared with that of healthy individuals as controls[44]. Other researchers identified a bacterial signature with the potential to predict the development of colorectal adenoma from healthy tissue and the progression from adenoma to carcinoma[45]. In other reports, the surface of intestinal and prostate tumours displayed higher levels of *Fusobacterium* spp. than the adjacent healthy tissue in patients[37,38,46]. Similarly, a specific microbial signature, including a decrease in α-diversity and in butyrate-producing bacteria, as well as an increase in lipopolysaccharide-producing bacteria, was able to distinguish patients with pancreatic cancer from healthy individuals as controls[47]. Moreover, in patients with hepatocellular carcinoma, 30 specific operational taxonomic unit markers were reported to have a statistically

significant diagnostic potential for early and advanced hepatocellular carcinoma[48].

Tolerability of therapies. Several cancer therapies (for example, chemotherapy, radiotherapy and immunotherapy) are known to produce clinically relevant adverse effects, most of which affect the gastrointestinal tract and are associated with an imbalance in the gut microbiota (dysbiosis). Thus, the identification of microbial signatures able to influence gastrointestinal drug toxicity or, in general, any adverse events could be clinically relevant to avoid the dosage reduction of chemotherapy.

Preclinical and clinical studies have shown that the gut microbiota is involved in the safety and tolerability of chemotherapeutics such as 5-fluorouracil, cyclophosphamide, irinotecan, oxaliplatin, gemcitabine and methotrexate[49]. The mechanisms of microbiota-associated drug toxicity can vary substantially depending on the treatment type and microbial species[50,51]. For example, the activation and metabolism of irinotecan, a drug widely used in the treatment of metastatic CRC, small-cell lung cancer and several other solid tumours, can cause severe diarrhoea[52,53]. Wallace et al. demonstrated that the presence in the gut of β-glucuronidase-positive bacteria leads to reactivation and subsequent release of the active SN-38 metabolite into the bowel lumen, generating diarrhoea in a mouse model[54]. In other studies, modulation of bacteria induced by antibiotics in mice suppressed irinotecan-dependent diarrhoea but also worsened methotrexate-dependent intestinal mucositis[55,56]. Other evidence suggests that platinum-based chemotherapy could cause the degeneration of intestinal mucosal cells, leading to microbial translocation into the blood circulation and mesenteric lymph nodes, resulting in septicaemia and systemic inflammation[57].

The gut microbiota seems to influence also the safety and tolerability of immune checkpoint inhibitors (ICIs)[58–63], and it has a pivotal role in radiation-induced bowel injury, as shown in mouse model studies[64–66].

Other studies involving patients with gynaecological, colorectal, anal and cervical malignancies confirmed that the gut microbiota is profoundly modified by radiotherapy and associated adverse effects, and that the baseline composition of the microbiota in patients can influence the occurrence and severity of adverse events (such as diarrhoea, mucositis and fatigue)[67–69]. In a study published in 2019, patients with radiation enteropathy had higher counts of *Roseburia*, *Clostridium* cluster IV and *Faecalibacterium* than patients without enteropathy[70]. This study emphasized the clinical relevance of the intestinal bacterial composition to predict patient responses to cancer therapy by reducing the risk of treatment-related toxicity[71].

Efficacy of therapies. The gut microbiota is also involved directly in the absorption and metabolism of antineoplastic drugs through the modulation of gene expression and physiological properties of the local mucosal barrier[72]. The activities of several chemotherapeutic agents have been tested in different microbial contexts. For instance, Lehouritis et al. demonstrated that non-pathogenic bacteria could influence the efficacy in mouse models of over 50% of tested chemotherapeutic drugs in several ways, including the inhibition of the antitumour activity of

Box 2 | **Common terminology for data-driven sciences in medicine**

Data science
Data science is the set of methodological principles and multidisciplinary techniques aimed at interpreting and extracting knowledge from data through the relative analytical phase. Thus, it is the field of study dedicated to the principled extraction of knowledge from complex data.

Data mining
Data mining is the actual extraction of usable information from large sets of data involving machine learning algorithms that incorporate data science principles. It is also known as 'knowledge discovery in databases'.

Data sharing
Data sharing is the ability to share data, metadata, products and information from multiple applications or users. The data sharing implies that the data are stored on one or more servers in the network. Data sharing is a main feature of a database management system. All those who produce, share and use data and metadata have the responsibility to guarantee the authenticity, quality and integrity of data, and to preserve privacy when appropriate.

Digital health
Digital health is the convergence of digital (web platforms and mobile health applications) and genomic technologies with the provision of health, health care and lifestyle to improve the efficiency of health-care delivery and make drugs more personalized and precise.

Precision medicine
Precision medicine is the tailoring of medical treatment to the specific characteristics of each patient. Precision medicine guides health decisions towards the most effective treatment for a specific patient and, therefore, improves the quality of care, while reducing the number of diagnostic tests and unnecessary therapies. Precision medicine describes a model for providing health care that relies heavily on data, analysis and information.

Personalized medicine
Personalized medicine is based on an approach to patients that considers their specificity, in terms of genetic, molecular and omics features, but takes into account their preferences, beliefs, attitudes, knowledge and social context.

Natural language processing
Natural language processing is a branch of artificial intelligence that deals with the interaction between computers and humans through natural language. The ultimate goal of natural language processing, which is based on machine learning techniques, is to read, decipher, understand and derive a meaning from human languages.

Learning health system
A learning health system is one in which knowledge generation processes are integrated into daily practice to produce continuous improvement in care.

Machine learning
Machine learning is the field of study that focuses on how computers learn and improve from data. The learning algorithms create models that can make predictions or decisions without being explicitly programmed to perform the task. The machine learning model can be as simple as a set of rules and thresholds that were automatically derived from the data.

Deep learning and deep neural network
Deep learning is a subfield of machine learning concerned with algorithms called artificial neural networks, which are inspired by the structure and function of the brain. Deep neural networks use several layers to transform data using sophisticated mathematical modelling techniques.

Domain expertise
Domain expertise is the understanding of problems in a given domain (for example, health care) that helps contextualize the application of data science to solve these problems in the real world.

gemcitabine or the induction of the prodrug CB1954 (REF.[73]). Moreover, *Parabacteroides distasonis* abrogated the antitumour effect of doxorubicin in a preclinical in vivo mouse model bearing subcutaneous cancers of both metastasizing melanomas and non-metastasizing fibrosarcoma cell lines[74], whereas in tumour-bearing germ-free mice the production of reactive oxygen species and concomitant tumour DNA damage were reduced in response to oxaliplatin, thereby impairing the efficacy of treatment[75,76]. This phenomenon was also common to cyclophosphamide treatment because of the increased permeability of the intestine in response to cyclophosphamide, which allows the translocation of specific species of Gram-positive bacteria to the lymph nodes and spleen of mouse models. These findings suggest that the gut microbiota can control treatment efficacy by blending chemotherapeutic effects and immune system activities in the tumour microenvironment[77,78]. In one study, two bacterial species, *Enterococcus hirae* and *Barnesiella intestinihominis*, were associated with longer progression-free survival in patients with ovarian or lung cancer who underwent chemotherapy and immunotherapy and improved the immunomodulatory properties of cyclophosphamide[78].

Specific microbial profiles have been associated with the clinical response to ICIs in several epithelial cancers (such as melanoma and renal-cell carcinoma)[79]. Tumours produce molecules, such as PD1, PDL1 and cytotoxic T lymphocyte-associated protein 4 (CTLA4) that downregulate the immune response and enable the immune tolerance and immune escape of cancer[80,81] Monoclonal antibodies against PD1, CTLA4 and PDL1 are FDA-approved ICIs that unleash the patient's immune response against tumours[82].

Preliminary studies also show that the increased abundances of specific bacteria, such as *Akkermansia muciniphila*, *E. hirae* and *Alistipes* spp., in renal-cell carcinoma and/or non-small-cell lung cancer, predict a more favourable response to ICIs[62] whereas, in melanoma mouse models, specific *Bacteroides* (*B. fragilis* and/or *Bacteroides thetaiotaomicron*) and *Burkholderiales cepacia* in the gut were associated with ICI responses[59]. Moreover, *Bifidobacterium* seems to increase the efficacy of anti-PD1 therapy in mouse models, reducing bladder tumour growth[80].

Finally, in patients with melanoma, Gopalakrishnan et al. found a relative abundance of bacteria of the Ruminococcaceae family in those responding to ICIs[61], whereas Frankel et al. reported that all ICI responders were enriched in *Bacteroides caccae*[60]. In this latter study, subjects undergoing combination treatment with ipilimumab and nivolumab were enriched with *Faecalibacterium prausnitzii, B. thetaiotaomicron* and *Holdemania filiformis*, while treatment with pembrolizumab was associated with higher levels of *Dorea formicogenerans*. Matson et al. instead found that eight species were enriched in patients with metastatic melanoma responding to ICI treatment: *Enterococcus faecium*, *Collinsella aerofaciens*, *Bifidobacterium adolescentis*, *Klebsiella pneumoniae*, *Veillonella parvula*, *Parabacteroides merdae*, *Lactobacillus* spp. and *Bifidobacterium longum*[83]. By contrast, non-responders were associated with *Ruminococcus obeum* and *Roseburia intestinalis*[83].

In addition to the work in melanoma, Routy et al. examined microbial associations in patients with non-small-cell lung cancer and renal-cell carcinoma and found that greater metagenomic species richness in the gut correlated with clinical response, and *A. muciniphila* was the most highly correlated species with a response to ICIs[62]. Faecal microbiota transplantation from patients with cancer who responded to ICIs into germ-free or antibiotic-treated mice ameliorated the antitumour effects of treatment[62]. Sivan et al, in another study in mice, demonstrated that the oral administration of *Bifidobacterium* alone improved melanoma control to the same degree as ICIs, and combination treatment nearly abolished the tumour outgrowth[63].

Finally, Peled et al. observed patterns of gut microbiota disruption characterized by loss of diversity and domination by a single taxon following allogeneic haematopoietic cell transplantation from 1,362 patients[84]. Subgroup analyses identified an association between lower intestinal diversity and higher risks of transplantation-related death and death attributable to graft-versus-host disease.

Microbiome big data in cancer

Although metagenomic sequencing technologies have revealed the presence of bacteria in several cancer samples and some cancers have been associated with specific microorganisms or microbial profiles[85–88], early available studies from single cohorts have not always shown consistent findings owing to confounding factors and small sample sizes. In the past few years, large-scale analysis of faecal and endothelial (mucosal) microbiomes has been applied to patients with cancer, not only to provide reliable sample sizes but also to compare samples from different populations (in terms of geographical origin, culture, diet and lifestyle, avoiding the risk of influence by possible confounders), with the aim of validating clinically relevant microbiome signatures associated with different cancer types[45,89].

