ELSEVIER

Teaser The discovery of reproducible biomarkers is the foremost challenge in precision medicine. The adaptation of appropriate unbiased data analytic strategies could be the key to solving this problem.

Unbiased data analytic strategies to improve biomarker discovery in precision medicine

Saifur R. Khan[1,2], Yousef Manialawy[1,2], Michael B. Wheeler[1,2,‡] and Brian J. Cox[3,4,‡]

[1] Endocrine and Diabetes Platform, Department of Physiology, University of Toronto, Medical Sciences Building, Room 3352, 1 King's College Circle, Toronto, ON M5S 1A8, Canada
[2] Advanced Diagnostics, Metabolism, Toronto General Hospital Research Institute, Toronto, ON, Canada
[3] Reproduction and Development Platform, Department of Physiology, University of Toronto, Medical Sciences Building, Room 3360, 1 King's College Circle, Toronto, ON M5S 1A8, Canada
[4] Department of Obstetrics and Gynecology, University of Toronto, Toronto, ON, Canada

Omics technologies promised improved biomarker discovery for precision medicine. The foremost problem of discovered biomarkers is irreproducibility between patient cohorts. From a data analytics perspective, the main reason for these failures is bias in statistical approaches and overfitting resulting from batch effects and confounding factors. The keys to reproducible biomarker discovery are: proper study design, unbiased data preprocessing and quality control analyses, and a knowledgeable application of statistics and machine learning algorithms. In this review, we discuss study design and analysis considerations and suggest standards from an expert point-of-view to promote unbiased decision-making in biomarker discovery in precision medicine.

Introduction

Biomarkers are measurable biomolecules (i.e., genes, proteins and metabolites) that can classify or identify patients from the healthy population [1]. Given a set of patients, a biomarker can identify those with a malignant versus benign tumor. Additionally, a biomarker could identify those cancer patients that best respond to a particular treatment [2]. Biomarkers can be most effective when used in a combination, termed a signature [3]. Driving efforts in biomarker discovery is the advancement in high-throughput techniques (i.e., omics), which has made it possible to acquire massive amounts of biological data for making data-driven health decisions via a precision medicine approach. Despite the promise of big datasets, as of 2018 only five biomarkers had been discovered and validated in humans (https://www.fda.gov), which pales in comparison to the roughly 4000 diseases where a molecular basis has been well-established [4]. As such, there is a desire to increase the number of validated biomarkers and their clinical uses in the interests of industries and policymakers. An often noted failing in biomarker discovery is a failure to hold up

Corresponding authors: Khan, S.R. (Saifur.khan@utoronto.ca), Cox, B.J. (b.cox@utoronto.ca)
‡ Michael B. Wheeler and Brian J. Cox are joint senior authors.

Saifur R. Khan is currently a Diabetes Canada postdoctoral fellow in the Department of Physiology, Faculty of Medicine, University of Toronto. His current research focuses on the discovery of prognostic and novel drug targets for type 2 diabetes using big data analytics and systems biology, followed by translational research. He earned his PhD in pharmaceutical sciences at the University of Alberta, where he studied isoniazid, a first-line antituberculosis drug, using omics. His work identified three pharmacological functions of isoniazid, for which he proposed a new mode of action. Previously, he was a senior scientist at Incepta Pharmaceuticals Ltd. where he formulated many innovative drug dosage forms.

Michael B. Wheeler is a full professor in the Faculty of Medicine, Departments of Physiology and Medicine at the University of Toronto. He is also a Senior Scientist at the University Health Network, Toronto General Hospital Research Institute in the Advanced Diagnostics Division. His research is focused on developing novel strategies to treat type 1 and type 2 diabetes using multidisciplinary approaches, which combine information gained from human omics studies, genetic models of diabetes, molecular biology and cell biology.

Brian J. Cox completed a Master's degree in applied biochemistry and a PhD and postdoctoral fellowship in molecular genetics. He has worked in the biotechnology industry (Affinium Pharmaceuticals) and federal government (Health Canada, Environment Canada) on projects spanning atmospheric chemistry, estrogen receptor biochemistry, mass-spectrometry-based proteomics and mouse genetic models. Currently, he is an associate professor in the departments of Physiology, and Obstetrics and Gynecology at the University of Toronto, Canada. He holds a Canada Research Tier II Chair in fetal and maternal health. His research group is focused on trophoblast development and placental pathologies. Projects on trophoblast development use a variety of genome-wide and systems biology approaches to identify the genetic regulatory mechanisms of cell fate specification and differentiation. Projects on placental pathology are using modeling and class discovery approaches on large-scale datasets of human-patient samples to develop patient stratification methods.

https://doi.org/10.1016/j.drudis.2019.05.018

GLOSSARY

Attribute A characteristic of a subjects, also known as variable in a dataset. Sometimes it has been described here as a factor
Confounding factors A factor that is not considered in a study has an influence that can affect the decision or result
Cohort A group of people under study
Stringent data Data that are acquired using very strict rules
Eigenvectors Appearance of new collective vectors after linear transformation of data. In data visualization statistics, these vectors are known as component. For example, in principal component analysis they are known as principal components
Dimension reduction A process of reducing the number of random variables (Wikipedia)
Probabilistic A nonconstructive (theoretical but without realistic proof) statistical method
Class Here class means study group. For example, case and control are two groups of classes
Class separation Statistically, the difference between two groups is represented as class separation
Algorithm Statistically, algorithm means rule
Imputation An estimation of the value based on the other available data
AUC Area under curve
Accuracy The ability of a test to differentiate between patient and healthy is known as test accuracy. It is calculated as a ratio of total correctly identified patients and healthy people against to total population
Sensitivity The ability of a test to identify patient correctly is known as test sensitivity. It is calculated as a ratio of total correctly identified patients against to total patients
Specificity The ability of a test to identify healthy people correctly is known as test specificity. It is calculated as a ratio of total correctly identified healthy people against total healthy people

to independent verification. This review will focus on how improvements in data analytic approaches could identify biomarkers with more-consistent behavior.

To establish a biomarker or signature as a diagnostic it must pass multiple phases of rigorous testing (Fig. 1). Phase 1 entails shortlisting potential biomarker candidates using unbiased methodologies and is often the most important stage in successful biomarker discovery. Statistically, a biased approach is an approach where a variable is selected without proper randomization. This could cause the selection of a variable that is not truly representative of the targeted population. By contrast, an unbiased approach where a proper randomization protocol is adopted means that the selected variable has a very good chance of being accurately representative of the targeted population. This unbiased candidate selection is heavily dependent on the proper use of analytic tools and a thorough understanding of the most applicable statistical methods. Generally, a few thousand candidates are initially evaluated in phase 1 to yield an optimized panel of biomarkers or a signature, which is usually composed of a small set of candidates. Given that biological-sample analysis-techniques (i.e., omics approaches) in phase 1 are usually semiquantitative in nature, they rely on an approximation or relative quantification against a set of standards. Candidates must be further refined using a robust quantitative method in phase 2, also known as the 'method development' phase. Once the most robust evaluation methods for each candidate are determined, the finalized biomarker or signature is further investigated over different cohorts and different time intervals in phase 3. The different phases of biomarker discovery have been illustrated in Fig. 1.

The key reasons underlying the lack of validated biomarkers all arise in phase 1, thereby detracting from the efficacy of downstream analyses. These include: (i) low statistical power, which fails to properly address high individual variabilities; (ii) a lack of absolute biological data quantification methods; and (iii) gaps in the understanding of data analytics, particularly in their application, preprocessing, optimization and interpretation. Generally, the inclusion of more-diverse participants helps to address the problem of low statistical power of a study [5]. In addition, methods for acquiring big biological data are always improving gradually with regard to quantification process and volume. As such, pharmaceutical giants have shown interest in precision medicine approaches for their future discovery projects. Pfizer is employing machine learning data analytics – the famous IBM Watson – for the discovery of novel immune-oncology drugs. Sanofi is using another data analytic AI platform to discover novel therapies for metabolic diseases [6]. Additionally, government funding for precision medicine has been growing exponentially in virtually every research-intensive country including the USA, Canada and the UK. However, there is a clear shortage of skilled biological scientists who are needed to play a significant part in study design, big data analysis and interpretations, as well as in decision-making. This shortage can mainly be attributed to the lack of coding skills among biologists (e.g., R, Python, SAS, etc.) needed to effectively work with rapidly evolving data analytics strategies within the health sector. However, extensive coding knowledge is becoming less necessary with the advent of graphical user interface (GUI) software (e.g., Metaboanalyst, Weka, etc.) and the wide availability of source codes for various functions in R and other languages. Therefore, it is time for biological scientists to train in the effective implementation of data analytics.

A survey of published manuscripts within the biomarker discovery field reveals several common shortcomings: choice of cohort and study design is often arbitrary, assessment of dataset quality might be lacking or entirely absent and the selection of a particular validation protocol might be unjustified, among other issues. As a result, phase 1 biomarkers are mostly inconsistent and irreproducible in class separation from cohort to cohort. Some stakeholders argue that this issue largely stems from the lack of a standard protocol for biomarker discovery; because of this, industries have pushed regulatory authorities (e.g., FDA) to standardize the process. However, given the wide variety in biomarker studies and resource availability, a one-size-fits-all regulatory protocol would probably hinder many studies by overlooking study-specific considerations. Promoting comprehensive knowledge on statistics, algorithms (see Glossary) and the biological field of study would probably prove to be the better approach to allow flexibility. Bearing these factors in mind, in this review, we aim to characterize an unbiased data analytics approach that covers key knowledge and considerations in phase 1 biomarker discovery for precision medicine. This would be expected to aid in unbiased decision-making, which is a prerequisite for successful biomarker discovery.

