



REVIEW

Epidemiology/Genetics

Transforming Big Data into AI-ready data for nutrition and obesity research

Diana M. Thomas¹  | Rob Knight² | Jack A. Gilbert³ | Marilyn C. Cornelis⁴ | Marie G. Gantz⁵ | Kate Burdekin⁵ | Kevin Cummiskey¹ | Susan C. J. Sumner⁶ | Wimal Pathmasiri⁶ | Edward Sazonov⁷ | Kelley Pettee Gabriel⁸ | Erin E. Dooley⁸ | Mark A. Green⁹  | Andrew Pfluger¹⁰ | Samantha Kleinberg¹¹

¹Department of Mathematical Sciences, United States Military Academy, West Point, New York, USA

²Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, California, USA

³Department of Pediatrics and Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, USA

⁴Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA

⁵Biostatistics and Epidemiology Division, Research Triangle Institute International, Research Triangle Park, North Carolina, USA

⁶Department of Nutrition, Nutrition Research Institute, University of North Carolina Chapel Hill, Kannapolis, North Carolina, USA

⁷Electrical and Computer Engineering Department, The University of Alabama, Tuscaloosa, Alabama, USA

⁸Department of Epidemiology, The University of Alabama at Birmingham, Birmingham, Alabama, USA

⁹Department of Geography & Planning, University of Liverpool, Liverpool, UK

¹⁰Department of Geography and Environmental Engineering, United States Military Academy, West Point, New York, USA

¹¹Computer Science Department, Stevens Institute of Technology, Hoboken, New Jersey, USA

Correspondence

Diana M. Thomas

Email: diana.thomas@westpoint.edu

Funding information

U.S. Department of Defense Environmental Security Technology Certification Program, Grant/Award Number: EW22-7278; the National Institutes of Health (NIH) Interagency Agreement, Grant/Award Number: AOD22022001; National Institutes of Health, Grant/Award Numbers: U54TR004279, U24HD107676, U24CA268228, U24DK131617-01, UG1HD107688, U24CA268153

Abstract

Objective: Big Data are increasingly used in obesity and nutrition research to gain new insights and derive personalized guidance; however, this data in raw form are often not usable. Substantial preprocessing, which requires machine learning (ML), human judgment, and specialized software, is required to transform Big Data into artificial intelligence (AI)- and ML-ready data. These preprocessing steps are the most complex part of the entire modeling pipeline. Understanding the complexity of these steps by the end user is critical for reducing misunderstanding, faulty interpretation, and erroneous downstream conclusions.

Methods: We reviewed three popular obesity/nutrition Big Data sources: microbiome, metabolomics, and accelerometry. The preprocessing pipelines, specialized software, challenges, and how decisions impact final AI- and ML-ready products were detailed.

Results: Opportunities for advances to improve quality control, speed of preprocessing, and intelligent end user consumption were presented.

Conclusions: Big Data have the exciting potential for identifying new modifiable factors that impact obesity research. However, to ensure accurate interpretation of conclusions arising from Big Data, the choices involved in preparing AI- and ML-ready data need to be transparent to investigators and clinicians relying on the conclusions.

INTRODUCTION

The words “Big Data” in obesity and nutrition research are usually associated with innovation in artificial intelligence (AI), machine learning (ML), and precision nutrition [1–4]. Active nutrition and obesity research programs are integrating data collected by traditional methods (e.g., self-reported surveys) with new, rich data such as those extracted from body-worn sensors (measuring movement, sleep, nutrition, blood glucose, eating), the microbiome, and the metabolome. These new data sources enhance the potential to identify biomarkers and physiology of mechanisms underpinning disease and predictive features to explain interpersonal variation in response to diet [1–4]. However, major hurdles still exist in transforming and integrating these complex raw data to facilitate effective analysis and interpretation to gain new insights into nutrition and obesity. Unlike lifestyle survey instruments and traditional nutrition and obesity data (e.g., body composition), these new data sources require substantial preprocessing to identify useful associations with the behavioral and physiological features. Therefore, data from sensors and multiomic methodologies (metagenomics, metatranscriptomics, metaproteomics, and metabolomics) require significantly more processing represented by a combination of automated methods and human judgment to unlock their valuable insights.

In order to clearly outline the steps required to transform raw Big Data into AI-ready data, we begin with definitions of Big Data, AI/ML, and AI-ready data.

What are Big Data?

The words “Big Data” were used for the first time in 1984 by sociologist Charles Tilly to describe large-sized datasets [5]. Today, the term “Big Data” reflects more than just the size of the dataset but the magnitude of the four “v’s” [6]. The first “v” is the volume of data and represents the large size of datasets. For example, data generated from the microbiome, metabolomics, and wearable devices represent large volumes of data that can pose challenges in being able to be manipulated and analyzed. The second “v” is the velocity or speed of how quickly the data are generated. An example of high-velocity data generated in real time that are important to obesity and nutrition are accelerometry data, which may be read by sensors several hundred times per second [7]. The third “v” is the variety of the data. Big data can be structured, semi-structured, or unstructured. Structured data have a clear format of type, length, and volume. An example of structured data are the movement data generated from accelerometers stored as meters per second squared. Semi-structured data are data in which the format is partially clear, such as electronic health record data. Some electronic health records store data such as problem lists in structured form, whereas others use text-based records in which data may follow a template, leading to semi-structured data. Finally, examples of unstructured data in obesity and nutrition are

Study Importance

What is already known?

- Obesity and nutrition studies have been increasingly including Big Data to gain new insights and develop new and often personalized guidance. Big Data are often used in concert with artificial intelligence (AI) and machine learning (ML) models. Making Big Data AI-ready requires a series of labor-intensive steps and decisions requiring human judgment. These steps are rarely addressed in reports of the studies in the literature, particularly in the context of nutrition.

What does this review add?

- We lay out the steps to preprocess three types of Big Data (microbiome, metabolomics, and accelerometry) into AI-ready data, emphasizing the points at which human judgment plays a role. These data types are increasingly used to derive new insights and conclusions in obesity and nutrition research. Areas for improvement and streamlining the preprocessing steps are identified.

How might these results change the direction of research or the focus of clinical practice?

- AI- and ML-developed algorithms relying upon Big Data are increasingly being used to guide nutrition recommendations. Understanding the steps and gaps required to transform Big Data into AI- and ML-ready data is critical, especially when human judgment is required, before applying models to inform patient care.

video food intake and video physical activity-monitoring data [8–11] or public comments to the Dietary Guidelines [12]. Aside from having a data type such as text or video, these data do not have any further organization nor do they have partially structured formats. The final “v” is veracity of Big Data. The veracity of Big Data can be defined as the underlying accuracy, or lack thereof, of the data in consideration, specifically impacting the ability to derive actionable insights and value out of it. Examples of poor veracity in the data can come from hacking or falsifying data or too many missing values [13, 14]. Additionally, the data may require new, or combinations of, analytic methods to extract findings that are useful. Finally, more data may not always lead to new insights [15]. Owing to these additional considerations, we suggest including a fifth “v,” value, to the itemized description because “Big Data” should extend or complement insights that are generated from smaller traditional datasets [13].