Armour et al. conducted the first integrative functional analysis of nearly 2,000 publicly available faecal metagenomic samples obtained from eight clinical studies, and beyond identifying characteristics of the gut microbiome that associate generally with disease they also found specific microbial functions that robustly stratified diseased individuals from healthy individuals as controls[90]. Other cohorts are derived from wide, multicentre genomics programmes that have included analysis of faecal metagenomes of patients with cancer. For example, The Cancer Genome Atlas performed RNA sequencing and/or whole-genome sequencing on >4,000 tumour samples to identify microbial signatures (including those from bacteria, viruses, fungi, bacteriophages and archaea) specifically associated with each cancer type[91–93]. Other projects were specifically directed towards the analysis of the human gut microbiota and included not only original studies including new cohorts of patients but also meta-analyses of cohorts from already published studies[94]. The first meta-analyses (most of which investigated the associations between the microbiota and CRC) were based both on 16S ribosomal RNA (rRNA) gene amplicon data[89,94] and data derived from shotgun sequencing of the faecal microbiota[45,95], and have reproducibly identified compositional and functional microbiome fingerprints that predict the presence of cancer. Among 16S rRNA studies, Shah et al. found that the abundance of specific taxa (including *Parvimonas micra* derived from ATCC 33270, *Streptococcus anginosus* and uncultured members of Proteobacteria) are consistently increased in patients with CRC[89]. CRC tissues from a Malaysian cohort of patients (especially those tissues originating from the right colon) also showed an increased presence of invasive biofilms, symbionts with tumorigenic properties (*B. fragilis*) and pathogenic microorganisms from the oral cavity (including *F. nucleatum*, *P. micra* and *Peptostreptococcus stomatis*)[94]. In 2019, a meta-analysis derived from shotgun sequencing studies associated CRC with higher microbial richness and activity in the gluconeogenesis,

putrefaction and fermentation pathways and an overabundance of the choline trimethylamine-lyase genes[96]. Similarly, Wirbel and colleagues discovered that a core set of 29 microorganisms is markedly enriched in patients with CRC, and functional analysis of CRC metagenomes revealed an overabundance of protein and mucin catabolism genes and depletion of carbohydrate degrading genes, in conjunction with increased production of secondary bile acids[45].

Most available studies have focused on identifying specific bacterial taxa or bacterial functional pathways associated with the development and progression of cancer. Interestingly, with the objective of clarifying the role of microbiota in CRC carcinogenesis and establishing the pathological drivers, a platform (ColPortal) that integrates omics and epigenetics data with demographic information, location, histology (including digital histological slides), tumour staging, tissue prognostic factors, molecular biomarker status and clinical outcome for each patient with CRC has been developed[96]. Furthermore, based on the influence of the gut microbiota on ICI efficacy[61,62,97], large cohort projects, such as the ONCOBIOME project, have focused their interest on identifying characteristics of the human gut microbiome that are associated not only with cancer development and progression, but also with tumour response to different therapies (including chemotherapeutics, ICIs and vaccines) and with the occurrence of treatment-related adverse effects.

In summary, the technologies that have enabled the identification of the microbiota have undergone a very important and rapid evolution. The use of multi-omic techniques has led to a collection of huge amounts of data and challenges for data storage and analysis, which will quickly create a further challenge to simplify the data into biologically relevant information.

Machine learning

The size of big data and the processing power required to analyse it coupled with the open source community has led to incredible breakthroughs in many areas of data analytics such as computer vision, speech recognition, medical diagnostics and gaming. Although statistics facilitate the understanding and interpretation of data, ML includes algorithmic methods that enable machines to solve problems without specific computer programming and, therefore, leads the way in predictive modelling tasks[98,99]. In other words, ML is a discipline in computer science wherein machines (that is, computers) are programmed to process the input data, learn or train the underlying model, and finally elaborate predictions on new data. Thus, ML is essentially an interplay between large datasets and a specific class of AI (BOX 2) that includes learning mechanisms (such as deep neural network) that can be trained to be highly accurate in finding complex patterns within big data and to transform inputs to more useful and interpretable information. ML is able to analyse large-scale data from different settings (for example, demographic, laboratory and imaging data) and combine them into predictions for disease risk, diagnosis, prognosis and appropriate treatments[19].

The application of ML tools requires substantial rigour in the careful creation of frameworks and experimental setups (in terms of data handling, selection of data for training and validation) so that the testing and identification of the questions can be answered by each data stream. The appropriate comparative framework is essential to distinguish resulting improvements based on actual contributions from those based on data manipulation, changes in metrics and overfitting (discussed below). In practice, the quality of input data and the amount of training data are key factors of the entire ML process. The ML model design process itself includes the search for the optimal set of model parameters that translate the features of data into accurate predictions of the labels[100]. The parameters are detected through a series of steps, in which the parameters are identified, the performance of the model is evaluated, the errors are identified and corrected, and then the process is repeated. This process is called training and will continue until it is not possible to further improve the performance of the model[101].

A given ML model can be trained to predict training data with high accuracy without being able to make accurate predictions on test data. In this case, the parameters for the model are so specifically adapted to complex training data (overfitting) that they do not provide predictive power outside of these data. On the contrary, the ML model could be too simple and not accurately predict training data (underfitting). Overfitting and underfitting are the main causal factors behind the poor performance of ML approaches[101]. In the specific field of gastroenterology, continuous variable fitting techniques can be used to predict quantity of a therapeutic response or the development of gastrointestinal disease prior to the manifestation of symptoms[102], or to provide computer vision in endoscopy to automatically detect lesions[103,104]. In the field of clinical IBD research, the advent of AI, ML and systems biology has paved the way for the efficient integration and interpretation of big datasets for discovering clinically translatable knowledge[105,106].

Moreover, there are two principal categories of ML methods[107]. In the unsupervised method, the labels on the input data are unknown, and the method learns only from patterns in the features of the input data. The goal of the unsupervised method is to group or cluster subsets of the data based on similar features. In this case, no independent predictive model is produced; examples include discovering patient clusters based on genetic signatures[108] or identifying novel signatures of health and disease[109]. In other studies, unsupervised computational methods were used for more comprehensive insights into biological systems by collecting and analysing diverse high-throughput data[110,111].

On the contrary, with the supervised approach, labels are available for the input data and are used to train the ML model to recognize patterns that are predictive of the data labels. The trained model in the supervised method is a predictive one, where an underlying inferred network is used to make predictions on a novel sample. Such approaches have been highly successful in the characterization of drug metabolism of action[112,113] or drivers of disease states[114] or applied to ageing research (for example, to find biomarkers of ageing)[115]. Semisupervised approaches can be useful when the labels on the input data are incomplete or only a small amount of the data is labelled[116]. This scenario can occur in biological contexts; for example, for a set of genes of interest or in multi-omics studies, only a small subset might be functionally annotated[117,118]. Finally, ensemble methods, combining multiple independent ML models that integrate multiple independent predictions into a single predictive model, can be used to obtain better and more robust predictions compared to an individual learner.

To conclude, ML methods are context specific and require careful experimental design to appropriately test a hypothesis. A great model for one problem might be inappropriate for another problem, and vice versa, and needs to be tailored for the dataset and research question being asked. The domain (such as life science) expert

knowledge together with ML expertise have a crucial role in designing ML solutions. The source of variability in estimation of the generalization error that arises from mistakes in the process of designing an ML solution is much larger than the variance that comes from the actual choice of core learning mechanism. FIGURE 1 illustrates the framework that should be followed when designing ML solutions to a particular problem[119]. As the field of bioinformatics becomes increasingly dependent on ML algorithms for many different tasks, such as cohort classification, outcome prediction, study participant clustering and/or stratification and in-depth feature analysis, it is important that researchers are aware of the different algorithms and tools available for such tasks and how they are implemented[120,121], which enables scientists with a cursory knowledge of ML techniques to implement many common methodologies.

Machine learning in oncology. ML is highly flexible compared with traditional biostatistical methods and can have countless potential applications. Some examples of application of ML in the areas of oncology already exist. CURATE.AI is an automatic learning application developed at the National University of Singapore, Singapore, and the University of California in Los Angeles, USA, that helps doctors to choose the optimal dosage and therapeutic combination for a given patient[122,123]. Many clinical studies are under way to evaluate the application of ML in different oncological environments, such as in improving diagnosis, treatment and prediction of metastasis in breast[124,125], prostate[126], ovarian[127], bladder[128], pancreatic[129] and brain cancer[130]. Interestingly, great expectations are also placed in the prevention[131,132] and prediction[133–135] of cancer. Ultimately, therefore, all these complex applications of ML learning tools in clinical settings offer great promise in terms of improving health-care delivery through personalized recommendations based on large amounts of clinical data[136]. ML methods could be developed to support precision treatment rules to understand how the effects of treatment are modified by clinical features (such as illness severity, disease subtypes and comorbidity), sociodemographic characteristics (such as patient age, sex and education), imaging and biomarkers (genetic, histopathological, omics and so on)[137]. Furthermore, ML algorithms and new computational models offer also the opportunity to generate computational drug networks to predict the efficacy of approved drugs against relevant carcinogenic targets (that is, drug repurposing), as well as to select patients with better response or disease biomarkers[138]. On the other hand, in the field of digital pathology, the advent of AI and ML tools enables mining of new morphometric phenotypes and could improve patient management across a range of cancer types[139]. Penson et al. developed an ML approach to predict tumour type from targeted panel DNA sequence data obtained at the point of care and were able to predict tissue of origin in oncologic practice and integrate the pathological diagnoses, often with important therapeutic implications[140]. Grewal et al., with an ML method using whole-transcriptome RNA sequencing data, have demonstrated the practical use of a whole transcriptome-based pan-cancer method in diagnosing primary and metastatic cancers (including colorectal and pancreatic adenocarcinoma, and cholangiocarcinoma) and resolving complex diagnoses (rare cancer types, primary cancers, treatment-resistant metastatic cancers and cancers of unknown primary origin, including cholangiocarcinoma)[141].