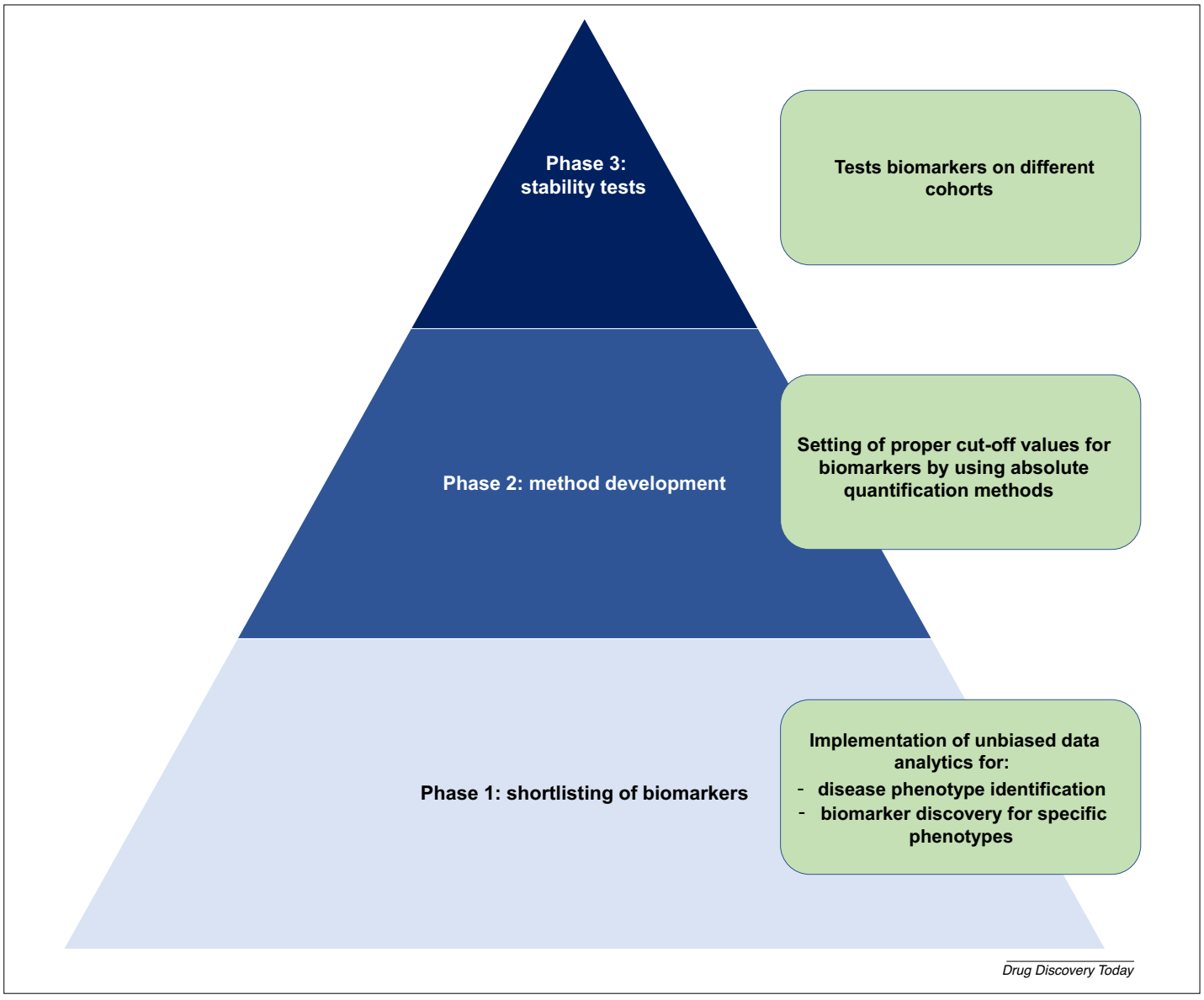

FIGURE 1

The different phases of biomarker discovery in precision medicine. In phase 1, big health data are explored using unbiased data analytics for disease phenotype identification and phenotype-specific biomarker discovery where top analytes are shortlisted. In phase 2, the cut-off value for classifying a disease state vs a healthy state is readjusted by using robust absolute quantification methods. This stage is also known as the method development phase. Subsequently, the stability of shortlisted analytes as biomarkers is reassessed on different cohorts with a larger population.

Precision medicine for biologically heterogenous diseases

Precision medicine is an initiative where disease treatments and prevention strategies are adapted based on individual variability in genes, environment and lifestyle [7]. These variations create different disease phenotypes, also termed as 'heterogenicity' in disease [8]. Heterogenicity has hindered efforts to identify biomarkers for the characterization of common disease risk as well as treatment outcomes. High heterogenicity is most notably observed in cancer, metabolic diseases and neurodegenerative diseases on account of significant genetic, environmental and lifestyle variabilities within patient populations. This makes such diseases ideal candidates for precision medicine, because different levels of stratification can be adopted to reduce their heterogenicity or dimensions.

In 2016, the Precision Medicine Initiative (PMI) was launched by the National Institutes of Health (NIH) to address the heterogenicity of many diseases for which we do not have a proven means of prevention or effective treatment (https://allofus.nih.gov). The PMI invested US$130 million to garner insight into the biological, environmental and behavioral influences that underlie intractable diseases. Approximately US$70 million was allocated for cancer research through the National Cancer Institute (NCI) to 'scale up efforts to identify genomic drivers in cancer and apply that knowledge in the development of more effective approaches to cancer treatment' [9]. Although this initiative has accelerated biomarker discovery to guide prognostic and therapeutic decisions of these heterogeneous diseases, only a handful of biomarkers has been successfully translated into routine clinical practice in light of

irreproducibility of most candidates in later studies. This irreproducibility is commonly attributed to problematic study design, issues with assay platforms and unavailability of specimens [10–12]. Many issues of biomarker discovery can be improved through better implementation of data analytics, and yet this area has received little attention overall. The focus of the remaining review will be on three aspects that influence data analytics, experimental design (patients, analytical platforms) and data processing. An emphasis will be placed on consideration of chemometric methods.

Considerations for biomarker discovery

A high-quality dataset is the essential prerequisite for data analytics in phase 1 of biomarker discovery in precision medicine. The acquisition of stringent raw data from a biological sample in an omics-derived high-throughput experiment is the starting point of this process. The schematic flow diagram of biomarker discovery for phase 1 is illustrated in Fig. 2. The raw data are subject to preprocessing which includes estimation or exclusion of missing data, filtering and normalization. This preprocessing converts raw data into final data. The final dataset must then be scrutinized for its qualities through chemometric analysis. In the chemometric analysis, dataset qualities are assessed for the presence of confounding factors, the existence of class separation and the presence of technical errors or batch effects. All of this information can help predict how successful the discovered biomarker will be, and where adjustments might be needed. Each of these steps is discussed in the corresponding sections of this review. However, before discussing biomarker discovery phase 1 in detail, we must first address important considerations in ensuring an unbiased approach.

Study design: a crucial matter

Study design plays a vital part in eliminating bias from the dataset. In a population study, pair-matching for clinical variables such as age, sex, race and ethnicity is considered a very effective strategy to

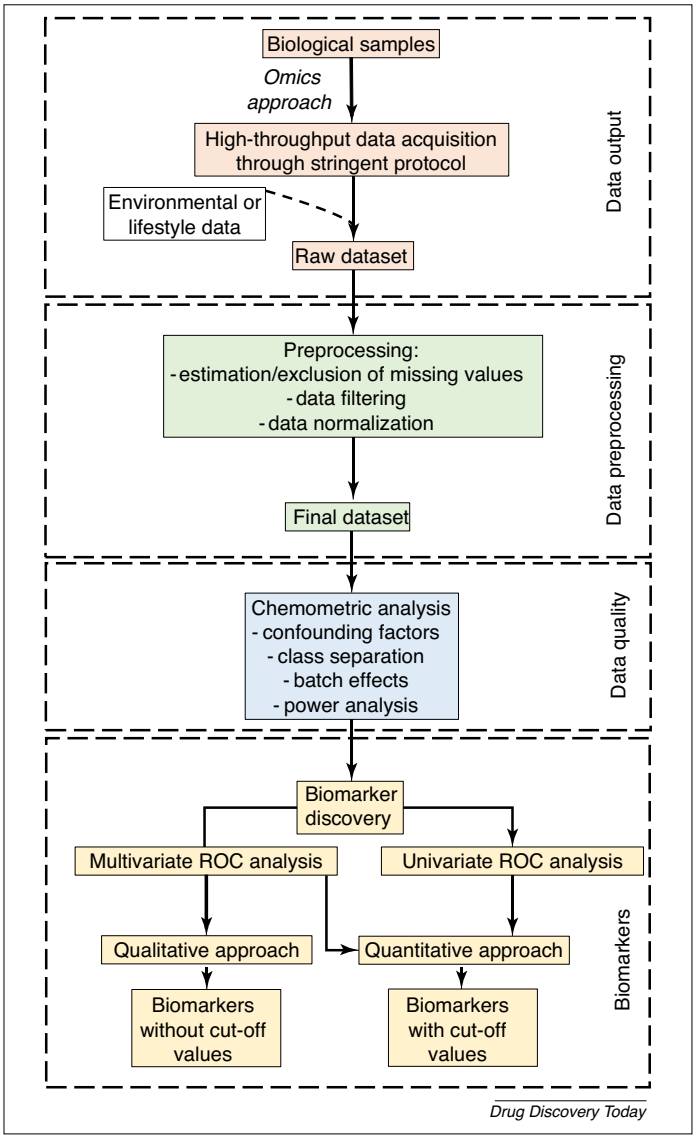

FIGURE 2

A flow diagram of unbiased data analytics for biomarker discovery in phase 1. It includes data output, data preprocessing, data quality assessment and use of algorithms for top candidate selection as potential biomarkers.

minimize the influence of many confounding factors [13–15]. However, over-matching (which is described as a matching scenario where case and control are matched based on an overly strong determinant of the target biomarker profile) could result in loss of a strong biomarker from the study and could decrease the discrimination power of biomarkers [16]. For example, the Body Mass Index (BMI) is a strong determinant in obesity-related metabolic disorders. Therefore, it is not recommended to neutralize the discriminating power of the BMI via a pair-matching strategy in studies related to obesity-induced metabolic diseases.