What are AI and ML?

Since the first seminal publications on AI [16, 17] and ML [18], there have been numerous definitions of AI and ML in which ML is sometimes considered a subset of AI, and the terms are frequently interchanged [19]. The term “AI” is often referred to in applications that have the capability of creating human-like cognition, usually from the application of multiple layers of ML models. For example, a user can have conversations with the generative AI platform ChatGPT, which relies on large language models that are based in deep-learning models (or sets of neural networks) called transformers [20]. Generative AI differs from ML because ML usually refers to a single ML model. For consistency in our discussion, we define ML as computer algorithms that improve automatically through experience [19, 21] and AI as an algorithm that can learn insights, adapt through feedback, be dynamic, respond to its environment, and problem-solve independently, with minimal human supervision [19, 22].

What are AI-ready data?

Big Data are typically not interpretable in raw form, and there are steps required that often rely on methods such as ML and AI combined with human judgment to make data AI-ready.

The definition of AI-ready data is evolving depending on the type of organization, field of study, and AI use-case involved [23]. Because of the wide range of AI applications, use-cases, and needs by level of user (e.g., individual, field, organization), a common definition of AI-ready has remained elusive [24]. The National Oceanic and Atmospheric Association (NOAA) has identified four attributes of AI-ready data: AI-ready data has been prepared (e.g., labeled, annotated), checked for quality, documented, and has defined criteria for access [25]. These attributes overlap with the principles of the Findable, Accessible, Interoperable and Reusable (FAIR) data principles [25]; however, some attributes such as treatment of missing data and outliers are not covered under FAIR. In order to cover the NOAA AI-ready data attributes in a broad, encompassing manner, we will apply the UK National AI Strategy definition of AI-ready data (referred to as data foundations) as the “characteristics of data that contribute to its overall condition, whether it is fit for purpose, recorded in standardized formats on modern, future-proof systems and held in a condition that means it is FAIR” [26].

The raw form of most sources of Big Data in nutrition and obesity research is not AI-ready. For example, raw sequence data generated from the microbiome are not AI-ready. This is because the raw sequence data do not correspond to the entities that need to be analyzed and related to phenotype information. The raw data need to be heavily preprocessed with iterative internal human judgment checks and specialized software to transform the sequenced data into an AI-ready dataset (e.g., operational genomic unit [OGU] tables, operational taxonomic unit [OTU] tables, amplicon sequence variants) [27–29].

Why should nutrition and obesity researchers care about Big Data processing?

The steps that transform Big Data into AI-ready datasets are the most complex of the entire modeling process and require the highest amount of labor and reflection. When consumers of the AI/ML models do not review the decisions and processes involved for transforming raw data into AI-ready data, misunderstanding and faulty interpretation of the analysis and conclusions can result downstream. Finally, failure to account for the true cost of the labor and focused attention required for quality Big Data preprocessing can result in substandard AI/ML models that are open to error [30].

Here, we present the preprocessing steps that make Big Data AI-ready using three specific examples of Big Data: the microbiome, metabolomics, and accelerometry. These three Big Data sources were selected because of their increased use within precision nutrition studies to discover previously unknown factors that predict response to diets [1, 2, 31, 32]. Therefore, understanding the process to transform these data sources into AI-ready datasets within the field of obesity and nutrition is particularly important. We note that the choice of these three data sources does not include issues such as hallucinations observed in generative AI products [33] or concerns regarding ethical issues in AI.

The purpose of this study comprised the following: 1) to highlight the current state of the field and opportunities for improved methods (e.g., apps/software that generate AI-ready data from raw data); 2) to present the importance of preprocessing Big Data in the AI/ML modeling pipeline; and 3) to familiarize the larger obesity and nutrition research community about the true labor, time, and attention involved for quality preprocessing of Big Data. Our article is designed to walk readers through the complex journey of transforming raw Big Data into AI-ready data.

BIG DATA EXAMPLES IN OBESITY AND NUTRITION RESEARCH

A summary of the Big Data types described in the next few sections (microbiome, metabolomics, and accelerometry), their unique challenges, advances required to help the end user, and challenges common to all appear in Table 1.

Microbiome data in obesity and nutrition research

Connections among the composition of gut microbiota, obesity, and obesity-related comorbidities have been long observed in both animal and human studies [36]. Although historical microbiome studies involving diet and obesity have relied on associations or interventions that did not have sufficient objective measurements of food intake, a recent in-residence intervention study demonstrated that diet-microbiome interactions do indeed impact energy balance [3]. In addition, data from the microbiome have been successfully used in several

TABLE 1 A summary of the Big Data types described (microbiome, metabolomics, and accelerometry), their unique challenges, advances required to help the end user, and challenges common to all.

Big Data type	Unique challenges transforming raw data into AI-ready data	Advances to help end user	Common challenges
Microbiome	Variation and errors in DNA sequences Noise and bias in sequence results Incomplete microbiome reference databases Lack of preprocessing approaches that produce voluminous, high-quality data needed to train models	Continued sequencing technology improvements Reduced sequencing costs Continued reference database improvement Advances in preprocessing approaches to develop high-quality, correctly labeled data able to train AI/ML models Software: QIIME2 [28] https://qiime2.org/	Developing common reference databases to annotate preprocessed data Developing preprocessing tools that are accessible to mainstream users
Metabolome	Peak picking Background subtraction Instrument drift and batch correction Peak identification and annotation Identification of dark matter (unknown but clearly visible signals) Quantitation Identifying biological relevance	Advanced data processing software Common standardized reference materials Creation of a national repository of non-commercial available standards Public spectral databases Public repositories of annotated data Software: ADAP [34] https://adap-big.github.io/	
Accelerometry	Classifying activity types (e.g., bicycling, yoga) or transitions from one activity type to the next (e.g., sit to stand) in real-life settings Predicting non-steady state energy expenditure predictions Reconciling time scales Comparability among devices and device placement Opacity of proprietary software State of the art ML-based preprocessing models are not accessible to mainstream users	Broad ground truth data with really good labels Ground truth for non-steady state energy expenditure Development of more user-friendly tools for model delivery Software: GGIR [35] https://cran.r-project.org/web/packages/GGIR/vignettes/GGIR.html	

Abbreviations: ADAP, Automated Data Analysis Pipeline; AI, artificial intelligence; ML, machine learning.

precision nutrition studies to identify connections between the composition of gut microbiota and postprandial response to foods [1, 2]. The exponential rise in gut microbiome nutrition and obesity studies speaks to a growing enthusiasm to include microbiome data within the field.