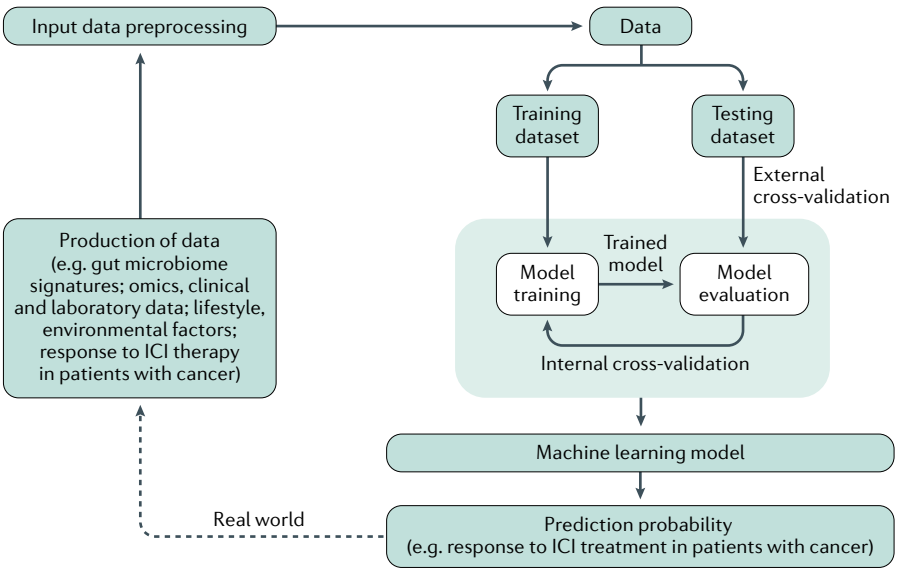

Fig. 1 | **Representation of a machine learning workflow.** A hypothetical scenario of searching for gut microbial signatures to predict response to immune checkpoint inhibitors (ICIs) in patients with cancer is presented. Through a process called data preprocessing, real-world data are converted to a clean dataset used to train the model. This process is fundamental to achieve good results from applying the machine learning model. The training dataset is the material through which the computer, by using an algorithm, learns how to process information. Training data are used to fit the model (typically *.fit*, *.train* functions in python). The testing dataset is a set of unseen data used only to assess the performance of the model; this dataset should be used once and there should be no feedback to the modelling process from the scores obtained on the testing data. External cross-validation can be used to minimize bias when estimating the generalization error. Validation data are used to test a model. Internal cross-validation can be used to estimate the accuracy of the machine learning model on unused parts of the training data to tune the system parameters. The scores obtained by the internal cross-validation (training test) are used to drive changes in the modelling process: selection of model hyperparameters, selection of features, selection of learning mechanisms (XGBoost, neural networks and so forth), selection of regularization strategy (for example, early stopping and learning rate adaptation), selection of data preprocessing and postprocessing routines and their parameters. The model evaluation scores the model, which can help to find the best model that represents the data or to estimate how it will work in the future. The probability is that which the model correctly classifies (for example, diagnostics) or that predicts condition, disease, treatment and so forth (prognostics).

Machine learning and the microbiome. The conversion of huge amounts of data arising from metagenomics on the human microbiome into notable mechanistic insights from a biological and clinical point of view remains a challenge, but represents a great opportunity for network biology approaches that include ML algorithms[4,100,142–146]. The Human Microbiome project, the ENCODE project consortium and countless genome projects are producing enormous amounts of data concerning humans and their microbiome. These datasets are essential to obtain in-depth information on biological systems and complex diseases, but their potential can be fully realized only through higher-level analysis. Thus, it is imperative to focus the

analytical approach on tools and techniques that are appropriate in managing large, heterogeneous, complex datasets. ML aims to solve complexities, providing next-level analysis that enables new views and new hypotheses on living systems.

In response to these and the challenges already mentioned, the ML4Microbiome COST Action was launched in 2019, with the aim of optimizing and standardizing the use of ML in microbiome research. Several other completed, high-profile projects have worked to increase the amount of freely available, fully annotated microbiome data[147]. The NIH Human Microbiome Project has been carried out over 10 years and has analysed microbiome and host activities in longitudinal studies of disease-specific cohorts (such as preterm birth, IBD and type 2 diabetes mellitus) by creating multi-omic datasets that were collected, integrated and distributed through public repositories as a community resource[148]. The Microbiome Learning Repo is a public, web-based ML repository for microbiome datasets that can be another important resource for algorithm developers[149].

In addition, a number of excellent studies have been published, covering experimental design and analysis of microbial community multi-omics[90,150], the best practices for analysing microbiomes[151] and the advances in understanding the connections of human gut microbiome populations to human health[152]. Studies have begun to explore the power of ML to use microbiome patterns in predicting host characteristics[153,154] or tracking microbiota-for-age development in children[155]. In some case, modelling features of the 'healthy' microbiome was considered a step towards defining general microbial dysbiosis[156]. Wirbel et al. reported training on data from a meta-analysis of eight geographically and technically diverse faecal shotgun metagenomic studies on CRC and identified a core set of 29 species that ameliorated the detection accuracy and disease specificity[45]. Thomas et al. combined analysis of heterogeneous CRC cohorts (including publicly available datasets and several new cohorts) and identified reproducible microbiome biomarkers and accurate disease-predictive models that can be useful to develop clinical prognostic tests and hypothesis-driven mechanistic studies[95]. An ML model has been developed to investigate the contribution of the gut microbiome to treatment outcomes in a heterogeneous cohort that included multiple cancer types[157]. ML methods have also been used to study antibiotic-resistance genes in the infant microbiome[63]. In the field of IBD, Zuo and colleagues[158] used ML-based clustering to define viral metacommunities in rectal mucosa samples derived from patients with ulcerative colitis.

As microbiome science is now progressively co-evolving with ML methodology, we are seeing the emergence of improved algorithms for both microbiome omics and other omics data. For instance, support vector machines have not only been used to classify samples into cohorts but also to identify molecular targets that could potentially be used for diagnostics, prognostics or interventions[159]. However, to satisfy the overall demand for an ML approach in microbiome research, software such as QIIME 2 (REF.[160]), MicrobiomeAnalyst[161] and USEARCH[162] have started to incorporate ML methods that can also be used by researchers who do not necessarily have bioinformatics training. The 'predomics' is another newly developed and innovative ML approach inspired by microbial ecosystem interactions and tailored for metagenomics data[163]. It is a new algorithm that helps in providing diagnostic decisions in the microbiome field by discovering accurate predictive signatures and providing interpretability of findings.

However, microbiome data are complex, and traditional ML methods can be limited by the representation ability of the models and cannot learn complex patterns from the data. In the past few years, sophisticated ML in microbiome analysis has been proposed[164], including using deep neural networks to predict response to anti-integrin biologic therapy in IBD by the analysis of gut microbiome function[165]. Deep learning has been widely applied to fields ranging from automatic driving to image recognition owing to its flexibility and high capacity to train enough data. However, deep learning is considered hardly practical in reality and is difficult to interpret[166] when used in microbiome-wide association studies[167–169].

Big data integration

The large-scale collection of biological, clinical and translational bioinformatics datasets supports the possibility of generating meta-metabolic network models, based on shared metabolites, for any given microbiota–host system. These models could be useful to predict and gain insights into the synergistic and dysbiotic relationships between hosts and their microbial components[170], as well as to enable microbiome-based predictions of phenotypic outcomes[4,171]. The microbiome is just one of many big data omics sources being used for classification modelling and, as increasing amounts of omics data become available, complex and combinatorial dynamics must be further elucidated. Although the combination of omics data types might be challenging[172], it also has great potential (as some examples earlier demonstrate) and presents opportunities to understand the underlying relationships between types of data and how they contribute, in combination, to a specific phenotype.

ML is a solution for data integration[173], and there are many methods for integrating data types and exploring their relationships. Other similar projects, including the collection and sharing of linked genetic, physical and clinical information on a population scale, have enabled researchers to quickly search for genetic associations and produce a large number of publications[174]. Many of these studies have aggregated different data to enable studies on a much larger scale. When combined, different types of omics data can provide more information than the sum of each analysed alone. Several studies have shown the utility of collecting and analysing diverse high-throughput data for more comprehensive insights into biological systems. Shomorony at al.[109], by using an unsupervised ML approach, were able to identify novel biomarker signatures of health and disease risk from multimodal data (including metabolome, microbiome, genetics and advanced imaging). Moreover, Perkins et al. introduced a platform of deep quantitative multimodal phenotyping (include whole-genome sequencing, advanced imaging, metagenomic sequencing, metabolome, and clinical laboratory, medical and family history) that seeks to provide a comprehensive, predictive and personalized assessment of an individual's health status by identifying previously undiagnosed disease states and also to identify early disease biomarkers[175]. Argelaguet et al. showed a need for such integrated analysis of heterogeneous data by demonstrating its utility in identifying major drivers of variation in chronic lymphocytic leukemia[111]. Price et al. revealed communities of related analytes associated with diseases (such as cardiometabolic disease and IBD) by integrated analysis on a multimodal dataset (including genome, clinical tests, metabolomes, proteomes and microbiomes)[110]. Finally, ML algorithms, by analysing multi-omics features and their links to childhood irritable bowel syndrome coupled with nutritional interventions, were able to identify associations between microorganisms, metabolites and

abdominal pain, potentially able to lead to new microbiome-guided diagnostic and therapeutic strategies[176].

For all these reasons, studies are moving towards multi-omic approaches, especially those focused on precision medicine. For instance, both the gut microbiota and the corresponding host genotype seem to have a role in insulin secretion and the induction of particular diet-associated metabolic phenotypes[177]. A combination of longitudinal genome, immunome, transcriptome, proteome, metabolome, microbiome and wearable monitor data were used to enhance the prediction of risk of cardiovascular disease[178]. Furthermore, the complex relationship between the microbiome and host DNA methylation in the pathogenesis of diseases such as diabetes and CRC has been highlighted in the past few years[86,171,179,180]. Finally, the 1,000 IBD project comprises data on >1,200 patients with IBD, including microbiome, host genotypes and transcriptomes from intestinal biopsies, along with phenotype data (diet, drug responses and environmental factors)[181]. Similarly, the IBD Multi-omics Database provides a comprehensive description of host and microbial activities in IBD[170].

Such rich and heterogeneous omics datasets present both great combinatorial opportunities for ML methods, but also present challenges as the number of unique features (such as the range of bacteria present, gene expression and single nucleotide polymorphisms) are much larger than the number of samples (patients, time points and so on). The power calculations and estimates of the sample size are important preconditions for testing the research hypothesis adequately and for drawing meaningful conclusions that generalize beyond the sample of patients in the study itself. The level of confidence about these generalizations is determined a priori by the researchers based on what they consider an acceptable level of error (that is, to considerably reduce the risk of obtaining false-positive or false-negative results). Recommendations published in 2020 suggest incorporating α-diversity and β-diversity metrics into power calculations[182]. Nevertheless, the actual data size can also be an issue for computational processing, especially for shotgun sequencing (typically two million 150 bp long paired-end reads, or 1.5 gigabytes, per sample) compared with 16S rRNA data (50,000 300 bp long paired-end reads, or 30 megabytes, per sample), where the former analysis often requires high-powered computing and storage servers.

Precision medicine

ML is rapidly becoming a key approach in the development of precision medicine[183]. Precision medicine describes a model of health care based on data, analysis and technology that takes into account people at an individual level. This model goes beyond genomics and, to be successful, must include the centrality and commitment of the patient, genomics, molecular and digital technologies, data sharing and analytical skills using ML algorithms (BOX 3).