The selection of a cohort is another important consideration for biomarker discovery from observational studies; selection strategies vary with study design (i.e., prospective study, retrospective study or cross-sectional study). Although prospective and retrospective studies are longitudinal studies, prospective studies follow the group of subjects over time for a given study factor whereas retrospective studies evaluate the historical data of the group of subjects for that study factor [17]. In longitudinal studies, a population is observed for an age range for a study factor (e.g., obesity in diabetes development) to be concluded (e.g., whether obese people develop diabetes earlier than non-obese people). These kinds of longitudinal studies can discover predictive biomarkers or prognostics [18]. By contrast, cross-sectional studies evaluate group data for a given study factor at a specific time-point. This is done to rapidly approximate a longitudinal study. However, the longitudinal study is the actual measurement over time of the same people so the inference is direct. Cross-sectional studies are more applicable for biomarker discovery of overt disease conditions [19]. In all cases, there is a need for balance between selecting a high-risk and no-risk cohort. The use of a particular cohort with a high risk of developing the disease to discover a biomarker is ultimately less representative of the general population [16]. In predictive biomarker discovery, it is important to design a study at a very early disease stage, where the class separation between case and control is very subtle or closely overlapped. The selection of a no-risk cohort is also a poor representation of the general population. The statistically best approach is to design a study that reflects the actual disease percentage of the whole population or a particular subpopulation. For example, if one disease has a 10% prevalence in the whole population the ideal pair-matched case-control study should match nine controls for each case. However, this kind of study design might not prove to be practical owing to patient availability. Taking this into account, the study design should reflect the actual disease percentage as closely as possible.

Biological specimen quality: imperative for success in biomarker discovery

The quality of biological specimens is impacted by collection processes, storage conditions, transportation and repeated use, which all greatly influence reproducibility of biomarker discovery. Biospecimen collection from participants of cohort studies should be performed under a standard protocol and with trained personnel to minimize the undesired influence of confounding factors. For example, overnight fasting is recommended before blood and urine collection to avoid the effects of food intake. The biospecimens should also be collected in specialized tubes that can ensure the preservation of bioanalytes. These tubes must then be stored in a well-designed biobank or biorepository to ensure the stability of bioanalytes for future bioanalytical studies. The recommended temperature of the biobank could range from −20 °C to −70 °C; however, most clinical settings where biospecimens are collected lack such a facility. Therefore, an appropriate storage system is needed in the clinic permanently or temporarily to hold biospecimens while collection is ongoing. Temperature control is also necessary while biospecimens are being transported. At every step, appropriate labeling of tubes is important for reliable retrieval of participant information [20,21]. In addition, the repeated usage of samples that require repeated freezing and thawing can cause degradation of some analytes. This element can result in an inherent batch effect. To understand the trend of degradation owing to repeated freeze–thaw cycles, a pilot study can be conducted using several time samplings through several freeze–thaw cycles. It can identify the freeze–thaw-cycle-liable analytes and can be used for batch effect correction (described below).

Which omics is the best?

Biomarkers can be any biomolecule (such as genes, proteins, microRNA and metabolites) that indicate any specific alteration in biology owing to a disease condition or drug toxicity. The disease manifests as a complex interaction between biology, the environment, genes and mutations whereby the body produces an observable response [22]. Similarly, body responses can be observed under drug-induced toxicity. In chronic diseases, such responses could be started earlier. The question is: which class of biomolecule can best represent such an alteration or response? In the case of gene and protein expression, several robust high-throughput analytical platforms are available. However, transcribed mRNA can be muted by microRNA or alternatively spliced and translated into a different protein isoform. After translation, post-translational modifications can influence protein activity; these make genes and proteins suboptimal biomolecular candidates for representing an immediate biological response. microRNAs (miRNA), which are conserved and stable against endogenous RNAses, are single-stranded, 19–22 nucleotide small RNA molecules that offer another biomolecule candidate for biomarker studies. However, studies showed that miRNA content is highly influenced by preanalytical and analytical methods such as the presence of blood cells that alter circulating miRNA content [23], which are prone to hemolysis [24], and influenced by extraction methods [25]. By contrast, metabolites (defined as small dynamic biomolecules with molecular weights <1500 Da) provide chemical phenotypes that are the net result of all biological processes including genomics, transcriptomics, miRNA and proteomics; however, they can be easily influenced by environmental variabilities (e.g., food intake, exercise, etc.). By selecting overnight fasting samples, the influence of food intake can be minimized. However, robust high-throughput analytical techniques for metabolites are still at an early development stage. Therefore, each type of biomolecule has its limitations. The choice of targeting biomolecule and corresponding omics method mainly depends on the availability of technologies and expertise.

A list of currently FDA-supported and -approved human biomarkers is displayed in Table 1. Although a majority (three out of five) of these biomarkers was analyzed by immunoassays, Murphy *et al.* used real-time quantitative reverse transcription PCR to discover Plasmodium falciparum 18S rRNA/rDNA in blood for monitoring the stage

TABLE 1

FDA-supported and -approved biomarkers for human diseases and treatment outcomesSource: FDA official site, 16 April 2019

Disease	Biomarker	Bioanalytical analysis	FDA approval/letter of support	Data analytic approach
Aspergillosis: a wide variety of fungal diseases caused by genus *Aspergillus*	Galactomannan: a cell-wall component of fungi, found in serum/broncho-alveolar lavage fluid	Immunoassay	Approved 14 Nov 2015	Univariate
Chronic kidney disease	CKD273, a peptide associated with chronic kidney disease	Urinary proteome analysis using capillary electrophoresis coupled with mass spectrometry (CE-MS)	Letter of support June 2016 [28]	Univariate
Prognostic biomarker for chronic obstructive pulmonary disease (COPD)	Fibrinogen (factor I): a circulating glycoprotein	Immunoassay	Approved 14 Sept 2016	Univariate
Prognostic biomarker with patient age and baseline glomerular filtration rate for Autosomal Dominant Polycystic Kidney Disease	Total kidney volume (TKV)	MRI and CT scan	Approved 15 Sept 2016	Univariate
Kidney tubular injury	Clusterin, cystatin-C, kidney injury molecule-1, *N*-acetyl-beta-D-glucosaminidase, neutrophil gelatinase-associated lipocalin and osteopontin	Urinary nephrotoxicity biomarker panel as assessed by immunoassays	Letter of support 25 July 2018	Multivariate
Monitoring biomarker of effective antimalarial drugs after ≥6 days of treatment initiation	*Plasmodium* 18S rRNA/rDNA	Measured in blood samples by a nucleic acid amplification test (real-time quantitative reverse transcription PCR)	Approved 12 Oct 2018	Univariate

of malaria infection [26]. It received FDA approval in 2018 as a drug-monitoring biomarker for antimalarial drugs. It is notable here that the use of an omics-derived big dataset is absent. This is in contrast to a more recent study, in which a peptide (CKD273) was discovered as a biomarker of chronic kidney disease (CKD) using high-resolution urinary proteomics (where capillary electrophoresis was coupled with mass spectrometry) from 53 of 76 non-anuric CKD patients. This study serves as an effective example of the utilization of an omics-derived big dataset in biomarker discovery [27]. Interestingly, a single peptide here classified CKD sufficiently. However, most prognostic variables are usually insufficient to classify a disease state individually, hence why multivariate data analytics are required. In mid-2016, this biomarker received an FDA letter of support for phase 3 studies after validation in >1000 participants in phase 2 [28]. In addition to proteomics, metabolomics (i.e., studying a range of metabolites) is another popular platform for biomarker discovery. In comparison with other omics technologies, metabolomics can provide the most comprehensive phenotypic picture of biological alterations under a given disease or treatment because it deals with metabolites (biochemicals) that are the ultimate determinant of phenotypes. This platform mainly falls under two broad categories: untargeted and targeted. Typically, discovery projects are best undertaken with untargeted platforms where all metabolites are profiled followed by identification steps. Although such a strategy is ideal, this often increases the complexity by 1000-fold, demanding broad skillsets and greater investments of time and money. To keep the focus on biomarker discovery, most scientists prefer a targeted metabolomics platform where only known metabolites (where molecular ion peaks are already determined) are profiled. Because the field of metabolomics is relatively new, the profiles of known metabolites in common conditions remain largely uncategorized. This provides ample opportunity for discovery. However, the main shortcoming of targeted metabolomics platforms is that they are low- to mid-level in throughput, mainly as a result of the huge diversity in the chemical reactivity and physicochemical properties of metabolites. This makes it challenging to analyze them together.