What do raw microbiome data look like, and why are they not AI-ready?

Raw data for microbiome analyses typically come in the form of DNA sequences [37]. DNA sequences may be generated by targeted approaches such as amplicon sequencing [38], in which the sequence between two conserved regions of DNA that serve as primers for polymerase chain reaction is amplified, or by untargeted approaches such as shotgun metagenomics [39], in which all of the DNA is fragmented and sequenced. Sequences can be either “short-read” (typically 100–300 base pairs long) such as those from Illumina platforms or “long-read” such as those from PacBio or Oxford Nanopore platforms (thousands of bases; note that this is still much shorter than a

typical bacterial genome, which is ~5 million bases) [40, 41]. The DNA sequences can be thought of as strings on a four-character alphabet, with additional characters such as “N” used to indicate positions that cannot be assigned to a single base [42, 43]. Typically, each position is also assigned a quality score, which takes the form of a small integer and is stored alongside the sequence data in a format called FASTQ [44, 45]. Instrument output ranges from a few million to a few billion sequences that can be either single-end (i.e., read from only one end of the DNA fragment) or paired-end (i.e., read from both ends of the fragment). Paired-ends can be either overlapping or non-overlapping [46].

What are the steps to transform raw microbiome data into AI-ready data?

DNA sequences obtained from a single sequencing run are unordered. Typically, barcoding approaches are used to introduce synthetic DNA sequences into the start of each DNA molecule before sequencing so that DNA from different biological specimens can be mixed together

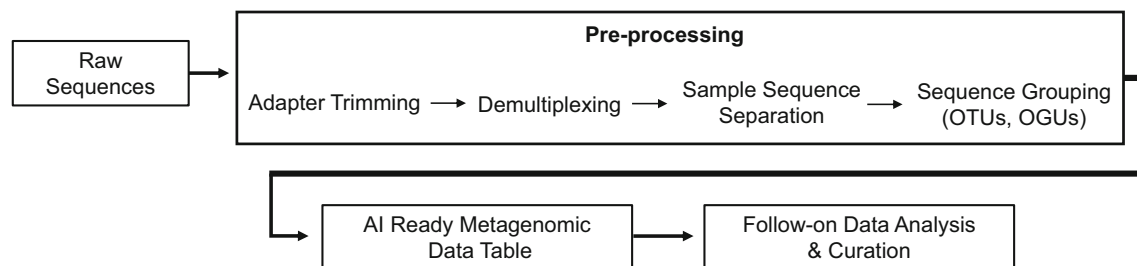


FIGURE 1 Flow diagram of the preprocessing steps to make microbiome data artificial intelligence (AI)-ready. OGU, operational genomic unit; OTU, operational taxonomic unit.

on the same sequencing run (because each sequencing run is expensive and produces more sequences than are required to interpret a single sample) [47]. Although the number of sequences obtained from each sample is related to the DNA concentration in the input, this relationship is not exact, and there is often considerable variation in the yield of the number of sequences in each sample (sometimes two orders of magnitude). This problem is not under the control of the researcher providing the samples or the laboratory doing the sequencing. Each sequencing run can produce slightly different results (both noise and bias); therefore, it is undesirable to combine information across sequencing runs, although it is unavoidable on large projects [48].

Because most study designs are multiplexed, the first step is to trim the strings (“adapters”), which are introduced to allow the sequencing process to proceed, from the recovered DNA. The barcodes also need to be identified and removed in a process called demultiplexing, and all the sequences associated with each sample need to be separated [49, 50].

The raw DNA sequences contain errors, with the error rate depending on platform. Furthermore, for shotgun metagenomics, the start and end points are random; therefore, it is not useful to count identical sequences directly [51, 52]. Instead, the sequences that correspond to the “same” biological entity for a given level of analysis (strain, species, genus, etc.) need to be identified. This can be done for amplicon data by grouping into OTUs or for shotgun metagenomic data by grouping into OTUs or into OGUs, a related concept that relies on grouping to a known reference database [27]. In general, methods for grouping these sequences can be reference-based, i.e., aligning to a known but incomplete set of reference sequences, or *de novo*, i.e., determining the groupings entirely from the data themselves. All of these procedures are error-prone, especially when the databases are highly incomplete, but are nonetheless useful for community analysis.

The end product of “preprocessing” is an AI-ready table in which each cell corresponds to the count of a particular entity (e.g., species or gene function) in a particular biological specimen [27, 53]. Follow-on steps include host DNA filtering (for metagenomic samples because targeted amplicon approaches typically do not pick up any host DNA); adaptor trimming; barcode identification and demultiplexing; trimming (if the sequences are heterogeneous in length or if a sequencing run needs to be combined with sequences from

a different run with different length characteristics); read stitching (if paired-end reads need to be merged); removal of low-quality sequences (by quality scores endogenous to the data or by comparison to a reference database of known good sequences); and taxonomy assignment [48, 51]. Function assignment is completed after taxonomy assignment (e.g., by using a tool such as PICRUSt to impute function from phylogeny) or by direct matching to a database of reference sequences of known function [54]. A description of the pipeline to transform microbiome data into AI-ready data appears in Figure 1.

Metabolomics data in obesity and nutrition research

Metabolomics involves the detection of the low molecular weight compounds in cells, tissues, and biological fluids and the investigation of how levels of metabolites differ among diverse populations, states of health and wellness, behaviors, dietary intake, physical activity, and other exposure or lifestyle factors [55–58]. Metabolomics has been increasingly used to understand how changes in diet may lead to changes in metabolism [55] and how dietary interventions may improve health [59]. Metabolomics studies are typically referred to as targeted or untargeted [60, 61]. Historically, targeted methods have measured known or identified metabolites in studies designed to test hypotheses. With increasingly large and comprehensive metabolite panels, targeted studies may also be hypothesis-generating. Untargeted methods enable the measurement of tens of thousands of signals that are identified or annotated using physical standards or Big Data analytics. Untargeted metabolomics is ideally suited for hypothesis-generating.

How are metabolomics data acquired?

Nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) are the two most common approaches used to acquire metabolomics data [62]. No one metabolomics method can capture the entire metabolome, and, when possible, multiple platforms are used to increase the coverage of detected analytes.