As high-throughput methods for data generation become faster and less expensive, researchers are increasingly gaining access to a wealth of molecular information from human cohorts. The use of these multidimensional data requires the development of standardized methods of aggregation and data analysis and interdisciplinary translation of emerging computational techniques, such as ML, natural language processing, AI and other data-driven sciences (BOX 2).

Box 3 | **Rationale for AI-enabled health care in cancer**

- Specific intestinal microbial signatures have the potential to affect cancer development and influence safety and tolerability or even efficacy of cancer therapies.
- Large-scale microbiome datasets of ever increasing size and complexity are continuously being produced.
- Big data imposes the use of dedicated analytical and statistical approaches to move from the bioinformatician's bench to the patient's bedside.
- Machine learning, a type of artificial intelligence (AI), is able to analyse different large-scale data setting and combine them into predictions for disease risk, diagnosis, prognosis and appropriate treatments.
- Potential machine learning-based applications in cancer might include the development of predictive biomarkers, precision medicine and tailored modulation of the gut microbiome.
- Challenges include substantial rigour in careful creation of machine learning frameworks and experimental setups, in terms of data handling, validation and testing of results, to prove their worth in real-world settings.

With the application of these new analytical methods, and when these (anonymised) data become publicly available, robust classification models can be developed that can potentially be used, in conjunction with traditional medical practices and clinical information (phenotypic meta-data), to define the dynamic patterns of health and disease and to promote tailored medical approaches[19,184] (FIG. 2). For instance, penalized linear regression using either oral[179] or faecal microbiota sampling[185] has successfully been used to differentiate patients with early-stage CRC from healthy individuals with a greater precision than the traditional faecal occult blood test method. Similarly, Zeevi et al, using an ML algorithm and by integrating laboratory parameters, dietary habits, anthropometrics, physical activity and gut microbiota, showed that it is possible to predict a personalized postprandial glycaemic response to real-life meals[186]. The authors, with a blinded randomized controlled dietary intervention based on their algorithm, obtained a substantially lower postprandial glycaemic response and consistent alterations to gut microbiome configuration. In IBD, the gut microbiome sampled from different regions of the intestine has been used to classify subtypes of disease[187] and in combination with metabolomics data[188]. Non-intestinal microbiome markers have also been shown to have prognostic and diagnostic potential. For example, microbiome diversity in the respiratory tract is a predictor of lung function in patients with cystic fibrosis[189]. Similarly, microbiome markers in psoriasis can be used to stratify clinically relevant skin types[190]. In addition, the gut microbiome is increasingly being linked to drug efficacy and function — most notably, it has been shown to affect response rates for drugs for lung[62], kidney[62] and skin[83] cancer and be functionally implicated in the response to drugs for diabetes[191,192], cardiac failure[193] and depression[194].

The potential of microbiome data in precision medicine is, therefore, substantial, and it will continue to grow along with intensified research in these areas. Hopefully, advances in understanding tumour development and treatment responses can benefit from a combined analysis generated by the integration of parallel datasets (genomics, microbiomics, proteomics and transcriptomics). Analysis data are further enhanced by combining many of these networks into larger sets of cohort-based meta-data (disease targeting and therapeutic responses), providing unique insights into the elusive field of cancer in terms

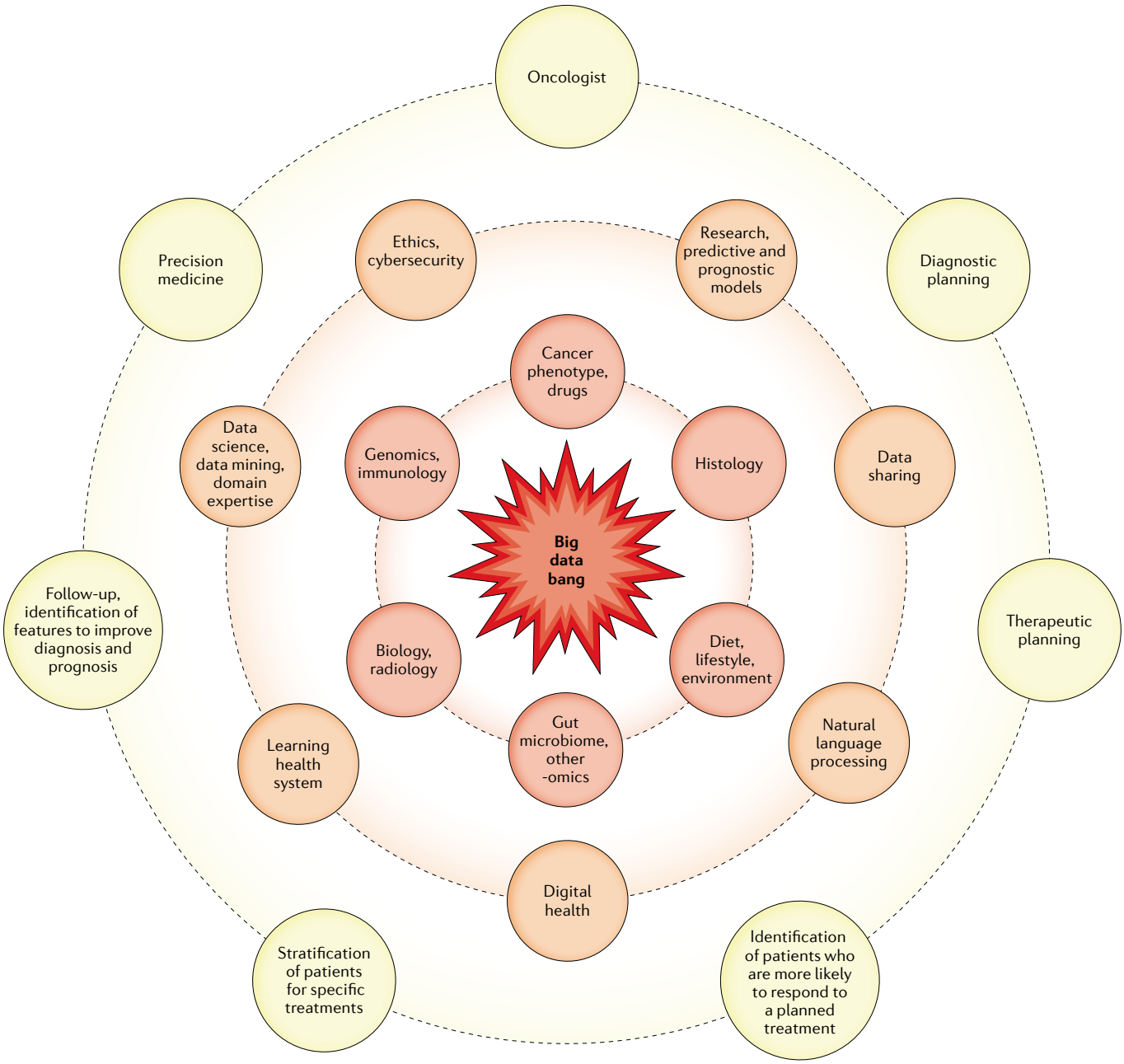

Fig. 2 | **From big data to precision medicine: moving through the data science.** The figurative analogy with the 'big bang' is justified by the deluge of big data that has prompted the rapid and exponential growth of data science in medicine. The intersection of data science, analytics and precision medicine in generating a learning health system has the potential to carry out research in the context of clinical care and at the same time to utilize and improve the tools and information used for delivering improved patient outcomes. In the representation, the circles are of different increasing amplitude to give the idea of the generative temporal expansion of big data first in data science and then, on the clinical level, in the pathways of precision medicine. The different colour coding of the different circles moving outwards indicate the different categories of features.

of gene expression, dysregulation of the epigenome and the response to treatment. By generating and mapping omics profiles for each individual patient through these meta-networks, it will be possible to achieve the goal of true precision therapy. This approach can therefore offer the possibility of new therapeutic options and the development of innovative models of disease.

Potential for improving therapy

The use of big data with ML approaches could help discover which microbial signatures are consistently and reproducibly effective to predict or treat cancer in patients. To date, although the modulation of the gut microbiome is advocated as a revolutionary option to improve the outcomes of patients with cancer, there are still several gaps to translate this possibility and available evidence into clinical practice. We are still not aware of the complete landscape of interactions between the gut microbiota and drug metabolism in cancer as current data are based on small population samples. The gut microbiome has considerable variability not only among different geographical populations but also across different periods of life and is influenced by a plethora of factors,

principally related to health status, diseases, intake of medications and diet[195,196].

This biological complexity currently prevents the implementation of microbiome-based precision medicine into clinical practice. For instance, despite multiple clinical trials reporting some efficacy of probiotics in counteracting the gastrointestinal adverse effects of radiotherapy[197–201] and chemotherapy[202–207], a systematic review of 12 studies from the Cochrane group concluded that the evidence supporting the use of probiotics to prevent and/or treat diarrhoea associated with cancer treatments is low or very low, as available studies are biased by low statistical power and heterogeneity[208]. In that context, ML-based studies training integrated multimodal omic big data could decrease the heterogeneity of findings, allowing the identification of probiotic strains that would be consistently and reproducibly effective in ameliorating chemotherapy-related and radiotherapy-related adverse effects, which often lead to a decrease in or the suspension of therapy[209], and which are usually inconsistently managed[210].

Targeted manipulation of the gut microbiota could also improve the response of patients to different antineoplastic therapies (chemotherapy, immunotherapy and radiotherapy)[62,66]. This objective can be directly achieved by increasing the efficacy of treatments (in patients whose microbial composition predicts response failure) or indirectly by improving the compliance to therapies reducing drug-related adverse effects[66,202–207,211]. The more futuristic and visionary field of microbiome modulation is related to the extensive microbial repopulation that is promised by faecal microbiota transplantation[212] and targeted microbial therapies, such as next-generation probiotics[213], synthetic microbial consortia[214] or genetically modified microorganisms[215,216]. Despite the absence of reliable clinical data of the effectiveness of these strategies, there is a lot of enthusiasm from the scientific community on this topic, as shown by the increasing number of active trials (see Supplementary Table 1). However, novel consistent associations between the human microbiome and health and disease may now emerge from ML analysis of big multi-omics data. Peculiar host–microbiome interactions can be targeted for new diagnostics and therapeutics to enhance precision medicine.

Difficulties in clinical translation

ML approaches have various limitations or requirements that must be addressed before they can be effectively used in health care (BOX 4). One of the most challenging tasks is to accurately estimate the true real-life performance of the model. The performance assessment framework is often chosen to maximize performance, which can result in a lack of comparability and reproducibility and difficulty in assessing the approaches that will work in the real world and those that will not. New areas of ML application, such as microbiome data, tend to suffer from the lack of skilled ML experts in the research community. In addition, a larger-scale assessment of microbiome changes in cancer is clearly required. Available cancer microbiome studies are few and small in size[185,217], and differences in their methodologies prevent the translation of their results into clinical practice. Evaluation of the metagenomes of large human cohorts can derive clear and consistent data[95], with a larger influence not only on clinical research but also on the health outcomes of the general population. Future studies should have an adequate sample size, appropriate to the specific study scenario, to enable valid conclusions and to be sensitive to the heterogeneity of cancer in terms of histology and stage, and define clear clinical outcomes so that the results can be easily applied to clinical practice.