To improve accessibility in metabolomics, many government-funded institutions and private biotech companies are aiming to fulfill this demand. One such institution is the Canadian-Government-funded 'Metabolomics Innovation Centre' (https://www.metabolomicscentre.ca), a pioneer in providing broad-spectrum metabolomics services using GC–MS, LC–MS, NMR, HPLC and combined approaches, and which is extensively engaged in further innovation in the field of targeted metabolomics. Biocrates Life Sciences (https://www.biocrates.com/), a private biotech company, is another example providing a plate-based targeted metabolomics platform where up to 500 metabolites can be absolutely quantified (i.e., MxP® Quant 500 kit). Whereas the platform remains at mid-level throughput, absolute quantification is a major boon in promoting reproducibility. However, Metabolon® (hhttps://www.metabolon.com) boasts of having the most comprehensive platform thus far, allowing simultaneous quantification of >1000 metabolites from a single sample. This platform relies on an in-house reference library that was developed from >3000 standards of known metabolite structures and their LC/MS ion peaks. The metabolites are relatively quantified (i.e., semi-quantitative) here against spiked-in standards using ultra-high-performance liquid chromatography/tandem accurate mass spectrometry (UHPLC/MS/MS) methods. This platform undoubtedly offers higher throughput for targeted metabolomics; however, the

relative quantification method increases the chance of bias and irreproducibility in candidate selection in phase 1.

Biological samples for biomarker discovery

The optimal choice of biological samples for omics depends on prior knowledge of the disease and system interactions. Broadly, biological samples are either invasive or noninvasive. Invasive biological samples include blood plasma, amniotic fluid, synovial fluid, cerebral spinal fluid, bile and tissue extracts, whereas noninvasive biological samples include urine, saliva, breast milk, seminal plasma and vaginal secretion. Blood plasma and urine are the most popular biological samples in the diagnostic field owing to their suitability in collection and for reflecting the global health state. Blood plasma provides an 'instantaneous' readout of the whole-body metabolic state including catabolic and anabolic processes, whereas urine provides an 'averaged' estimation of easily-excretable polar metabolites from catabolic process [29]. Therefore, a disease related to catabolic processes (e.g., injury, infections, diabetic ketoacidosis, etc.) can be diagnosed from urine. However, urine samples perform poorly in detecting diseases related to anabolic processes (e.g., metabolic diseases, cardiovascular diseases, etc.). For organ- and tissue-specific diseases, site-specific fluids and tissue extracts are better biological samples.

Health big data preprocessing

A big dataset, generated from any omics platform, needs to go through preprocessing steps where raw data are transformed into an understandable format using appropriate data mining techniques. Raw datasets are generally incomplete and problematic, typically containing missing values, outliers and variables with noisy and noninformative data; they can also lack certain important variables entirely, among other issues. Such incompetence in the dataset is handled using different data mining tools and techniques in data preprocessing, which can be referred to as data cleaning. Essential tools and techniques for effective cleaning of datasets in unbiased biomarker discovery are described in the following section.

Stringent data output and missing value imputation

Stringent data – where the highest confidence score is selected to produce raw data – is the method of choice for data output in biomarker discovery. Although stringent raw data are the most technically robust, these data usually have a high volume of missing values. Missing values constitute one of the most common problems for big datasets and can hamper downstream analyses [30]. These missing values usually arise owing to the concentration or signal of a sample being below the detection limit (e.g., very low concentration of a specific lipid in lipidomics) or missing information (e.g., missing clinical information such as BMI, age, race, sex, etc.). There are two main options in handling missing values: (i) simple removal of missing values from the dataset; and (ii) the estimation of missing values. The best option depends on the aims of the study, and on having a proper understanding of the dataset structure and limitations of the analytical methods [31,32]. The percentage of missing values of a given variable is another crucial consideration in choosing the best option. Although there is no fixed rule for deciding when to remove a variable containing missing values, usually variables with >10–15% of their data missing are considered excludable (although this depends on how big the dataset is).

In the case of quantitative biomarker discovery, it is preferable to remove variables with missing values if this is caused by subthreshold concentration or signal. However, in the case of qualitative biomarker discovery (i.e., ranking the biomarkers), these missing values can be estimated. This strategy is also preferable in putative pathway clustering analyses. There are various approaches to accomplish this. For estimating values below the detection limit, it is standard to substitute half of the minimum positive value that was detected for the subthreshold variable. For the other cases, missing values can also be imputed in one of the following ways: variable mean, K-nearest neighbor (KNN) or regression substitution.

Data filtering: is it advisable in biomarker discovery?

Data filtering is a step to identify and remove variables with noisy and noninformative data, thereby increasing the power of analysis in detecting true differences between classes [33]. Various techniques have been developed to filter out these noisy data. There are two broad classes of data filtering: model-free techniques (e.g., mean, interquartile range, standard deviation, relative standard deviation, online multiscale filtering, etc.) and model-based techniques (e.g., Kalman filtering, unscented Kalman filtering, particle filtering, etc.) [34]. In the case of predictive biomarker discovery, it is not recommended to filter out noninformative data from a small dataset until data do not interfere with the minimum signal threshold. For large datasets where variable noise is usually high, data filtering is a worthy option. In the case of instrument-derived noise, data filtering would be necessary to remove such noise. For example, an untargeted NMR dataset (e.g., peak list, spectral binning data) usually has many close-to-baseline peaks that might be technical noise [35]. These close-to-baseline peaks can be identified using the mean or median calculation. The standard deviation or interquartile range (IQR) can identify the homeostasis variables (i.e., variables containing nearly constant value throughout the study), which should be excluded from the dataset. In the data output from the analytical devices (e.g., LC–MS, GC–MS, etc.), the relative standard deviation (RSD = SD/mean) is calculated using reference samples. Variables with high relative standard deviation percentage are considered to be noise [35].

Data normalization for better object-to-data mapping

Does every dataset need to be normalized in biomarker discovery? Before answering this question, it is important to understand the general scenario of a dataset. For example, consider a small metabolomics dataset with a few variables such as glucose (concentration range 0–15 mmol/l), ceramide (0–0.01 mmol/l) and cholesterol (0–7 mmol/l) using samples from 1000 people. These three variables have widely different concentration ranges. In multivariate analyses on unnormalized data, glucose will intrinsically influence the result owing to its larger value, but might not represent a true result (i.e., false positive). If we introduce other variables such as age (0–100 years) and income (>25 000 US$) in the same dataset, then income will become the most influential variable here. From this simple example it is clear that an unnormalized dataset is prone to biased results; thus, data need to be distributed within the same scale to eliminate such influence. The desired outcome is data that are normally (or semi-normal) distributed and with a similar scale over the whole dataset.

Data normalization is a systematic approach where rules are adopted to reduce the data redundancy and to scale all variables in

an equal spread. It increases data cohesion, which aids in modeling of different relationships for proper interpretation of data. Strategies for data normalization include applying rules, data transformation and data scaling. Sometimes it is sufficient to use one or two of these strategies if they succeed in rendering the data normal. Normalized datasets that are free of significant data redundancy and are properly scaled usually have a proper (or semi) bell-shaped distribution, referred to as a normal or Gaussian distribution (Fig. 3a). A simple data distribution analysis (or a normality test) can be applied to measure the statistical normality