NMR involves the placement of a biological sample (e.g., urine, serum, saliva, organ tissue) or its extract in a magnetic field, excitation of the nuclei within the sample via radio-frequency pulses, and

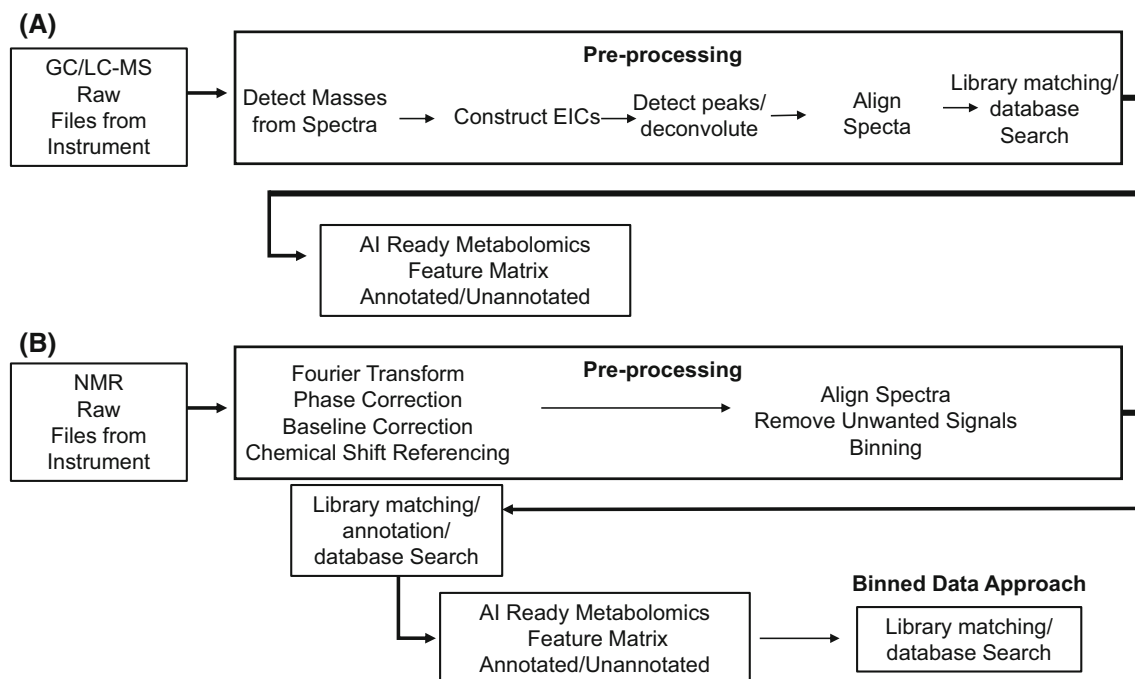


FIGURE 2 (A) Pipeline flowchart depicting the transformation from raw mass spectrometry (MS) metabolomics data to data that are artificial intelligence (AI)-ready (extracted ion chromatogram [EIC]) [111]. Pipeline modified from figure presented in [34]. (B) Pipeline flowchart depicting the transformation from nuclear magnetic resonance (NMR) spectroscopy metabolomics data to data that are AI-ready. Data matrix can be prepared using binned data or annotated/library matched data. For binned data approach, library matching/database search is performed on select important NMR bins. GC, gas chromatography; LC, liquid chromatography; MS, mass spectrometry.

detection of the resultant electromagnetic signal at a frequency that is related to the chemical structure surrounding the nuclei as it returns to its equilibrium state (referred to as a free-induction decay). Following Fourier transformation of the free-induction decay signal, the NMR spectrum is a plot of the intensity of the signal at a specific chemical shift (or resonant frequency of the nuclei) relative to a reference standard [63].

Instrument-level software tools facilitate automated NMR data acquisition, as well as data preprocessing via Fourier transformation, phasing, baseline correction, data normalization, and binning [64]. Artifacts in the preprocessed data may be a result of instrument-specific (i.e., incorrect phases of the spectra) or sample-specific (i.e., different dilutions) effects and must be addressed prior to data analysis. The region of water suppression, i.e., signals from solvents, or known drugs or medications may be removed before data analysis.

The structure and quantity of the metabolite can be determined using chemical shift data, intensity (or area under the peak), number of atoms under the peak, and information on the interaction among nuclei (such as spin-spin coupling patterns). NMR laboratories use commercial software (e.g., Chenomx Inc., Edmonton, Alberta, Canada) or source commercial standards of candidate metabolites to create in-house libraries that contain relevant structural information that can be used to match to signals in the NMR spectra of biological samples [64]. Spiking biospecimens with known concentrations of a standard are also used to facilitate identification and quantitation of metabolites. In

addition, there are several freely available academic programs that have been developed to semi-automate compound identification or quantification using NMR spectral datasets (e.g., BATMAN, Baysil, AQuA, ASICS, ASICS 2.0) [65].

NMR metabolomics studies have typically reported the identification of 50 to 200 metabolites in biospecimens such as urine, serum, plasma, saliva, and stool. Although NMR metabolomics has less resolution and sensitivity than MS methods, it is nondestructive to the sample, is highly reproducible within and among laboratories (particularly advantageous for analyzing samples generated from larger epidemiological studies, e.g., longitudinal studies), detects molecules that are difficult to ionize by MS, and can be used to determine the structure of the molecule.

Targeted or untargeted MS approaches are typically coupled with a separation technique such as gas chromatography (GC) or liquid chromatography (LC). Typically, a sample is injected onto a column for chromatographic separation and the elutant that enters the MS is bombarded with electrons to produce ion fragments. These ions are separated by their mass-to-charge ratio, resulting in a spectrum of intensities of each detected ion signal and its mass-to-charge ratio. Instrument-level software can be used for data preprocessing (e.g., peak picking, data alignment, data filtering; Figure 2A,B). The data are presented as the chromatographic retention time, the mass-to-charge ratio, and the fragmentation or isotopic pattern. For targeted methods, the peaks detected in a biological specimen are compared with those of an authentic standard for each analyte, and

quantification is preformed using concentration response curves for each analyte.

High-resolution MS platforms detect 10,000 to 40,000 signals in biospecimens such as urine, serum, plasma, saliva, or stool. Identification of these signals requires the development of libraries using authentic standards run under identical conditions to the biological specimen. Such in-house physical standards libraries are limited by the number of commercial standards available, the expense associated with the purchase of standards, the data acquisition, and the construction of the library. Metabolite identifications can be approached through first the annotation of a signal and then the purchase or synthesis of standards for confirmation. Signals can be annotated by matching to public databases that contain hundreds of thousands of experimental or empirical data for individual compounds or by comparison to metabolomics spectra stored in public repositories (e.g., <https://www.metabolomicsworkbench.org/>) for which an annotation, or identification, has been reported. Quantitation in untargeted metabolomics is challenging due to the large number of peaks detected, many of which are identified by matching to in-house physical standards libraries using one concentration or limited to annotation using public databases.

What are the steps to transform raw metabolomics data into AI-ready data?