Furthermore, a shift in the mindset of all professionals involved in the wide field of cancer microbiomics is necessary[218]. Physicians must learn new skills and new ways of interpreting knowledge for application in the clinical setting. The algorithms require specific expertise in preprocessing data and the ability to handle large datasets to direct them to the solution of an actual clinical problem. Although it is not necessary for physicians to know in detail the mathematical procedures underlying an ML algorithm, they could instead be instructed on the dataset used in a given clinical context and on the relative weights assigned to each type of data. Just as clinicians know that a certain drug does not always work, they will also have to know that AI-enabled software will not be perfect or equally reliable in every patient. Similar to the results of a clinical trial, clinicians could consider characteristics such as the sensitivity and specificity of a given dataset in responding to a particular clinical question, such as a particular disease risk or treatment response. The training should ultimately evolve in the curricula of medical schools or through continuous medical training, so that clinicians work effectively in the context of AI. Research bioinformaticians should also change their research activity, not limiting their efforts only to the sterile technical implementation of such systems but aiming to find appropriate ways to integrate them into clinical practice. Finally, the adoption of an AI-based approach in clinical practice also requires a clear identification of the levels of liability in the event of medical error, or malfunctioning of the systems, or in evaluating the training of doctors on the use of ML tools. Currently, there is no clear definition of liability when a ready-to-use ML algorithm is wrong (is it down to the physician, the hospital or the manufacturing company). On the other hand, patients should be also aware

Box 4 | **Limitations and difficulties in translation of AI-assisted health care**

Lack of large-scale data and validated clinical prediction models
The available cancer microbiome studies are still few and small in size.

Professional training
Specific training should evolve so that physicians and bioinformaticians can communicate and be effective in an artificial intelligence (AI)-embedded landscape.

Liability
It is crucial to establish the levels of liability for medical errors or training of the professional figures.

Consent of patients
Patients should be informed about how their information will be used and shared with third parties.

Privacy
It is crucial to find new paradigms to protect patients' privacy according to national and international laws.

Cybersecurity
Cybersecurity must always be considered at every stage of the development of the AI-enabled system.

Control
A careful planning of who controls and authorizes the use or transfer of health data is fundamental.

of how ML tools are used, and by which doctors with which specific training, for treating their health problems. In all those situations, data controllers must ensure that patients are well informed and are asked for their consent, especially if their data records are to be used for commercial purposes by national companies or by other countries. The identification of new paradigms to protect patient privacy is another parameter of fundamental importance. Developers of the AI health-care approach, while seeking to potentially acquire large amounts of patient data to use with the ML algorithms, must take into account the various national and international laws that protect the privacy of health information. Understanding how to implement AI algorithms, using patient data while preserving people's privacy, and at the same time promoting appropriate management practices by companies remain real challenges for bioinformatics experts[219,220]. Another important aspect concerns the cybersecurity that must always be considered at every stage of the design, development, implementation and maintenance of the systems themselves. Fortifying the characteristics of inviolability in an AI system will require a precise understanding of the relevant cyber threats and a careful planning of who controls and authorizes the use or transfer of health data. This is particularly challenging when considering the storage of complex international medical data.

Conclusions

The volume of information and knowledge in the field of medical activity, wellness, diseases, treatments and prevention increases enormously each year and far exceeds the ability of doctors to process and translate them on a practical level in the clinic. Specifically, the research into the cancer microbiome has attracted much attention and expectations for improving the lives of patients with cancer through different avenues. In this context, precision medicine, by targeting the microbiota with diverse strategies (including nutrition, antibiotic selection, probiotic administration or faecal microbiota transplantation) is going to become one of the next frontiers for patients, providing new opportunities with tailored therapies for individual patients (BOX 3).

AI-assisted health-care and data-driven sciences have the potential to clarify the landscape of findings and enable clinicians to move these findings from the bioinformatician's bench to the patient's bedside[221] (FIG. 2). However, to realize these exciting prospects, it is crucial to face the great challenges underlying a safe and effective technological innovation in this area by developing consensus standards through the identification and discussion of short-term and long-term priority challenges. Changes in cultural and educational paradigms at various levels are required, including the shift towards the sharing of data. Only if the research community is conceptually ready to share and integrate data worldwide will AI tools be able to meet high expectations and contribute favourably to the advancement of biomedical research.

Giovanni Cammarota[1]✉, *Gianluca Ianiro*[1], *Anna Ahern*[2], *Carmine Carbone*[3], *Andriy Temko*[4,5], *Marcus J. Claesson*[2], *Antonio Gasbarrini*[1] *and Giampaolo Tortora*[3]

[1]*Gastroenterology Department, Fondazione Policlinico Universitario Agostino Gemelli-IRCCS, Università Cattolica del Sacro Cuore, Rome, Italy.*

[2]*School of Microbiology and APC Microbiome Ireland, University College Cork, Cork, Ireland.*

[3]*Oncology Department, Fondazione Policlinico Universitario Agostino Gemelli-IRCCS, Università Cattolica del Sacro Cuore, Rome, Italy.*

[4]*School of Engineering, University College Cork, Cork, Ireland.*

[5]*Qualcomm ML R&D, Cork, Ireland.*

✉*e-mail: giovanni.cammarota@unicatt.it*

https://doi.org/10.1038/s41575-020-0327-3

Published online 9 July 2020

1. Marchesi, J. R. et al. The gut microbiota and host health: a new clinical frontier. *Gut* **65**, 330–339 (2016).
2. Yue, B. et al. Inflammatory bowel disease: a potential result from the collusion between gut microbiota and mucosal immune system. *Microorganisms* **7**, E440 (2019).
3. Zhang, Z. et al. Impact of fecal microbiota transplantation on obesity and metabolic syndrome — a systematic review. *Nutrients* **11**, E2291 (2019).
4. Thaiss, C. A. et al. Persistent microbiome alterations modulate the rate of post-dieting weight regain. *Nature* **540**, 544–551 (2016).
5. Mullish, B. H. & Williams, H. R. *Clostridium difficile* infection and antibiotic-associated diarrhoea. *Clin. Med.* **18**, 237–241 (2018).
6. van der Giessen, J. et al. Modulation of cytokine patterns and microbiome during pregnancy in IBD. *Gut* **69**, 473–486 (2020).
7. Konstantinov, S. R., van der Woude, C. J. & Peppelenbosch, M. P. Do pregnancy-related changes in the microbiome stimulate innate immunity? *Trends Mol. Med.* **19**, 454–459 (2013).
8. Maguire, M. & Maguire, G. Gut dysbiosis, leaky gut, and intestinal epithelial proliferation in neurological disorders: towards the development of a new therapeutic using amino acids, prebiotics, probiotics, and postbiotics. *Rev. Neurosci.* **30**, 179–201 (2019).
9. Tang, W. H. et al. Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N. Engl. J. Med.* **368**, 1575–1584 (2013).
10. Vivarelli, S. et al. Gut microbiota and cancer: from pathogenesis to therapy. *Cancers* **11**, 38 (2019).
11. Bi, J. H. et al. ClickGene: an open cloud-based platform for big pan-cancer data genome-wide association study, visualization and exploration. *BioData Min.* **12**, 12 (2019).
12. Rodriguez-Martin, B. et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* **52**, 306–319 (2020).
13. Zhang, J. et al. The International Cancer Genome Consortium data portal. *Nat. Biotechnol.* **37**, 367–369 (2019).
14. Brown, J. A., Ni Chonghaile, T., Matchett, K. B., Lynam-Lennon, N. & Kiely, P. A. Big data-led cancer research, application, and insights. *Cancer Res.* **76**, 6167–6170 (2016).
15. Evans, B. J. & Krumholz, H. M. People-powered data collaboratives: fueling data science with the health-related experiences of individuals. *J. Am. Med. Inform. Assoc.* **26**, 159–161 (2019).
16. Provost, F. & Fawcett, T. Data science and its relationship to big data and data-driven decision making. *Big Data* **1**, 51–59 (2013).
17. Sanchez-Pinto, L. N., Luo, Y. & Churpek, M. M. Big data and data science in critical care. *Chest* **154**, 1239–1248 (2018).
18. Gruson, D., Helleputte, T., Rousseau, P. & Gruson, D. Data science, artificial intelligence, and machine learning: opportunities for laboratory medicine and the value of positive regulation. *Clin. Biochem.* **69**, 1–7 (2019).
19. Rajkomar, A., Dean, J. & Kohane, I. Machine learning in medicine. *N. Engl. J. Med.* **380**, 1347–1358 (2019).
20. Sender, R., Fuchs, S. & Milo, R. Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans. *Cell* **164**, 337–340 (2016).
21. Lozupone, C. A. et al. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220–230 (2012).
22. Conlon, M. A. & Bird, A. R. The impact of diet and lifestyle on gut microbiota and human health. *Nutrients* **7**, 17–44 (2014).
23. Imhann, F. et al. The influence of proton pump inhibitors and other commonly used medication on the gut microbiota. *Gut Microbes* **8**, 351–358 (2017).
24. Thomas, S. et al. The host microbiome regulates and maintains human health: a primer and perspective for non-microbiologists. *Cancer Res.* **77**, 1783–1812 (2017).
25. Fessler, J., Matson, V. & Gajewski, T. F. Exploring the emerging role of the microbiome in cancer immunotherapy. *J. Immunother. Cancer* **7**, 108 (2019).
26. Scott, A. J. et al. International Cancer Microbiome Consortium consensus statement on the role of the human microbiome in carcinogenesis. *Gut* **68**, 1624–1632 (2019).
27. Lazar, V. et al. Aspects of gut microbiota and immune system interactions in infectious diseases, immunopathology, and cancer. *Front. Immunol.* **9**, 1830 (2018).
28. Pagliari, D. et al. Gut microbiota–immune system crosstalk and pancreatic disorders. *Mediators Inflamm.* **2018**, 7946431 (2018).
29. Bingula, R. et al. Desired turbulence? Gut–lung axis, immunity, and lung cancer. *J. Oncol.* **2017**, 5035371 (2017).
30. Gopalakrishnan, V. et al. The influence of the gut microbiome on cancer, immunity, and cancer immunotherapy. *Cancer Cell* **33**, 570–580 (2018).
31. Rugge, M. et al. Gastric cancer as preventable disease. *Clin. Gastroenterol. Hepatol.* **15**, 1833–1843 (2017).
32. Parsonnet, J. et al. *Helicobacter pylori* infection and the risk of gastric carcinoma. *N. Engl. J. Med.* **325**, 1127–1131 (1991).
33. Garrett, W. S. Cancer and the microbiota. *Science* **348**, 80–86 (2015).
34. Wu, S. et al. A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. *Nat. Med.* **15**, 1016–1022 (2009).
35. Raza, M. H. et al. Microbiota in cancer development and treatment. *J. Cancer Res. Clin. Oncol.* **145**, 49–63 (2019).
36. Pushalkar, S. et al. The pancreatic cancer microbiome promotes oncogenesis by induction of innate and adaptive immune suppression. *Cancer Discov.* **8**, 403–416 (2018).
37. Kostic, A. D. et al. *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* **14**, 207–215 (2013).
38. Rubinstein, M. R. et al. *Fusobacterium nucleatum* promotes colorectal carcinogenesis by modulating E-cadherin/beta-catenin signaling via its FadA adhesin. *Cell Host Microbe* **14**, 195–206 (2013).
39. Yoshimoto, S. et al. Obesity-induced gut microbial metabolite promotes liver cancer through senescence secretome. *Nature* **499**, 97–101 (2013).