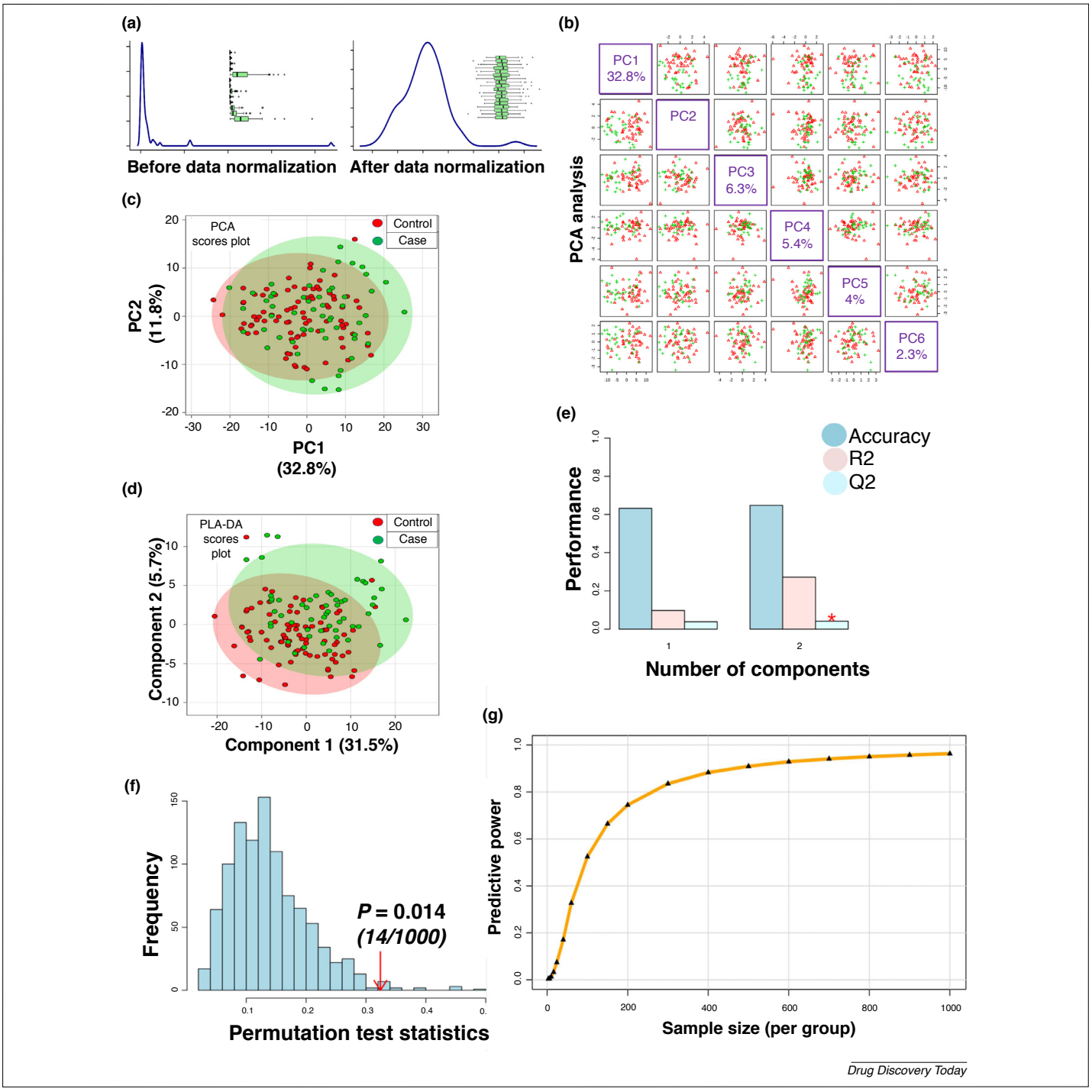

FIGURE 3

Normalization and data quality assessment. **(a)** The representative data distribution before and after normalization. The dataset was first (quantile) normalized, then followed by log2-transformation and mean-centric scaling. **(b)** The principal component analysis (PCA) on representative data. Two major principle components (PCs) contributed 44.6% eigenvalues. Others were non-major PCs. None of the non-major PCs has a high eigenvalue (i.e., ≥10%). **(c)** The PCA score plot shows overlapping tendency between PC1 and PC2. **(d)** The PLS-DA score plot shows subtle class separation. **(e)** Q^2 calculation shows that PC1 has 63% accuracy and PC2 has 64% accuracy. **(f)** Permutation analysis shows that this subtle class separation between case and control is not a random event. In 1000 permutations, it is valid for 986 events. **(g)** Power analysis showed that 275 participants per group could reach 80% predictive power (i.e., 80% chance to represent the whole population).

of the data. Normalized data (or normal data) are better for linear methods and applying more-standard statistical approaches. In cases where a normal distribution of data cannot be achieved, there are nonparametric statistical methods.

There are many rules available for data normalization. Broadly they are of two types: (i) local normalization rules; and (ii) global normalization rules. Local normalization is carried out based on reference variables. Reference variables are considered to have standard data, to which all other data are aligned. The question is: what are the criteria for selection of suitable reference variables against which normalization can be carried out? Ultimately, the criteria for reference selection are similar to those of other bioanalytical techniques. Importantly, reference variables should be stable over the course of disease states. For example, beta-actin which is a stable house-keeping gene can be used as a reference variable to normalize the transcriptional expression of other genes. By contrast, creatinine and/or total protein in renal diseases cannot be used for reference because they are affected by levels of proteinuria and would be better suited as potential biomarkers. The utilization of variables as a reference will create inherent bias within the study. When deciding between multiple variables, a simple linear regression of the variables can quantify their individual classification abilities. The variable with the weakest classification ability is considered the most stable and can be used for normalization. Instead of individual references, sometimes pooled reference samples or features can also be calculated from a group of references by using either a sum, median or mean.

In comparison to local normalization, global normalization does not rely on any specific reference. It counts all samples and creates its own reference based on certain rules. The Z-score (which calculates the number of standard deviations from the mean of a data point) is a popular global data normalization technique [36–39]. Here, data are logarithmically ($\log_2$) transformed and scaled to make a mean of zero and a standard deviation of one. Quantile normalization (Q-normalization) is another popular multisample normalization technique that considers the observed variabilities across samples as being the result of technical rather than biological reasons. Q-normalization calculates the average of each quantile across samples. Each quantile average is then used as a reference and is applied to equalize all observed distributions into one average distribution for all samples [40]. Because this Q-normalization does not consider the biological groups or conditions that can produce variabilities in the observations across samples, this strategy could come at the cost of losing some differential power of the analyses. As a solution to this shortcoming, Hicks *et al*. proposed 'smooth quantile normalization', which assumes that the statistical distribution of each sample should be the same within a given biological group or condition, but not across all samples. This alteration maintains the biological variabilities between groups or conditions [41].

Sometimes data cannot be successfully normalized by applying rules, requiring different kinds of mathematical transformation. A popular data transformation process is the logarithm ($\log_2$) transformation. The cube root transformation is another data transformation process. In a flow cytometry dataset, the hyperbolic arcsine (arcsinh) is a popular data transformation method owing to its capability in handling zero and negative values. If simple data transformation cannot establish a normal distribution, it is necessary to apply rules further or employ scaling. Scaling can be achieved using one of the following techniques: mean-centric scaling; mean-centered divided by standard deviation or standard error of each variable; and mean-centered divided by the range of each variable.

Chemometric analysis for dataset quality assessment

Before embarking upon performing any health data analyses, the experimental design needs to be optimized in terms of the influence of confounding factors, the power of the study population and the presence of batch effects. Several mathematical and statistical analyses, collectively known as chemometric analysis, are widely used to understand the data quality and make appropriate adjustments.

Confounding factors: an unavoidable problem for population studies

In population-based health data, it is impossible to consider every influencing factor in decision-making. Therefore, the chance of some confounding factors existing is unavoidable. High influence of confounding factors in a dataset causes biased results or interpretations [42]. Although this cannot be entirely eliminated, effective strategies can minimize this unwanted influence. These strategies might be better thought of as visualization methods of high-dimensional data (i.e., datasets containing many variables for many subjects). A better explanation is that these visualization methods create a set of independent variables, known as components, from collections of variables that have different dependencies and relationships. These components visualize the high-dimensional data in an understandable manner by removing redundancy. Some of them are also able to identify the presence of confounding factors with their weighted influences. There are many statistical visualization tools to identify the influence of confounding factors in a dataset. Among them, principal component analysis (PCA) is the most popular method. PCA clusters multidimensional data orthogonally in an unsupervised manner to determine their linear relationship in terms of principal components (PCs) or eigenvectors (i.e., influencing factors), with their magnitude or eigenvalues (i.e., the percentage of contributions) (Fig. 3b). The major PCs that exhibit the highest contribution percentage (or eigenvalues) in the data usually represent the known biological conditions (e.g., case and control) (Fig. 3c). The contribution percentage of each non-major PC is usually small. Any non-major PC with at least 10% contribution percentage is considered a major confounding factor, which has a substantial effect on final data interpretation and produces biased results. In the presence of such confounding factors, it is recommended to redesign the study (e.g., by adding more pair-matching criteria). In addition to characterizing the influence of confounding factors, PCA could also be used to selectively exclude the identified confounding-factor-related clusters from the dataset. This filtering operation is known as dimensionality reduction. Usually, PCs with $\leq$2 eigenvalues are removed to clean the dataset [43]. However, in biomarker discovery, this dimension reduction is not recommended because it can yield biased results. Thus, in the case of non-major PCs with small eigenvalues, it is not necessary to exclude them from the dataset via the dimension reduction process. The alternative to PCA is t-distributed stochastic neighbor

embedding (t-SNE), which is sometimes used for very complex datasets as an exploratory option to visualize nonlinear relationships that might not be observable via PCA. Because t-SNE uses a probabilistic formula [44], it is usually stochastic (random or noisy) and sometimes nonreproducible, whereas PCA is deterministic.

The class separation varies with the objective of the study

The cumulative contribution percentage (eigenvalues) of major PCs should represent the majority of the variance in the disease model dataset, whereas this might not be very prominent in the predictive model dataset. This depends on how early the disease has been targeted in the study. If the cumulative contribution percentage of major PCs is too low there is a great chance that downstream predictive analysis will be unsuccessful. Therefore, it is necessary to understand the strength of class separation between a case and control within the dataset using the Empirical Bayes method. First, a separation between two experimental classes is estimated using a supervised statistical clustering method such as PLS-DA, sPLS-DA or orthoPLS-DA. PLS-*DA* – a popular supervised statistical method – is used to estimate the strength of class separation between the experimental classes (e.g., case and control) (Fig. 3d). At a later point, this estimated class separation must be reconfirmed using the rigorous permutation protocol of empirical Bayes methods. This permutation test can find whether the performance is formed randomly. An empirical *P*-value which calculates the proportion of no class separation probability within this random parmutation tests should be at least <0.5 to ensure the class separation is not a random event. Sometimes the Q^2-value is calculated for PLS-DA to measure the certainty of the estimated class separation (Fig. 2e). However, the Q^2-value is influenced by between-class separation and within-class variability. This makes it difficult to set an acceptable cut-off for a Q^2-value from which effective classification can be identified. Therefore, a permutation analysis through empirical Bayes is recommended to prove that the observed class separation was not random (Fig. 3f). The test provides a *P*-value based on type-1 error.