A product of the metabolomics pipeline is an (annotated) feature matrix. After platform-specific processing (Figure 2A,B), additional preanalysis processing is required: summarization of technical replicates (if applicable), filtering of metabolites (i.e., excess missingness), imputation of missing values, transformation of data, and normalization and batch correction [66]. A missing value can be due to the metabolite not being present or present but below the detection threshold or for technical reasons. A missing value can also be due to true missingness, e.g., related to inborn error in metabolism, genetic polymorphisms, or differences in exposures. Thus, the cause for missingness, although often unknown, may impact bias introduced by different imputation methods [66].

Challenges for transforming metabolomics data into AI-ready data: scaling

Statistical analysis approaches used in metabolomics studies include univariate and multivariate methods, hierarchical cluster analysis, or logistic regression modeling. Multivariate models align with AI and are applied to the full matrix of variables, i.e., seeking patterns or relationships. The performance of statistical models and interpretation of the results depend significantly on the data preprocessing, scaling, and normalization methods used [64]. For example, a consequence of spectra reduction could be that strong changes of low-concentrated metabolites could be masked by slight changes of highly concentrated metabolites if these metabolites have adjacent or overlapping peaks.

Data scaling aims to standardize the range of data values and is often performed prior to projection methods such as principal component analysis. Scaling renders the influence of each data point on the multivariate model equally important. Any dominance of highly concentrated metabolites on projection methods is thereby diminished. However, scaled spectra are highly distorted and do not reflect the concentration–intensity relationship [64].

Challenges for transforming metabolomics data into AI-ready data: batch effects

Metabolomics investigations, particularly large-scale studies requiring data to be acquired over years, are faced with ensuring that the inherent variability in sample preparation, data acquisition (including drift in instrument response), and data processing is effectively mitigated to ensure rigor and reproducibility (see preprocessing workflow diagram in Figure 2A,B). Metabolomics data are acquired for biospecimens that are randomized into batches. The sample preparation phase includes spiking internal standards into each individual biospecimen to use as a reference of instrument response for that biospecimen. Quality control (QC) study pools (QCSPs) are created for each batch of samples through combining an aliquot of each individual study sample. The QCSP, a reference material (such as a National Institute of Standards and Technology [NIST] standard), and blanks (water) are prepared identical to the individual study samples. Data are acquired for the individual study samples with the QCSPs, the reference material, and the blanks interspersed (e.g., QC samples run after every 10 biological samples). The QC samples are used to subtract background, filter data to remove inherent variability within a batch, and to filter and align data across batches. In addition, data for QC standards are acquired prior to and following each batch to assess instrument drift and performance.

Challenges for transforming metabolomics data into AI-ready data: standardized reference materials

Challenges in the field of metabolomics, and particularly to providing Big Data for large-scale epidemiological investigations, include ensuring that data are captured using the range of QC standards and reference materials that are needed for harmonization within and across laboratories and across analytical technologies. Because laboratories use different compounds to create in-house physical standards libraries, and because public databases contain a variety of experimental and empirical data for metabolite annotations, it is imperative that the evidence behind metabolite identifications and annotations are provided, and not only a list of peak matches. The use of standard compound identifiers (IDs) such as the human metabolome database (HMDB) identifier, Chemical Abstracts Service number, Refmet ID, and InChI key are needed to ensure harmonization of the data by ID rather than only compound name due to many synonyms that exist for chemicals. Labeling peak identifications and annotations of

isomeric compounds to reflect the base structure (e.g., with the reference list of metabolite IDs: Refmet IDs or InChI key) [67], rather than the specific isomeric forms, will increase harmonization across studies.

Advancing the field of metabolomics requires more rapid data processing for large-scale epidemiological studies, with improved algorithms for peak picking, data alignment, and metabolite identification and annotation of knowns and unknowns (e.g., nutritional and environmentally relevant dark matter). In recent years, open-source software (e.g., the Automated Data Analysis Pipeline [ADAP]) [34] has been developed that can handle large-scale epidemiology data that cannot be readily processed with instrument-level software, as well as the development of rapid algorithms for matching signals to millions of public database spectra [34, 68–70]. Additional advances are needed in pathway-mapping tools that extend past mapping signals to only host or microbial metabolites. Although metabolomics signatures are rich in metabolites of host and microbial metabolism, they are also rich in low molecular weight metabolites that are derived from the intake of drugs and medications, tobacco use, exposures to smoke and pollutants, exposure to chemicals in products used in the home and garden, or chemicals added to foods and beverages. Many of these signals may represent dark matter, e.g., in cases in which data for standards have not been included in accessible libraries. Current pathway-mapping tools focusing on host and microbial limit the ability to assess the impact of the exposome on human health and response to intervention.

Accelerometry data in obesity and nutrition research

Since 1983, accelerometers have been used in obesity and nutrition research to estimate physical activity-related measurements (e.g., physical activity-related energy expenditure, steps, and volume within and across intensity categories) in community-dwelling humans [71, 72]. This first use of accelerometers to measure physical activity has expanded into almost routine use of accelerometers in obesity and nutrition research and has been employed by the National Health and Nutrition Examination Survey (NHANES) during the 2003 to 2006 (waist-worn; waking hours only) and 2011 to 2014 cycles (wrist-worn; 24-h) [73, 74].

What does raw accelerometry data look like, and why are they not AI-ready?

Accelerometers produce high-velocity data because the sampling frequency is typically set to be at least twice of that of the highest biomechanically possible [75]. If the accelerometer is initialized to sample at 100 Hz, it means that measurements are obtained at 100 evenly spaced intervals every second. This means that 10 participants wearing an accelerometer for 3 days (72 h) will result in 259,200,000 data points. Some studies such as the UK Biobank Study (collected at 100 Hz for over 100,000 participants) [76] and the 2011 to 2014 NHANES cycles (collected at 80 Hz for over 18,000 participants) [77]

include large population sample sizes, vastly increasing data volume. The key to analyzing this volume of data is preprocessing that distills the data into forms that generate AI-ready datasets.

What are the steps to transform raw accelerometry data into AI-ready data?

Commercially available devices such as the Fitbit [78, 79] or Garmin [79] transform raw triaxial accelerometry data to display for consumers a variety of physiologically meaningful variables typically summarized per 24-h period, such as steps, sleep duration, sleep quality, and time spent at different activity levels. Additionally, heart rate and pulse oximetry are typically measured using photoplethysmography with light and photodetectors to measure variations of blood circulation [80]. Each device uses proprietary algorithms to preprocess the data, and, surprisingly, some of the distilled metrics differ greatly among devices and clinical measurements [79]. On the other hand, raw data from research grade accelerometers such as the ActiGraph GT3X and GT9X can be distilled into AI-ready datasets that require the user to transform the data.

Raw accelerometer data can be transformed into AI-ready data using the commercial software that includes the internal algorithms within the software [81], such as ActiLife (or CentrePoint) or an open-access software such as the R package GGIR [35]. Raw accelerometer signals may also be used directly as inputs into ML models (e.g., deep neural networks) [82]. In the case in which ML is used to transform raw accelerometer data, human judgment needs to be used to correctly label the input data and evaluate output data [30]. In the case in which a package such as GGIR is used, human decisions are made in order to set parameters.