40. Raskov, H., Burcharth, J. & Pommergaard, H. C. Linking gut microbiota to colorectal cancer. *J. Cancer* **8**, 3378–3395 (2017).
41. Li, S., Peppelenboscha, M. P. & Smits, R. Bacterial biofilms as a potential contributor to mucinous colorectal cancer formation. *Biochim. Biophys. Acta Rev. Cancer* **1872**, 74–79 (2019).
42. Belkaid, Y. & Hand, T. W. Role of the microbiota in immunity and inflammation. *Cell* **157**, 121–141 (2014).
43. Yu, J. et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70–78 (2017).
44. Zackular, J. P. et al. The gut microbiome modulates colon tumorigenesis. *mBio* **4**, e00692-13 (2013).
45. Wirbel, J. et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).
46. Sobhani, I. et al. Microbial dysbiosis in colorectal cancer (CRC) patients. *PLoS One* **6**, e16393 (2011).
47. Ren, Z. et al. Gut microbial profile analysis by MiSeq sequencing of pancreatic carcinoma patients in China. *Oncotarget* **8**, 95176–95191 (2017).
48. Ren, Z. et al. Gut microbiome analysis as a tool towards targeted non-invasive biomarkers for early hepatocellular carcinoma. *Gut* **58**, 1014–1023 (2019).
49. Pouncey, A. L. et al. Gut microbiota, chemotherapy and the host: the influence of the gut microbiota on cancer treatment. *Ecancermedicalscience* **12**, 868 (2018).
50. Alexander, J. L. et al. Gut microbiota modulation of chemotherapy efficacy and toxicity. *Nat. Rev. Gastroenterol. Hepatol.* **14**, 356–365 (2017).
51. Touchefeu, Y. et al. Systematic review: the role of the gut microbiota in chemotherapy- or radiation-induced gastrointestinal mucositis — current evidence and potential clinical applications. *Aliment. Pharmacol. Ther.* **40**, 409–421 (2014).
52. Mathijssen, R. H. et al. Clinical pharmacokinetics and metabolism of irinotecan (CPT-11). *Clin. Cancer Res.* **7**, 2182–2194 (2001).
53. Ma, M. K. & McLeod, H. L. Lessons learned from the irinotecan metabolic pathway. *Curr. Med. Chem.* **10**, 41–49 (2003).
54. Wallace, B. D. et al. Alleviating cancer drug toxicity by inhibiting a bacterial enzyme. *Science* **330**, 831–835 (2010).
55. Kodawara, T. et al. The inhibitory effect of ciprofloxacin on the beta-glucuronidase-mediated deconjugation of the irinotecan metabolite SN-38-G. *Basic Clin. Pharmacol. Toxicol.* **118**, 333–337 (2016).
56. Frank, M. et al. TLR signaling modulates side effects of anticancer therapy in the small intestine. *J. Immunol.* **194**, 1983–1995 (2015).
57. Hooper, L. V. & Macpherson, A. J. Immune adaptations that maintain homeostasis with the intestinal microbiota. *Nat. Rev. Immunol.* **10**, 159–169 (2010).
58. Quince, C. et al. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
59. Vetizou, M. et al. Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science* **350**, 1079–1084 (2015).
60. Frankel, A. E. et al. Metagenomic shotgun sequencing and unbiased metabolomic profiling identify specific human gut microbiota and metabolites associated with immune checkpoint therapy efficacy in melanoma patients. *Neoplasia* **19**, 848–855 (2017).
61. Gopalakrishnan, V. et al. Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science* **359**, 97–103 (2018).
62. Routy, B. et al. Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science* **359**, 91–97 (2018).
63. Sivan, A. et al. Commensal *Bifidobacterium* promotes antitumor immunity and facilitates anti-PD-L1 efficacy. *Science* **350**, 1084–1089 (2015).
64. Gerassy-Vainberg, S. et al. Radiation induces proinflammatory dysbiosis: transmission of inflammatory susceptibility by host cytokine induction. *Gut* **67**, 97–107 (2018).
65. Kumagai, T., Rahman, F. & Smith, A. M. The microbiome and radiation induced-bowel injury: evidence for potential mechanistic role in disease pathogenesis. *Nutrients* **10**, E1405 (2018).
66. Cui, M. et al. Faecal microbiota transplantation protects against radiation-induced toxicity. *EMBO Mol. Med.* **9**, 448–461 (2017).
67. Manichanh, C. et al. The gut microbiota predispose to the pathophysiology of acute postradiotherapy diarrhea. *Am. J. Gastroenterol.* **103**, 1754–1761 (2008).
68. Nam, Y. D. et al. Impact of pelvic radiotherapy on gut microbiota of gynecological cancer patients revealed by massive pyrosequencing. *PLoS One* **8**, e82659 (2013).
69. Wang, A. et al. Gut microbial dysbiosis may predict diarrhea and fatigue in patients undergoing pelvic cancer radiotherapy: a pilot study. *PLoS One* **10**, e0126312 (2015).
70. Reis Ferreira, M. et al. Microbiota- and radiotherapy-induced gastrointestinal side-effects (MARS) study: a large pilot study of the microbiome in acute and late-radiation enteropathy. *Clin. Cancer Res.* **25**, 6487–6500 (2019).
71. Lam, S. Y., Peppelenbosch, M. P. & Fuhler, G. M. Prediction and treatment of radiation enteropathy: can intestinal bugs lead the way? *Clin. Cancer Res.* **25**, 6280–6282 (2019).
72. Roy, S. & Trinchieri, G. Microbiota: a key orchestrator of cancer therapy. *Nat. Rev. Cancer.* **17**, 271–285 (2017).
73. Lehouritis, P. et al. Local bacteria affect the efficacy of chemotherapeutic drugs. *Sci. Rep.* **5**, 14554 (2015).
74. Viaud, S. et al. The intestinal microbiota modulates the anticancer immune effects of cyclophosphamide. *Science* **342**, 971–976 (2013).
75. Ghiringhelli, F. et al. Activation of the NLRP3 inflammasome in dendritic cells induces IL-1beta-dependent adaptive immunity against tumors. *Nat. Med.* **15**, 1170–1178 (2009).
76. Ozben, T. Oxidative stress and apoptosis: impact on cancer therapy. *J. Pharm. Sci.* **96**, 2181–2196 (2007).
77. Iida, N. et al. Commensal bacteria control cancer response to therapy by modulating the tumor microenvironment. *Science* **342**, 967–970 (2013).
78. Daillere, R. et al. *Enterococcus hirae* and *Barnesiella intestinihominis* facilitate cyclophosphamide-induced therapeutic immunomodulatory effects. *Immunity* **45**, 931–943 (2016).
79. Fyza, Y., Gills, J. & Sears, C. L. Impact of the microbiome on checkpoint inhibitor treatment in patients with non-small cell lung cancer and melanoma. *EBioMedicine* **48**, 642–647 (2019).
80. Seidel, J. A., Otsuka, A. & Kabashima, K. Anti-PD-1 and anti-CTLA-4 therapies in cancer: mechanisms of action, efficacy, and limitations. *Front. Oncol.* **8**, 86 (2018).
81. Darvin, P., Toor, S. M., Sasidharan Nair, V. & Elkord, E. Immune checkpoint inhibitors: recent progress and potential biomarkers. *Exp. Mol. Med.* **50**, 1–11 (2018).
82. Yang, B. et al. Progresses and perspectives of anti-PD-1/PD-L1 antibody therapy in head and neck cancers. *Front. Oncol.* **8**, 563 (2018).
83. Matson, V. et al. The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. *Science* **359**, 104–108 (2018).
84. Peled, J. U. et al. Microbiota predictor of mortality in allogeneic hematopoietic-cell transplantation. *N. Engl. J. Med.* **382**, 822–834 (2020).
85. Schwabe, R. F. & Jobin, C. The microbiome and cancer. *Nat. Rev. Cancer* **13**, 800–812 (2013).
86. Elinav, E. et al. The cancer microbiome. *Nat. Rev. Cancer.* **19**, 371–376 (2018).
87. de Martel, C. et al. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol.* **13**, 607–615 (2012).
88. Fais, T. et al. Targeting colorectal cancer-associated bacteria: a new area of research for personalized treatments. *Gut Microbes* **7**, 329–333 (2016).
89. Shah, M. S. et al. Leveraging sequence-based faecal microbial community survey data to identify a composite biomarker for colorectal cancer. *Gut* **67**, 882–891 (2018).
90. Armour, C. R., Nayfach, S., Pollard, K. S. & Sharpton, T. J. A metagenomic meta-analysis reveals functional signatures of health and disease in the human gut microbiome. *mSystems* **4**, e00332-18 (2019).
91. Segata, N. et al. Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).
92. Bhatt, A. S. et al. Sequence-based discovery of *Bradyrhizobium enterica* in cord colitis syndrome. *N. Engl. J. Med.* **369**, 517–528 (2013).
93. Kostic, A. D. et al. Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma. *Genome Res.* **22**, 292–298 (2012).
94. Drewes, J. L. et al. High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. *NPJ Biofilms Microbiomes* **3**, 34 (2017).
95. Thomas, A. M. et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).
96. Esteban-Gil, A. et al. ColPortal, an integrative multiomic platform for analysing epigenetic interactions in colorectal cancer. *Sci. Data* **6**, 255 (2019).
97. Derosa, L. et al. Negative association of antibiotics on clinical activity of immune checkpoint inhibitors in patients with advanced renal cell and non-small-cell lung cancer. *Ann. Oncol.* **29**, 1437–1444 (2018).
98. Li, Y., Wu, F. X. & Ngom, A. A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.* **19**, 325–340 (2018).
99. Alanazi, H. O., Abdullah, A. H. & Qureshi, K. N. A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. *J. Med. Syst.* **41**, 69 (2017).
100. Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. & Collins, J. J. Next-generation machine learning for biological networks. *Cell* **173**, 1581–1592 (2018).
101. Zhang, Y. et al. Machine learning performance in a microbial molecular autopsy context: a cross-sectional postmortem human population study. *PLoS One* **14**, e0213829 (2019).
102. Ruffle, J. K., Farmer, A. D. & Aziz, Q. Artificial intelligence-assisted gastroenterology — promises and pitfalls. *Am. J. Gastroenterol.* **114**, 422–428 (2019).
103. Saito, H. et al. Automatic detection and classification of protruding lesions in wireless capsule endoscopy images based on a deep convolutional neural network. *Gastrointest. Endosc.* **92**, 144–151 (2020).
104. Lui, T. K., Guo, C. G. & Leung, W. K. Accuracy of artificial intelligence on histology prediction and detection of colorectal polyps: a systematic review and meta-analysis. *Gastrointest. Endosc.* **92**, 11–22 (2020).
105. Seyed Tabib, N. S. et al. Big data in IBD: big progress for clinical practice. *Gut* https://doi.org/10.1136/gutjnl-2019-320065 (2020).
106. Olivera, P., Danese, S., Jay, N., Natoli, G. & Peyrin-Biroulet, L. Big data in IBD: a look into the future. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 312–321 (2019).
107. Noor, E., Cherkaoui, S. & Sauer, U. Biological insights through omics data integration. *Curr. Opin. Syst. Biol.* **15**, 39–47 (2019).
108. Lopez, C., Tcker, S., Salameh, T. & Tucker, C. An unsupervised machine learning method for discovering patient clusters based on genetic signatures. *J. Biomed. Inform.* **85**, 30–39 (2018).
109. Shomorony, I. et al. An unsupervised learning approach to identify novel signatures of health and disease from multimodal data. *Genome Med.* **12**, 7 (2020).
110. Price, N. D. et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nat. Biotechnol.* **35**, 747–756 (2017).
111. Argelaguet, R. et al. Multi-omics factor analysis — a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
112. Bisikirska, B. et al. Elucidation and pharmacological targeting of novel molecular drivers of follicular lymphoma progression. *Cancer Res.* **76**, 664–674 (2016).
113. Costello, J. C. et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **32**, 1202–1212 (2014).
114. Mezlini, A. M. & Goldenberg, A. Incorporating networks in a probabilistic graphical model to find drivers for complex human disease. *PLoS Comput. Biol.* **13**, e1005580 (2017).
115. Fabris, F., Magalhaes, J. P. & Freitas, A. A. A review of supervised machine learning applied to ageing research. *Biogerontology* **18**, 171–188 (2017).
116. Yu, Z. et al. Progressive semisupervised learning of multiple classifiers. *IEEE Trans. Cybern.* **48**, 689–702 (2018).
117. Huang, H., Vangay, P., McKinlay, C. E. & Knights, D. Multi-omics analysis of inflammatory bowel disease. *Immunol. Lett.* **162**, 62–68 (2014).
118. Lio, W. et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J. Med. Internet Res.* **18**, e323 (2016).