Power analysis: understanding of population representation

Power analysis is a statistical method to comprehend the populational representation of a study in terms of predictive power, which is presented on a scale of 0–1 or using a percentage scale and should be performed at the study design stage. It determines the strength of a study in detecting an effect within a given degree of confidence. In other words, this analysis can predict the required sample size to achieve an effect within a certain degree of confidence (Fig. 3g). This makes sense given that the power of a dataset is mainly influenced by effect size and sample size. In general, high effect size and high sample size increase the power of study [45]. Although a power of 0.8 is considered as a benchmark for an effective study, low power might be acceptable as a pilot study if it is on the exponential phase of the curve. This phase represents a very high effect size and increasing a few-fold of sample size would be enough to attain a power of 0.8 or 80%.

Batch effects: a silent killer

The batch effect is defined as systematic, nonbiologic variation between groups of samples (or batches) owing to experimental artifacts that can arise from changes in laboratory conditions, platforms, reagent lots and personnel [46]. In most studies, batch effects are ignored. However, this phenomenon often compromises the integration and interpretation of data and could translate the results into false discovery. Therefore, it is inappropriate to combine two or more batches without batch-effect analysis. To address the presence of batch effects, the standard protocol is to run a few of the same samples from each study-group in every batch. These samples are called reference samples. The linear comparison between the different batches of these reference samples can be applied to visualize and estimate differences caused by batch effects. If a batch effect is observed between different batches it must be adjusted for accordingly. Although the aforementioned simple linear approach is very useful for understanding batch effects, it is not advisable to use this estimation for batch correction through a linear model [47].

ComBat (the combating batch effect algorithm) is the most popular method for highly effective batch-effect visualization and correction of different types of biological datasets (Fig. 4a). This algorithm uses reference samples from which it calculates the variance using the distance of a given variable in one batch from the mean of the variable across all batches. It then corrects for the variance by adjusting scale using the empirical Bayes algorithm [47]. However, batch-effect correction and/or adjustment is not always smooth in the absence of reference samples or reference variables. In such cases, the unwanted sources of variation remain unknown; surrogate variable analysis (SVA) can instead be used to estimate this unknown variation. In SVA, the surrogates of reference samples are determined from the uncommon differentially expressed variables in different batches. From a biological point of view, theoretically, the differentially expressed variables between case and control of different batches should be consistent. However, the presence of these uncommon variables is probably caused by batch effects – these are referred to as surrogate variables. Analysis of these surrogate variables can be used to estimate the batch effects that are then corrected by the ComBat algorithm.

Biomarker discovery strategies

After chemometric analyses on preprocessed data, the dataset is ready for the final operation: biomarker discovery. Strategically, there are two types of biomarker discovery in phase 1: qualitative and quantitative. The qualitative approaches employ variable ranking methods, whereas quantitative approaches select variables with concentration cut-off values. Examples of qualitative methods include the support vector machine (SVM), multiple logistic regression (MLR), PLS-DA, random forest, and so on. Among the qualitative approaches, the PLS-DA ranking method and MLR are popular. In PLS-DA, variables are ranked based on their variable importance in projection (VIP) score, which estimates the importance of each variable in the projection. VIP-score is numeric, and a VIP-score >1 is considered important; however, a VIP-score >2 is preferable for variable (biomarker) selection. In MLR, variables are ranked based on the area under the curve (AUC) of each variable. From these top-ranked variables, some variables are combined by screening out the data redundancy to increase the AUC as a signature. Although these qualitative methods do not provide an instant scale for direct decision-making on new subjects, it is the best way to narrow down biomarkers for further studies in the following phases. Logistic regression is a popular univariate quan-

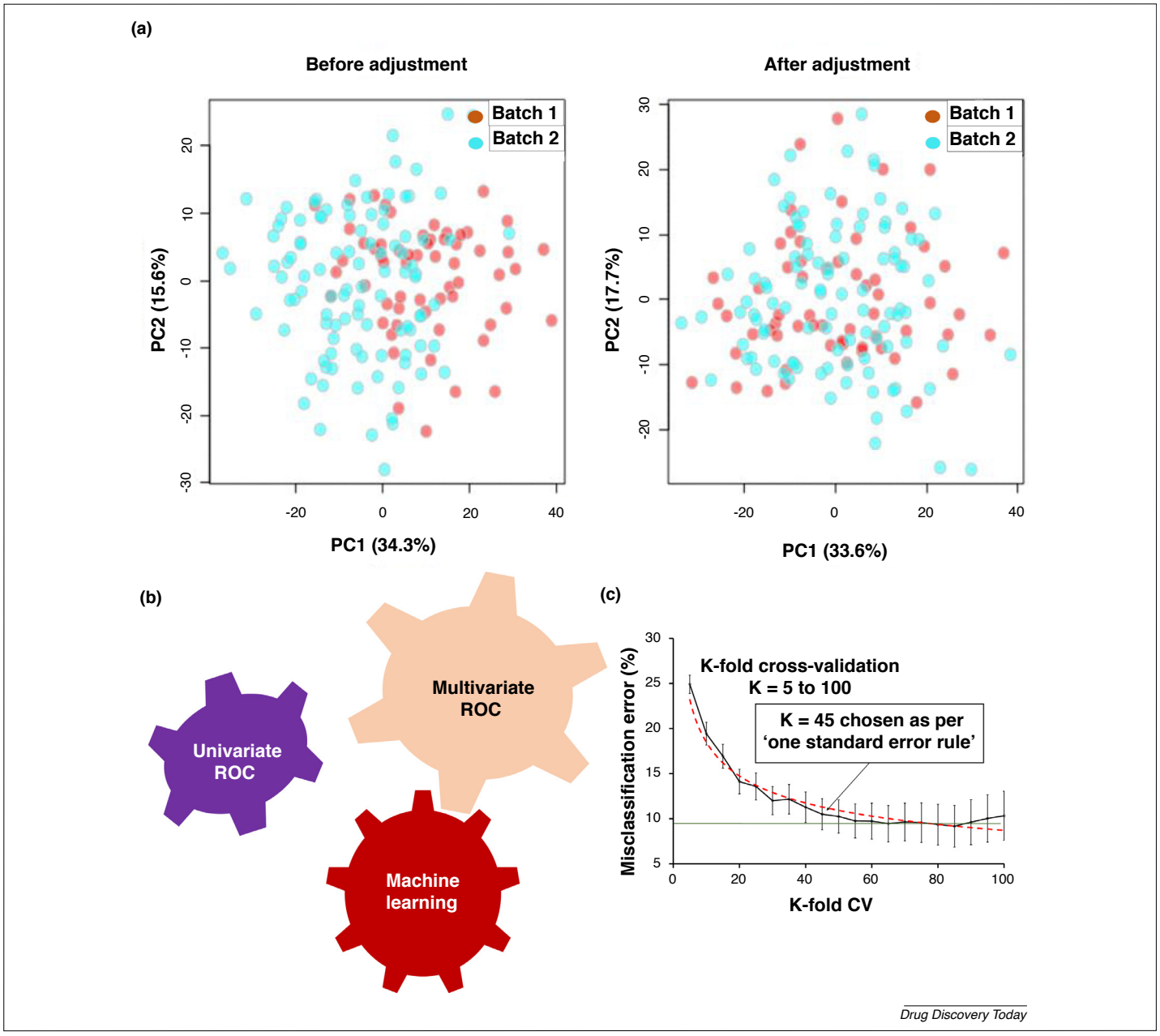

FIGURE 4

Batch effect adjustment and biomarker discovery. **(a)** The visualization of the batch effect before adjustment, and the neutralized batch effect after adjustment using ComBat. **(b)** Tools for biomarker discovery. **(c)** The application of the 'one standard rule' to optimize the K-value in K-fold cross-validation. K = 85 showed the highest accuracy owing to over-fitting. The optimized value is K = 45 owing to its presence outside of the over-fitted zone (K = 60–90) without losing significant discriminating power.

titative biomarker discovery method where a single variable is used to classify the population. In the case of multivariate quantitative biomarker discovery, a decision tree is a popular method where multiple biomarkers are clustered in a tree format with quantitative values and directions.

Data analytic tools for biomarker discovery

Statistical tools

A simple regression analysis (e.g., logistic regression, linear regression, etc.) can measure the class separation capacity of a single variable (i.e., biomolecule, environmental factor, lifestyle factor, etc.). The use of a single variable in class separation analysis is known as 'univariate ROC analysis' whereas multivariate ROC analysis uses multiple variables together (Fig. 4b). Although a single variable biomarker is most desirable owing to its suitability in clinical settings, this is largely unrealistic, especially for chronic diseases like diabetes, cardiovascular diseases, cancer, among others. Therefore, multivariate analyses are more appropriate for selecting multiple variables in generating a biomarker signature. The stepwise (forward, backward or bidirectional) multiple logistic regression (sMLR) is a classic, popular statistical tool for multivariate ROC analysis. It ranks the best predictors (biomarkers) through

logistic regression for each variable. It then combines the variables to achieve maximum predictability [48]. Sometimes there can be a strong linear dependency among two or more explanatory variables in the dataset, which raises the problem of multicollinearity (i.e., more than one variable provides the same classification for the same subjects). This can cause the selection of more variables for the same population in sMLR and lead to data redundancy. Therefore, the selected signature in sMLR should be scrutinized for data redundancy through permutation analysis. The aforementioned PLS-DA method is another statistical tool for the feature (or variable) ranking in multivariate ROC analysis. Although it is prone to data-overfitting in dimension-reduction analysis, it can deal with data redundancy in the ranking process using a proper bootstrapping method (discussed in the model optimization section) [49].