For example, aggregation of the raw accelerometer data into manageable intervals called epochs along with what time-window constitutes a day are some of the human decisions required. An epoch set at one minute would average the raw data within one minute. The decision of how to set the time-window for the epoch can change the values in final AI-ready data for physical activity metrics (e.g., physical activity bouts) [83] or sleep metrics [84]. Human judgment may also be required to determine parameters related to sleep algorithms and physical activity algorithms. Specifically, thresholds need to be set to distinguish non-wear time from sleep [85, 86], and similar thresholds need to be defined to classify each epoch as time spent within a distinct physical activity intensity category (e.g., sedentary, light, moderate, and vigorous-intensity physical activity) [87]. Finally, decisions involving QC of the accelerometry data, such as what thresholds are needed for non-wear time, to classify when data should be retained for metrics aggregated in a day (e.g., steps per day) [88]. A pipeline flow diagram summarizing the preprocessing steps that do not rely on an ML classifier appear in Figure 3.

ML/AI methods may also be used for more detailed analysis of physical activity from raw accelerometry data, such as recognition of the type of activity being performed (e.g., sitting, standing, walking, eating) and estimation of related metrics such as

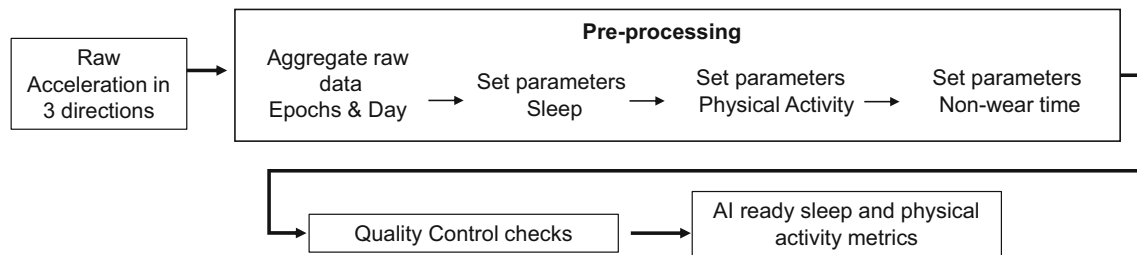


FIGURE 3 Pipeline flowchart depicting the transformation from raw accelerometry data into data that are artificial intelligence (AI)-ready.

energy expenditure [82]. Development and validation of such ML/AI methods rely on accurate and time-aligned labeling of raw accelerometer signals performed by a human annotator or derived from a reference instrument such as an indirect calorimeter. In this case, the labeling becomes a vital part of preparing an AI-ready dataset.

Despite the increasing use of accelerometers within obesity and nutrition studies and the availability of open-source programs such as GGIR, transforming raw accelerometry data into AI-ready datasets continues to be a challenge for individual investigators involved in obesity research. Investigators can choose to use commercial product output; however, because the preprocessing steps and algorithms are proprietary and closed or semi-closed, there is little insight into the underlying mechanisms that may lead to the variation and precision of the output. Specifically, the accuracy of physical activity and sleep information are not transparent. On the other hand, using actigraphs and relying on software such as GGIR provides more transparency. However, although using GGIR may not require knowledge of the programming language R, adjusting parameters and running the GGIR program may be challenging for investigators who are unfamiliar with R.

One solution that would mitigate a requisite R programming background is to develop an interface for the GGIR package similar to the commercial product ActiLife [81]. Although incredibly challenging, integrating a research-ready open software that includes mature ML models for data distillation and a user-friendly graphical interface would increase the application of accelerometry for physical activity measurement and analysis. Despite the challenge to improve software such as GGIR in this manner, the existence of trained ML models and the community open-source GGIR framework suggests that such tool development is within reach.

Although most of our discussion involves collecting raw accelerometry data in a single study, accelerometry has been used in many existing studies that could be technically combined to generate a sample size large enough for quality ML modeling [89]. However, currently, there is no standardization of accelerometer placement that can impact estimations of sleep and physical activity metrics [90, 91]. In addition, there does not exist standardization of algorithms used to estimate sleep and physical activity metrics. If data are preprocessed and raw accelerometry data are unavailable, the final estimates of sleep and physical activity metrics differ and cannot be combined [92].

INTEGRATION OF BIG DATA INTO AN AI/ML MODEL

The purpose of transforming Big Data into AI-ready datasets is to ultimately use the AI-ready data to derive human interpretable conclusions [28]. However, even after we have access to AI-ready data, the volume and complexity of the data may be extremely large and unmanageable. Therefore, we still need to select which model variables, combination of model variables, or transformation of model variables (referred to as data features) will contribute meaningfully to model predictions.

In most cases, AI/ML models require large sample sizes [30]. The required sample size is a function of the model complexity. This is one of many reasons behind reducing the AI/ML model complexity. One way to make this reduction is by identifying the relevant features that contribute to predictions. Moreover, including data features that do not impact predictions can lead to poorer model performance.

There are many available methods for feature selection [93], and which methods to choose may be dependent on a multitude of factors specific to the application. We refer the reader to other sources for specific methods [93] and only note here that this is a critical step after data preprocessing that is required before input of AI-ready data into AI/ML models.

Combining multiple Big Data source to predict a primary end point

After feature selection has been performed on AI-ready data, a common next step is to combine all features into a tensor to predict a primary outcome [94]. A tensor is an n -dimensional array that contains the data for a single participant. A final input dataset with Big Data variables such as the microbiome, accelerometry, and metabolomics may also include traditional nutrition and obesity variables such as blood pressure and questionnaire data. The different multimodal data can present challenges for AI/ML modeling. For example, ordinal categorical data from questionnaires are usually modeled with logit models or other specific methods designed to handle this type of data [95]. Some data may be conveyed as relative abundances as in OTU or OGU tables [28]. Other data may be more familiar in obesity research, such as energy intake consumed in kilocalories per day or fat-free mass in kilograms. Merging these different data types into

one model can be complicated and requires care lest an input is entered “illegally” into a model. For example, some types of ML models (ones designed to minimize variance such as principal component analysis) [96] are designed for continuous variable input. Users can still enter ordinal categorical data and force a prediction even though the prediction and model may not make sense [97].

Recently, several studies have used random forests or gradient-boosted regression for predicting postprandial glycemic response from Big Data [1,2]. Random forests and gradient-boosted regression both depend on decision trees. Decision trees accept multimodal data inputs; however, applying a single decision tree alone often results in poor predictions. Both random forests and gradient-boosted regression improve accuracy by developing predictions over many decision trees and iterations [98].