119. Doostparast Torshizi, A. & Petzold, L. R. Graph-based semi-supervised learning with genomic data integration using condition-responsive genes applied to phenotype classification. *J. Am. Med. Inform. Assoc.* **25**, 99–108 (2018).
120. Lin, Y. et al. Computer-aided biomarker discovery for precision medicine: data resources, models and applications. *Brief. Bioinform* **20**, 952–975 (2019).
121. Tang, B., Pan, Z., Yin, K. & Khateeb, A. Recent advances of deep learning in bioinformatics and computational biology. *Front. Genet.* **10**, 214 (2019).
122. Londhe, V. Y. & Bhasin, B. Artificial intelligence and its potential in oncology. *Drug. Discov. Today* **24**, 228–232 (2019).
123. Ngiam, K. Y. & Khor, I. W. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol.* **20**, e262–e273 (2019).
124. Babarenda Gamage, T. P. et al. An automated computational biomechanics workflow for improving breast cancer diagnosis and treatment. *Interface Focus.* **9**, 20190034 (2019).
125. Tseng, Y. J. et al. Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. *Int. J. Med. Inform.* **128**, 79–86 (2019).
126. Goldenberg, S. L., Nir, G. & Salcudean, S. E. A new era: artificial intelligence and machine learning in prostate cancer. *Nat. Rev. Urol.* **16**, 391–403 (2019).
127. Paik, E. S. et al. Prediction of survival outcomes in patients with epithelial ovarian cancer using machine learning methods. *J. Gynecol. Oncol.* **30**, e65 (2019).
128. Kouznetsova, V. L. et al. Recognition of early and late stages of bladder cancer using metabolites and machine learning. *Metabolomics* **15**, 94 (2019).
129. Jin, Y. et al. The diversity of gut microbiome is associated with favorable responses to anti-PD-1 immunotherapy in Chinese non-small cell lung cancer patients. *J. Thorac. Oncol.* **14**, 1378–1389 (2019).
130. Qian, Z. et al. Differentiation of glioblastoma from solitary brain metastases using radiomic machine-learning classifiers. *Cancer Lett.* **451**, 128–135 (2019).
131. Leatherdale, S. T. & Lee, J. Artificial intelligence (AI) and cancer prevention: the potential application of AI in cancer control programming needs to be explored in population laboratories such as COMPASS. *Cancer Causes Control.* **30**, 671–675 (2019).
132. Veselkov, K. et al. HyperFoods: machine intelligent mapping of cancer-beating molecules in foods. *Sci. Rep.* **9**, 9237 (2019).
133. Zhao, W. et al. Toward automatic prediction of EGFR mutation status in pulmonary adenocarcinoma with 3D deep learning. *Cancer Med.* **8**, 3532–3543 (2019).
134. Sato, M. et al. Machine-learning approach for the development of a novel predictive model for the diagnosis of hepatocellular carcinoma. *Sci. Rep.* **9**, 7704 (2019).
135. Feng, Q. X. et al. An intelligent clinical decision support system for preoperative prediction of lymph node metastasis in gastric cancer. *J. Am. Coll. Radiol.* **16**, 952–960 (2019).
136. You, J., McLeod, R. D. & Hu, P. Predicting drug–target interaction network using deep learning model. *Comput. Biol. Chem.* **80**, 90–101 (2019).
137. Kessler, R. C., Bossarte, R. M., Luedtke, A., Zaslavsky, A. M. & Zubizarreta, J. R. Machine learning methods for developing precision treatment rules with observational data. *Behav. Res. Ther.* **120**, 103412 (2019).
138. Mottini, C., Napolitano, F., Li, Z., Gao, X. & Cardone, L. Computer-aided drug repurposing for cancer therapy: approaches and opportunities to challenge anticancer targets. *Semin. Cancer Biol.* https://doi.org/10.1016/j.semcancer.2019.09.023 (2019).
139. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **16**, 703–715 (2019).
140. Penson, A. et al. Development of genome-derived tumor type prediction to inform clinical cancer care. *JAMA Oncol.* https://doi.org/10.1001/jamaoncol.2019.3985 (2019).
141. Grewal, J. K. et al. Application of a neural network whole transcriptome-based pan-cancer method for diagnosis of primary and metastatic cancers. *JAMA Netw. Open* **2**, e192597 (2019).
142. Subramanian, S. et al. Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* **510**, 417–421 (2014).
143. Vervier, K. et al. Large-scale machine learning for metagenomics sequence classification. *Bioinformatics* **32**, 1023–1032 (2016).
144. Fernandez-Navarro, T. et al. Exploring the interactions between serum free fatty acids and fecal microbiota in obesity through a machine learning algorithm. *Food Res. Int.* **121**, 533–541 (2019).
145. Thompson, J. et al. Machine learning to predict microbial community functions: an analysis of dissolved organic carbon from litter decomposition. *PLoS One* **14**, e0215502 (2019).
146. Shinn, L. et al. Applying machine-learning to human gastrointestinal microbial species to predict dietary intake. *Curr. Dev. Nutr.* **3**, https://doi.org/10.1093/cdn/nzz040.P20-040-19 (2019).
147. Turnbaugh, P. J. et al. The human microbiome project. *Nature* **449**, 804–810 (2007).
148. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project. *Nature* **569**, 641-648.
149. Vangay, P., Hillmann, B. M. & Knights, D. Microbiome learning repo (ML Repo): a public repository of microbiome regression and classification tasks. *Gigascience* **8**, giz042 (2019).
150. Mallick., H. et al. Experimental design and quantitative analysis of microbial community multiomics. *Genome Biol.* **18**, 228 (2017).
151. Knight, R. et al. Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* **16**, 410–422 (2018).
152. Cani, P. D. Human gut microbiome: hopes, threats and promises. *Gut* **67**, 1716–1725 (2018).
153. Knights, D., Costello, E. K. & Knight, R. Supervised classification of human microbiota. *FEMS Microbiol. Rev.* **35**, 343–359 (2011).
154. Moitinho-Silva, L. et al. Predicting the HMA–LMA status in marine sponges by machine learning. *Front. Microbiol.* **8**, 752 (2017).
155. Bockulich, N. A. et al. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl Med.* **8**, 343ra82 (2016).
156. Pasolli, E., Truong, D. T., Malik, F., Waldron, L. & Segata, N. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* **12**, e1004977 (2016).
157. Heshiki, Y. et al. Predictable modulation of cancer treatment outcomes by the gut microbiota. *Microbiome* **8**, 28 (2020).
158. Zuo, T. et al. Gut mucosal virome alterations in ulcerative colitis. *Gut* **68**, 1169–1179 (2019).
159. Larsen, P. E. & Dai, Y. Metabolome of human gut microbiome is predictive of host dysbiosis. *Gigascience* **4**, 42 (2015).
160. Bokulich, N. et al. q2-sample-classifier: machine-learning tools for microbiome classification and regression. *J. Open Source Softw.* **3**, 934 (2018).
161. Dhariwal, A. et al. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.* **45**, W180–W188 (2017).
162. Edgar, R. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
163. Prifti, E. et al. Interpretable and accurate prediction models for metagenomics data. *Gigascience* **9**, 1–11 (2020).
164. Zhou, Y.-H. & Gallins, P. A review and tutorial of machine learning methods for microbiome host trait prediction. *Front. Genet.* **10**, 579 (2019).
165. Ananthakrishnan, A. et al. Gut microbiome function predicts response to anti-integrin biologic therapy in inflammatory bowel diseases. *Cell Host Microbe* **21**, 603–610 (2017).
166. Zhu, Q., Jiang, X., Zhu, Q., Pan, M. & He, T. Graph embedding deep learning guides microbial biomarkers' identification. *Front. Genet.* **10**, 1182 (2019).
167. Angermueller, C., Parnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **12**, 878 (2016).
168. Eraslan, G., Avsec, Z., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* **20**, 389–403 (2019).
169. LaPierre, N., Ju, C. J., Zhou, G. & Wang, W. MetaPheno: a critical evaluation of deep learning and machine learning in metagenome-based disease prediction. *Methods* **166**, 74–82 (2019).
170. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
171. Stols-Goncalves, D. et al. Epigenetic markers and mMicrobiota/mMetabolite-induced epigenetic modifications in the pathogenesis of obesity, metabolic syndrome, type 2 diabetes, and non-alcoholic fatty liver disease. *Curr. Diab. Rep.* **19**, 31 (2019).
172. Palsson, B. & Zengler, K. The challenges of integrating multi-omic data sets. *Nat. Chem. Biol.* **6**, 787–789 (2010).
173. Zitnik, M. et al. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. *Inf. Fusion.* **50**, 71–91 (2019).
174. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
175. Perkins, B. A. et al. Precision medicine screening using whole-genome sequencing and advanced imaging to identify disease risk in adults. *Proc. Natl Acad. Sci. USA* **115**, 3685–3691 (2018).
176. Hollister, E. B. et al. Leveraging human microbiome features to diagnose and stratify children with irritable bowel syndrome. *J. Mol. Diagn.* **21**, 449–461 (2019).
177. Kreznar, J. H. et al. Host genotype and gut microbiome modulate insulin secretion and diet-induced metabolic phenotypes. *Cell Rep.* **18**, 1739–1750 (2017).
178. Schussler-Fiorenza Rose, S. M. et al. A longitudinal big data approach for precision health. *Nat. Med.* **25**, 792–804 (2019).
179. Flemer, B. et al. The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* **67**, 1454–1463 (2018).
180. Zackular, J. P. et al. The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prev. Res.* **7**, 1112–1121 (2014).
181. Imhann, F. et al. The 1000IBD project: multi-omics data of 1000 inflammatory bowel disease patients; data release 1. *BMC Gastroenterol.* **19**, 5 (2019).
182. Casals-Pascual, C. et al. Microbial diversity in clinical microbiome studies: sample size and statistical power considerations. *Gastroenterology* **158**, 1524–1528 (2020).
183. Shenoi, S.J., Ly, V., Soni, S. & Roberts, K. Developing a serach engine for precision medicine. *AMIA Jt. Summits Transl Sci. Proc.* **2020**, 579–588 (2020).
184. Goecks, J., Jalili, V., Heiser, L. M. & Gray, J. W. How machine learning will transform biomedicine. *Cell* **181**, 92–101 (2020).
185. Zheng, Y. et al. Specific gut microbiome signature predicts the early-stage lung cancer. *Gut Microbes* **11**, 1–12 (2020).
186. Zeevi, D. et al. Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1094 (2015).
187. Gevers, D. et al. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382–392 (2014).
188. Franzosa, E. A. et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2019).
189. Coburn, B. et al. Lung microbiota across age and disease stage in cystic fibrosis. *Sci. Rep.* **5**, 10241 (2015).
190. Alekseyenko, A. V. et al. Community differentiation of the cutaneous microbiota in psoriasis. *Microbiome* **1**, 31 (2013).
191. Wu, H. et al. Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. *Nat. Med.* **23**, 850–858 (2017).
192. Forslund, K. et al. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature* **528**, 262–266 (2015).
193. Haiser, H. J. et al. Predicting and manipulating cardiac drug inactivation by the human gut bacterium *Eggerthella lenta*. *Science* **341**, 295–298 (2013).
194. Cussotto, S., Clarke, G., Dinan, T. G. & Cryan, J. F. Psychotropics and the microbiome: a chamber of secrets. *Psychopharmacology* **236**, 1411–1432 (2019).
195. Claesson, M. J. et al. Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Natl Acad. Sci. USA* **108**, 4586–4591 (2011).
196. Bibbo, S. et al. The role of diet on gut microbiota composition. *Eur. Rev. Med. Pharmacol. Sci.* **20**, 4742–4749 (2016).
197. Qiu, G. et al. The significance of probiotics in preventing radiotherapy-induced diarrhea in patients with cervical cancer: a systematic review and meta-analysis. *Int. J. Surg.* **65**, 61–69 (2019).
198. Liu, M. M. et al. Probiotics for prevention of radiation-induced diarrhea: a meta-analysis of randomized controlled trials. *PLoS One* **12**, e0178870 (2017).