Machine learning

The artificial-intelligence-assisted machine learning approach that has become widely employed in precision medicine is another type of multivariate ROC analysis (Fig. 4b). It learns from a subset of data (training set) using statistical or various algorithm-based models and later validates its predictive model on another subset of data (testing set) using computational power. The sample size and the relationships between variables and constants are two important factors in choosing appropriate algorithms for machine learning. In general, linear algorithms are suitable for classifying a linear relationship (a straight-line relationship between a variable and a constant) using small sample sizes, whereas nonlinear algorithms are capable of classifying complex, nonlinear relationships using large sample sizes. For example, SVM is a supervised linear machine learning algorithm that can perform classification analysis on a relatively small dataset by finding a hyperplane that can differentiate between the two classes very well [50]. By contrast, a Bayesian network (BNet) (i.e., a nonlinear, probabilistic graphical model) can be used for complex diseases (e.g., cancer, diabetes, etc.) using moderate to large datasets. BNet can determine probabilities of causes depending on the probabilities of results. The directed acyclic graph with an underlying joint probability distribution in BNet can measure the uncertainty of a given situation (e.g., disease) [51]. The design of the genetic algorithm (GA) inspired by Darwin's theory of natural evolution is another example of a nonlinear algorithm that can handle small to large datasets [52]. The GA is particularly suitable when the outcome is highly unpredictable owing to complex internal processing; for example, it could be employed in the prediction of the time span needed for HIV to become drug resistant. Other popular nonlinear machine learning algorithms are decision-tree-based algorithms (e.g., J48, random forest, etc.), meta-algorithms (e.g., filtered classifier, voting, stacking, etc.), deep learning and so on. In general, nonlinear algorithms are flexible for sample size, and the incorporation of more data increases their capacity for accurate classification. However, each strategy has a certain limit after which the classification accuracy plateaus, with the exception of deep learning. This sample size range depends on the type of data. These algorithms are also nonparametric, meaning they can determine the required parameters to build a classification model by themselves.

It is important to note that many algorithms are well-suited only to specific types of data. Thorough knowledge of algorithms makes it possible to make simple tweaks to the dataset that would make it compatible with chosen algorithms. For example, the J48 algorithm of the Weka platform allows nominal classification but not numeric classification [53]. The nominal conversion of classifier labels would be sufficient to make the dataset compatible with the J48 algorithm. Exploring a few different algorithms for the same dataset is important to identify the best-suited algorithms to build the most reliable models in terms of their AUC, accuracy, sensitivity and specificity. However, these models generally need to be further optimized.

Model optimization strategies

Model optimization is an essential part of biomarker discovery. Every model can suffer from data-overfitting and bias selection. To reduce the chance of this occurring, a series of validation processes and parameter tunings are recommended. The choice of validation depends on the size of the dataset. For large datasets (i.e., >200 subjects) subjects can be divided into a training dataset (≥60% subjects) and independent testing dataset (≤40% subjects). For small datasets (<200 subjects) there is a lower chance of attaining an optimized model if subjects are used for creating a testing dataset. It is instead recommended to include all available data (or subjects) in the training set and perform a rigorous validation protocol. In the case of statistical models (e.g., logistic regression, sMLR, PLS-DA, etc.) using a small dataset, bootstrapping offers rigorous validation for biomarker and signature discovery. In bootstrapping, a random sampling with replacement technique is carried out to generate models. The final result is calculated from the average of the bootstrapping models [54,55]. In the case of a machine learning approach using a small dataset, K-fold cross-validation (K-fold CV) is the method of choice for validation. Usually, higher-fold (K) can lead to an overfit-model, whereas low values fit poorly or produce highly variable and biased models. The choice of K is optimized by using the 'one standard error rule' (Fig. 4c). As per this rule, the optimized K should meet two criteria. First, the optimized K should have a relatively low error rate with an insignificant loss of predictability from the highest possible discrimination. Second, it should not be behind the saturation point of accuracy to ensure that the model is not overfitted [56]. For a large dataset, the optimized model in the training dataset is further checked by using the independent testing dataset. The model optimization strategies have been illustrated in Fig. 5. In principle, the optimized model in the training dataset should behave similarly in the independent testing dataset within the same study cohort. If they vary significantly, this indicates that the model in the training dataset was not optimized properly.

Concluding remarks

The proper implementation of unbiased data analytics is indispensable for successful phase 1 biomarker discovery in precision medicine. In addition to reducing the chance of false biomarker discovery by eliminating biases and overfitting problems, it also provides a better understanding of disease heterogenicity and options for further stratification. Outside of analysis, increasing study population sizes in tandem with better incorporation of quantitative high-throughput technologies is also essential. The Center for Drug Evaluation and Research (CDER) division of the FDA adopted a 'biomarker qualification program' (BQP) to develop biomarkers as effective drug development tools in cooperation

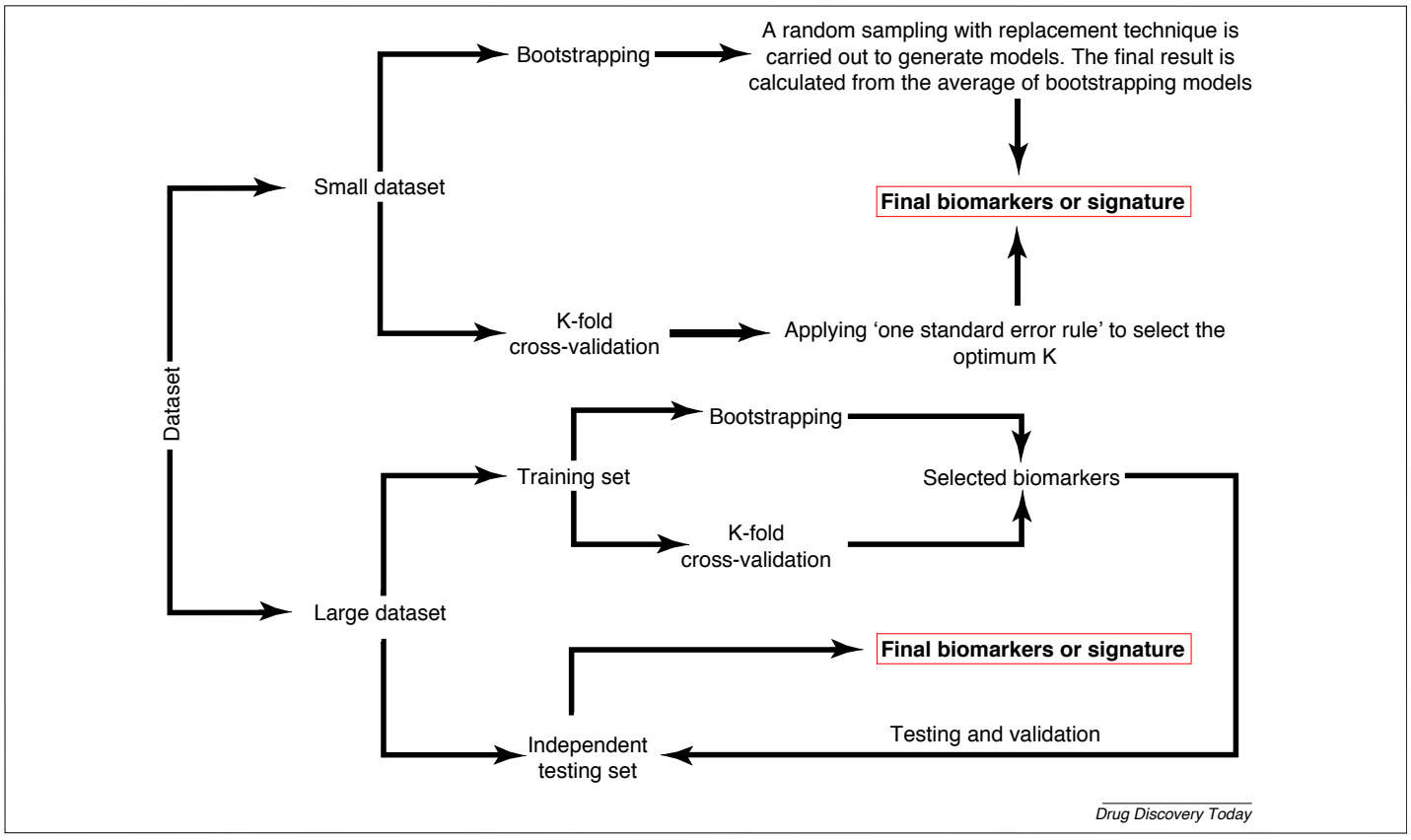

FIGURE 5

The model optimization strategies based on dataset size. In the large dataset, data are divided into two parts: 60–80% as a training set and the rest of as an independent testing set. In the testing set, biomarkers are selected by adopting proper validation protocols such as K-fold cross-validation and bootstrapping. The discovered biomarkers in the training set are further validated in an independent testing set. In a small dataset, it is not feasible to have an entirely independent dataset for testing. So, the reliability of biomarkers needs to be accounted for through either K-fold cross-validation or bootstrapping.

with external stakeholders. However, it is imprudent to conclude that a biomarker would necessarily make a good drug target. Most of the time, biomarkers are indirectly related to disease pathophysiology, and thus it is typically better to keep them separate. If a systems biology approach identifies a pathway in which biomarker molecules reside, only then can biomarkers be used as drug development tools.