Integration of multiple data sources collected over relatively long periods of time (weeks, months) also poses a challenge when accurate syncing of the predictor and response variables is required. Computerized instruments such as wearable accelerometers and continuous glucose monitors rely on embedded clocks that may drift several seconds per day if not periodically reset to a high-accuracy reference clock such as a Network Time Protocol (NTP) server or Global Positioning System (GPS) signal. The time-drift may aggregate over the period of the study and lead to changes in observed dynamics of predictor and response variables. For these types of scenarios, time-drift compensation should be an integral part of preparation of AI-ready data. The literature provides little available methods to synchronize devices, but an approach is to apply an algorithm to minimize the overall clock drift among devices [99]. Practices should be an integral part of preparation of AI-ready data.

The need for causal inference from Big Data

Although prediction and classification are core AI/ML tasks, they can be limiting for translational purposes. For example, if we learn that frequent consumption of frozen foods predicts a particular health outcome, this could be explained by an unmeasured factor influencing both this food choice and the outcome, and, even when consuming the food does cause the outcome, it could be explained by many mechanisms (e.g., high sodium in frozen foods, high calorie content, low fiber). Without this understanding, it becomes challenging to deliver effective guidance to individuals or populations or to create public health policies to improve health. That is, without understanding whether there is a causal relationship and what its mechanism is, we cannot know whether it would be more effective to, for instance, reduce sodium content or increase fiber.


Establishing the causal link between diet and health is the holy grail for nutrition research, but large observational data bring challenges in establishing causality. Cross-sectional Big Data sources such as the NHANES [100] pose two core challenges for causal inference: the temporal order of exposure and outcome are unknown, and, because the data are a snapshot in time, each individual's cumulative or lifetime exposure is uncertain. However, more recent methods for automatically

obtaining dietary data through wearables (accelerometers, microphones, combinations of sensing modalities) [101–104] and image-capture modalities (glasses that capture video of food consumption and remote food photography) [8, 105, 106] are changing this. It is now possible to collect meal-level data along with activity and short-term responses (e.g., using continuous glucose monitoring to understand glycemic responses to individual meals), making it possible to disentangle the order of events and more closely tie exposures to outcomes. However, transforming these sources to AI-ready datasets require similar intense preprocessing steps to the three Big Data sources presented here. For example, images from food photography [105, 106] or food consumption video [8] represent unstructured data. Current preprocessing steps are not fully automated and require human judgment and extensive labor and effort to convert into AI-ready datasets.

Many methods exist for inferring causal relationships from observational data, including methods for static [107] and time series [108] data. Importantly, these methods can discover causal relationships from the data rather than testing hypotheses. However, they rely on strong assumptions, and the degree to which inferences can be trusted to be genuinely causal is proportional to the degree to which these assumptions can be trusted to be true. First, it is critical that all potential confounders (causes of two or more observed variables) are measured. If they are not, this may lead to spurious inferences. Although methods have been developed to handle latent variables, few have been successfully extended to time series data. Recent developments [109] have leveraged prior knowledge to address this, but such knowledge is not always available. Second, if observations are not representative of the true distributions of variables (e.g., sampling from a restricted range) or true dependencies, incorrect relationships may be identified, or correct ones may be missed. Individual methods may make other assumptions such as requiring causal relationships to be stationary (i.e., not changing over time). Often, the assumptions made by causal inference methods are violated in health data. Third, an underappreciated problem is the difficulty of determining the variables to be analyzed [110]. The correctness of inference and assumptions made are all relative to having the “right” set of variables, but, as discussed throughout, it is up to the researcher to determine whether, e.g., weight should be analyzed continuously, discretized into ranges, or transformed into body mass index. Similarly, there is not yet a method, to our knowledge, for conducting power analyses for causal inference. Nevertheless, as larger Big Data nutrition sources become available and causal inference methods continue to advance and loosen their assumptions, we have the potential for going beyond prediction from Big Data to causation.

CONCLUSION

Big Data have the exciting potential for identifying new modifiable factors that impact obesity and nutrition research. However, analyzing Big Data through AI/ML models requires the critical step of processing Big Data into AI-ready data followed by selection of the most necessary data features. Both require enormous time, effort, and

human judgment. In order to ensure accurate rigorous interpretation and implementation of conclusions arising from Big Data, the steps and choices involved in transforming Big Data into AI-ready data need to be transparent and accessible to investigators and clinicians relying on the conclusions. 

AUTHOR CONTRIBUTIONS

Diana M. Thomas conceived the idea for this study and drafted the overall manuscript. Diana M. Thomas and Samantha Kleinberg identified the need and conclusions of this study. Rob Knight, Jack A. Gilbert, and Andrew Pfluger developed drafts of the microbiome section. Diana M. Thomas and Edward Sazonov developed the section on the definition of AI-ready. Susan C.J. Sumner and Wimal Pathmasiri drafted the metabolomics section. Kelley Pettee Gabriel, Erin E. Dooley, and Edward Sazonov developed the accelerometry section. Samantha Kleinberg developed drafts of the causal inference section. Marilyn C. Cornelius contributed to and reviewed multiple drafts of the microbiome and metabolomics sections. Andrew Pfluger reviewed and revised the microbiome section. Andrew Pfluger, Susan C.J. Sumner, and Edward Sazonov developed the summary table. Kevin Cumiskey reviewed all mathematics/statistics described in the manuscript. MAG wrote the first draft of definitions of big data. Marie G. Gantz generated the first drafts of definitions of Big Data, artificial intelligence, and machine learning. Marie G. Gantz and Kate Burdekin reviewed multiple versions of the data integration section. All authors reviewed multiple drafts of the manuscript.

FUNDING INFORMATION

Diana M. Thomas and Kevin Cumiskey were supported by the National Institutes of Health (NIH) Interagency Agreement AOD22022001. Samantha Kleinberg was supported in part by the NIH under U54TR004279. Marie G. Gantz was supported by U24HD107676. Edward Sazonov was supported by U24CA268228, and Rob Knight and Jack A. Gilbert were supported by 1U24DK131617-01. Kelley Pettee Gabriel and Erin E. Dooley were supported by UG1HD107688. Andrew Pfluger was supported by the US Department of Defense Environmental Security Technology Certification Program (ESTCP) grant EW22-7278, and Susan C. J. Sumner was supported by U24CA268153.

CONFLICT OF INTEREST STATEMENT

The authors declared no conflict of interest.