199. Wang, Y. H. et al. The efficacy and safety of probiotics for prevention of chemoradiotherapy-induced diarrhea in people with abdominal and pelvic cancer: a systematic review and meta-analysis. *Eur. J. Clin. Nutr.* **70**, 1246–1253 (2016).
200. Delia, P. et al. Use of probiotics for prevention of radiation-induced diarrhea. *World J. Gastroenterol.* **13**, 912–915 (2007).
201. Henson, C. C. et al. Nutritional interventions for reducing gastrointestinal toxicity in adults undergoing radical pelvic radiotherapy. *Cochrane Database Syst. Rev.* **11**, CD009896 (2013).
202. Reyna-Figueroa, J. et al. Probiotic supplementation decreases chemotherapy-induced gastrointestinal side effects in patients with acute leukemia. *J. Pediatr. Hematol. Oncol.* **41**, 468–472 (2019).
203. Osterlund, P. et al. *Lactobacillus* supplementation for diarrhoea related to chemotherapy of colorectal cancer: a randomised study. *Br. J. Cancer* **97**, 1028–1034 (2007).
204. Wada, M. et al. Effects of the enteral administration of *Bifidobacterium breve* on patients undergoing chemotherapy for pediatric malignancies. *Support. Care Cancer* **18**, 751–759 (2010).
205. Tian, Y. et al. Effects of probiotics on chemotherapy in patients with lung cancer. *Oncol. Lett.* **17**, 2836–2848 (2019).
206. Mego, M. et al. Prevention of irinotecan induced diarrhea by probiotics: a randomized double blind, placebo controlled pilot study. *Complement. Ther. Med.* **23**, 356–432 (2015).
207. Chitapanarux, I. et al. Randomized controlled trial of live *Lactobacillus acidophilus* plus *Bifidobacterium bifidum* in prophylaxis of diarrhea during radiotherapy in cervical cancer patients. *Radiat. Oncol.* **5**, 31 (2010).
208. Wei, D. et al. Probiotics for the prevention or treatment of chemotherapy- or radiotherapy-related diarrhoea in people with cancer. *Cochrane Database Syst. Rev.* **8**, CD008831 (2018).
209. Jonasch, E. et al. Phase II study of two weeks on, one week off sunitinib scheduling in patients with metastatic renal cell carcinoma. *J. Clin. Oncol.* **36**, 1588–1593 (2018).
210. Andreyev, J. et al. Guidance on the management of diarrhoea during cancer chemotherapy. *Lancet Oncol.* **15**, e447–e460 (2014).
211. Wang, Y. et al. Fecal microbiota transplantation for refractory immune checkpoint inhibitor-associated colitis. *Nat. Med.* **24**, 1804–1808 (2018).
212. Cammarota, G. et al. Fecal microbiota transplantation: a new old kid on the block for the management of gut microbiota-related disease. *J. Clin. Gastroenterol.* **48**, S80–S84 (2014).
213. O'Toole, P. W., Marchesi, J. R. & Hill, C. Next-generation probiotics: the spectrum from probiotics to live biotherapeutics. *Nat. Microbiol.* **2**, 17057 (2017).
214. Song, H. et al. Synthetic microbial consortia: from systematic analysis to construction and applications. *Chem. Soc. Rev.* **43**, 6954–6981 (2014).
215. Yuvaraj, S. et al. *E. coli*-produced BMP-2 as a chemopreventive strategy for colon cancer: a proof-of-concept study. *Gastroenterol. Res. Pract.* **2012**, 895462 (2012).
216. Huibregtse, I. L. et al. Genetically modified *Lactococcus lactis* for delivery of human interleukin-10 to dendritic cells. *Gastroenterol. Res. Pract.* **2012**, 639291 (2012).
217. Pellegrini, M. et al. Gut microbiota composition after diet and probiotics in overweight breast cancer survivors: a randomized open-label pilot intervention trial. *Nutrition* **74**, 110749 (2020).
218. Moore, J. H. et al. Preparing next-generation scientists for biomedical big data: artificial intelligence approaches. *Per. Med.* **16**, 247–257 (2019).
219. Buruk, B., Ekmekci, P. E. & Arda, B. A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Med. Health Care Philos.* https://doi.org/10.1007/s11019-020-09948-1 (2020).
220. Price, W. N. II & Cohen, I. G. Privacy in the age of medical big data. *Nat. Med.* **25**, 37–43 (2019).
221. van den Bogert, B., Boekhorst, J., Provano, W. & May, A. On the role of bioinformatics and data science in industrial applications. *Front. Genet.* **10**, 721 (2019).
222. Wang, Y. & Qian, P. Y. Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLoS One* **4**, e7401 (2009).
223. Budding, A. E. et al. Automated broad-range molecular detection of bacteria in clinical samples. *J. Clin. Microbiol.* **54**, 934–943 (2016).
224. Franzosa, E. A. et al. Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl Acad. Sci. USA* **111**, e2329–e2338 (2014).
225. Jin, P. et al. Mining the fecal proteome: from biomarkers to personalised medicine. *Expert. Rev. Proteom.* **14**, 445–459 (2017).
226. Daliri, E. B. et al. The human microbiome and metabolomics: current concepts and applications. *Crit. Rev. Food Sci. Nutr.* **57**, 3565–3576 (2017).
227. Ranjan, R., Rani, A., Metwally, A., McGee, H. S. & Perkins, D. L. Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem. Biophys. Res. Commun.* **469**, 967–977 (2016).
228. Vuik, F. et al. Composition of the mucosa-associated microbiota along the entire gastrointestinal tract of human individuals. *United European Gastroenterol. J.* **7**, 897–907 (2019).
229. Li, S. et al. Pancreatic cyst fluid harbors a unique microbiome. *Microbiome* **5**, 147 (2016).

Acknowledgements
This publication has in part emanated from research conducted with the financial support of AIRC Foundation for Cancer Research (AIRC IG grant number 18599, MFAG grant number 23681) and Science Foundation Ireland (grant number SFI/12/RC/2273).

Author contributions
The authors contributed equally to all aspects of the article.

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Gastroenterology & Hepatology thanks M. Peppelenbosch and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information
Supplementary information is available for this paper at https://doi.org/10.1038/s41575-020-0327-3.

RELATED LINKS
ColPortal: https://colportal.imib.es
GenBank and Gene Expression Omnibus: https://www.ncbi.nlm.nih.gov/gds
IBD Multi-omics Database: http://ibdmdb.org/
Microbiome Learning Repo: https://knights-lab.github.io/MLRepo/
ML4Microbiome COST Action: https://www.cost.eu/actions/CA18131
ONCOBIOME Project: https://cordis.europa.eu/project/id/825410/it