Conflicts of interest

The authors declare that there are no conflicts of interest associated with this manuscript.

Acknowledgments

Saifur R. Khan is supported by Diabetes Canada postdoctoral fellowship.

References

1 Strimbu, K. and Tavel, J.A. (2010) What are biomarkers? *Curr. Opin. HIV AIDS* 5, 463–466
2 Palmer, A.M. (2014) The utility of biomarkers in CNS drug development. *Drug Discov. Today* 19, 201–203
3 Antoranz, A. *et al.* (2017) Mechanism-based biomarker discovery. *Drug Discov. Today* 22, 1209–1215
4 Martz, L. (2014) The FDA's push for better biomarkers. *SciBX* 7, 1060
5 Dumas-Mallet, E. *et al.* (2017) Low statistical power in biomedical science: a review of three human research domains. *R. Soc. Open Sci.* 4, 160254
6 Fleming, N. (2018) How artificial intelligence is changing drug discovery. *Nature* 557, S55–S57
7 Hyman, D.M. *et al.* (2015) Precision medicine at Memorial Sloan Kettering Cancer Center: clinical next-generation sequencing enabling next-generation targeted therapy trials. *Drug Discov. Today* 20, 1422–1428
8 Devi, G. and Scheltens, P. (2018) Heterogeneity of Alzheimer's disease: consequence for drug trials? *Alzheimers Res. Ther.* 10, 122
9 Collins, F.S. and Varmus, H. (2015) A new initiative on precision medicine. *N. Engl. J. Med.* 372, 793–795
10 Goossens, N. *et al.* (2015) Cancer biomarker discovery and validation. *Transl. Cancer Res.* 4, 256–269
11 Tzoulaki, I. *et al.* (2011) Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: meta-epidemiology study. *BMJ* 343, 6829
12 Ioannidis, J.P. and Panagiotou, O.A. (2011) Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. *JAMA* 305, 2200–2210
13 Rose, S. and van der Laan, M.J. (2009) Why match? Investigating matched case-control study designs with causal effect estimation. *Int. J. Biostat.* 5, 1
14 Rausch, L. *et al.* (2016) Matched-pair analysis: identification of factors with independent influence on the development of PTLD after kidney or liver transplantation. *Transplant. Res.* 5, 6
15 de Graaf, M.A. *et al.* (2011) Matching, an appealing method to avoid confounding? *Nephron. Clin. Pract.* 118, c315–c318
16 Rundle, A. *et al.* (2012) Better cancer biomarker discovery through better study design. *Eur. J. Clin. Invest.* 42, 1350–1359
17 Caruana, E.J. *et al.* (2015) Longitudinal studies. *J. Thorac. Dis.* 7, E537–E540

18 Stomrud, E. *et al.* (2015) Longitudinal cerebrospinal fluid biomarker measurements in preclinical sporadic Alzheimer's disease: a prospective 9-year study. *Alzheimers Demen. Diagn. Assess. Dis. Monit.* 1, 403–411

19 Kang, D. *et al.* (2005) Design issues in cross-sectional biomarkers studies: urinary biomarkers of PAH exposure and oxidative stress. *Mutat. Res. Fund. Mol. Mech. Mutagen.* 592, 138–146

20 Tworoger, S.S. and Hankinson, S.E. (2006) Collection, processing, and storage of biological samples in epidemiologic studies: sex hormones, carotenoids, inflammatory markers, and proteomics as examples. *Cancer Epidemiol. Biomark. Prev.* 15, 1578–1581

21 Dakappagari, N. *et al.* (2017) Recommendations for clinical biomarker specimen preservation and stability assessments. *Bioanalysis* 9, 643–653

22 Khan, S.R. *et al.* (2014) Current status and future prospects of toxicogenomics in drug discovery. *Drug Discov. Today* 19, 562–578

23 Pritchard, C.C. *et al.* (2012) Blood cell origin of circulating microRNAs: a cautionary note for cancer biomarker studies. *Cancer Prev. Res.* 5, 492–497

24 Tiberio, P. *et al.* (2015) Challenges in using circulating miRNAs as cancer biomarkers. *BioMed Res. Int.* 2015, 731479

25 McDonald, J.S. *et al.* (2011) Analysis of circulating microRNA: preanalytical and analytical challenges. *Clin. Chem.* 57, 833–840

26 Murphy, S.C. *et al.* (2012) Real-time quantitative reverse transcription PCR for monitoring of blood-stage *Plasmodium falciparum* infections in malaria human challenge trials. *Am. J. Trop. Med. Hyg.* 86, 383–394

27 Argilés, À. *et al.* (2013) CKD273, a new proteomics classifier assessing CKD and its prognosis. *PLoS One* 8, e62837

28 Pontillo, C. and Mischak, H. (2017) Urinary peptide-based classifier CKD273: towards clinical application in chronic kidney disease. *Clin. Kidney J.* 10, 192–201

29 Álvarez-Sánchez, B. *et al.* (2010) Metabolomics analysis I. Selection of biological samples and practical aspects preceding sample preparation. *TrAC Trends Anal. Chem.* 29, 111–119

30 Graham, J.W. (2009) Missing data analysis: making it work in the real world. *Annu. Rev. Psychol.* 60, 549–576

31 Baraldi, A.N. and Enders, C.K. (2010) An introduction to modern missing data analyses. *J. School Psychol.* 48, 5–37

32 Khan, S.R. *et al.* (2019) The discovery of novel predictive biomarkers and early-stage pathophysiology for the transition from gestational diabetes to type 2 diabetes. *Diabetologia* 62, 687–703

33 Hackstadt, A.J. and Hess, A.M. (2009) Filtering for increased power for microarray data analysis. *BMC Bioinf.* 10, 11

34 Nounou, M.N. *et al.* (2013) Model-based and model-free filtering of genomic data. *Netw. Model. Anal. Health Inf. Bioinf.* 2, 109–121

35 Xia, J. *et al.* (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res.* 37, W652–W660

36 Khan, S.R. *et al.* (2016) Cytoprotective effect of isoniazid against H2O2 derived injury in HL-60 cells. *Chem. Biol. Interact.* 244, 37–48

37 Khan, S.R. *et al.* (2015) Proteomic profile of aminoglutethimide-induced apoptosis in HL-60 cells: role of myeloperoxidase and arylamine free radicals. *Chem. Biol. Interact.* 239, 129–138

38 Khan, S.R. *et al.* (2016) Global protein expression dataset acquired during isoniazid-induced cytoprotection against H_2O_2 challenge in HL-60 cells. *Data Brief* 6, 823–828

39 Babu, D. *et al.* (2019) Isoniazid induces a monocytic-like phenotype in HL-60 cells. *Arch. Biochem. Biophys.* 664, 15–23

40 Hicks, S.C. and Irizarry, R.A. (2015) quantro: a data-driven approach to guide the choice of an appropriate normalization method. *Genome Biol.* 16, 117. http://dx.doi.org/10.1186/s13059-015-0679-0

41 Hicks, S.C. *et al.* (2018) Smooth quantile normalization. *Biostatistics* 19, 185–198

42 Skelly, A.C. *et al.* (2012) Assessing bias: the importance of considering confounding. *Evid. Based Spine Care J.* 3, 9–12

43 Razquin, C. *et al.* (2018) Plasma lipidomic profiling and risk of type 2 diabetes in the PREDIMED Trial. *Diabetes Care* 41, 2617–2624

44 van der Maaten, L. and Hinton, G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605

45 van Iterson, M. *et al.* (2009) Relative power and sample size analysis on gene expression profiling data. *BMC Genomics* 10, 439

46 Reese, S.E. *et al.* (2013) A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. *Bioinformatics* 29, 2877–2883

47 Price, E.M. and Robinson, W.P. (2018) Adjusting for batch effects in DNA methylation microarray data, a lesson learned. *Front. Genet.* 9, 83

48 Bursac, Z. *et al.* (2008) Purposeful selection of variables in logistic regression. *Source Code Biol. Med.* 3, 17

49 Lee, L.C. *et al.* (2018) Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. *Analyst* 143, 3526–3539

50 Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Mach. Learn.* 20, 273–297

51 Sesen, M.B. *et al.* (2013) Bayesian networks for clinical decision support in lung cancer care. *PLoS One* 8, e82349

52 McCall, J. (2005) Genetic algorithms for modelling and optimisation. *J. Comput. Appl. Math.* 184, 205–222

53 Galli, M. *et al.* (2016) Machine learning approaches in MALDI-MSI: clinical applications. *Expert Rev. Proteomics* 13, 685–696

54 Xia, J. and Wishart, D.S. (2016) Using MetaboAnalyst 3.0 for comprehensive metabolomics data analysis. *Curr. Protoc. Bioinf.* 55 14.10.11–14.10.91

55 Kaufmann, E. and Wittmann, W.W. (2016) The success of linear bootstrapping models: decision domain-, expertise-, and criterion-specific meta-analysis. *PLoS One* 11 e0157914

56 Krstajic, D. *et al.* (2014) Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminf.* 6, 10