ORCID

Diana M. Thomas  <https://orcid.org/0000-0003-2641-9304>

Mark A. Green  <https://orcid.org/0000-0002-0942-6628>

REFERENCES

- Zeevi D, Korem T, Zmora N, et al. Personalized nutrition by prediction of glycemic responses. *Cell*. 2015;163(5):1079-1094. doi:10.1016/j.cell.2015.11.001
- Berry SE, Valdes AM, Drew DA, et al. Human postprandial responses to food and potential for precision nutrition. *Nat Med*. 2020;26(6):964-973. doi:10.1038/s41591-020-0934-0
- Corbin KD, Carnero EA, Dirks B, et al. Host-diet-gut microbiome interactions influence human energy balance: a randomized clinical trial. *Nat Commun*. 2023;14(1):3161. doi:10.1038/s41467-023-38778-x
- Shen X, Kellogg R, Panyard DJ, et al. Multi-omics microsampling for the profiling of lifestyle-associated changes in health. *Nat Biomed Eng*. 2024;8:11-29. doi:10.1038/s41551-022-00999-8
- Tilly C. The old new social history and the new old social history. *Review (Fernand Braudel Center)*. 1984;7(3):363-406.
- Pendyala V. The big data phenomenon. In: Pendyala V, ed. *Veracity of Big Data: Machine Learning and Other Approaches to Verifying Truthfulness*. Apress; 2018:1-15.
- Bonomi AG. Towards valid estimates of activity energy expenditure using an accelerometer: searching for a proper analytical strategy and big data. *J Appl Physiol* (1985). 2013;115(9):1227-1228. doi:10.1152/jappphysiol.01028.2013
- Doulah A, Ghosh T, Hossain D, Imtiaz MH, Sazonov E. "Automatic ingestion monitor version 2" – a novel wearable device for automatic food intake detection and passive capture of food images. *IEEE J Biomed Health Inform*. 2021;25(2):568-576. doi:10.1109/JBHI.2020.2995473
- Doherty AR, Kelly P, Kerr J, et al. Using wearable cameras to categorise type and context of accelerometer-identified episodes of physical activity. *Int J Behav Nutr Phys Act*. 2013;10(1):22. doi:10.1186/1479-5868-10-22
- Farooq M, Doulah A, Parton J, McCrory MA, Higgins JA, Sazonov E. Validation of sensor-based food intake detection by multicamera video observation in an unconstrained environment. *Nutrients*. 2019;11(3). doi:10.3390/nu11030609
- Hossain D, Ghosh T, Sazonov E. Automatic count of bites and chews from videos of eating episodes. *IEEE Access*. 2020;8:101934-101945. doi:10.1109/access.2020.2998716
- Lindquist J, Thomas DM, Turner D, Blankenship J, Kyle TK. Food for thought: a natural language processing analysis of the 2020 dietary guidelines public comments. *Am J Clin Nutr*. 2021;114:713-720.
- Andreu-Perez J, Poon CC, Merrifield RD, Wong ST, Yang GZ. Big data for health. *IEEE J Biomed Health Inform*. 2015;19(4):1193-1208. doi:10.1109/jbhi.2015.2450362
- Kitchin R. *The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences*. 2nd ed. Sage Publications Ltd.; 2021.
- boyd d, Crawford K. Critical questions for Big Data. *Inf Commun Soc*. 2012;15(5):662-679. doi:10.1080/1369118X.2012.678878
- Haenlein M, Kaplan A. A brief history of artificial intelligence: on the past, present, and future of artificial intelligence. *Calif Manage Rev*. 2019;61(4):5-14. doi:10.1177/0008125619864925
- Turing A. Computing machinery and intelligence. *Mind*. 1950;LIX(236):433-460.
- Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Dev*. 2000;44(1/2):206-226.
- Campeato O. *Artificial Intelligence, Machine Learning, and Deep Learning*. Mercury Learning and Information; 2020.
- Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30:6000-6010.
- Mitchell TM, Carbonell JG, Michalski RS, eds. *Machine Learning: A Guide to Current Research*. Springer; 1986.
- Russell S, Norvig P. *Artificial Intelligence: A Modern Approach*. 3rd ed. Prentice Hall; 2010.
- Holmström J. From AI to digital transformation: the AI readiness framework. *Bus Horiz*. 2022;65(3):329-339. doi:10.1016/j.bushor.2021.03.006
- Kirkpatrick C. FARR: FAIR in ML, AI Readiness, & Reproducibility Research Coordination Network. Presented at: Accelerating and Deepening Approaches to FAIR Data Sharing: A Workshop; April 20, 2023; Washington, DC. <https://www.nationalacademies.org/event/04-20-2023/accelerating-and-deepening-approaches-to-fair-data-sharing-a-workshop>

25. Robinson S, Rao D, Redmon R, Kihn E. *Developing Community Guidelines for AI-Ready Data*. Presented at: NCEI Science Council Meeting; April 14, 2022. https://cdn.iioos.noaa.gov/media/2022/05/Robinson_2022-06-15_IIOOS_AI-Ready.pptx
26. UK Office for Artificial Intelligence. National AI strategy. Published September 2021. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1020402/National_AI_Strategy_-_PDF_version.pdf
27. Zhu Q, Huang S, Gonzalez A, et al. Phylogeny-aware analysis of metagenome community ecology based on matched reference genomes while bypassing taxonomy. *mSystems*. 2022;7(2):e0016722. doi:10.1128/msystems.00167-22
28. Estaki M, Jiang L, Bokulich NA, et al. QIIME 2 enables comprehensive end-to-end analysis of diverse microbiome data and comparative studies with publicly available data. *Curr Protoc Bioinformatics*. 2020;70(1):e100. doi:10.1002/cpbi.100
29. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J*. 2017;11(12):2639-2643. doi:10.1038/ismej.2017.119
30. Thomas DM, Kleinberg S, Brown AW, et al. Machine learning modeling practices to support the principles of AI and ethics in nutrition research. *Nutr Diabetes*. 2022;12(1):48. doi:10.1038/s41387-022-00226-y
31. LeVatte M, Keshteli AH, Zarei P, Wishart DS. Applications of metabolomics to precision nutrition. *Lifestyle Genom*. 2022;15(1):1-9. doi:10.1159/000518489
32. de Toro-Martín J, Arsenault BJ, Després JP, Vohl MC. Precision nutrition: a review of personalized nutritional approaches for the prevention and management of metabolic syndrome. *Nutrients*. 2017;9(8). doi:10.3390/nu9080913
33. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. *Crit Care*. 2023;27(1):120. doi:10.1186/s13054-023-04393-x
34. Du X, Smirnov A, Pluskal T, Jia W, Sumner S. Metabolomics data preprocessing using ADAP and MZmine 2. *Methods Mol Biol*. 2020;2104:25-48. doi:10.1007/978-1-0716-0239-3_3
35. Migueles JH, Rowlands AV, Huber F, Sabia S, van Hees VT. GGIR: a research community-driven open source R package for generating physical activity and sleep outcomes from multi-day raw accelerometer data. *J Measure Phys Behav*. 2019;2(3):188-196. doi:10.1123/jmpb.2018-0063
36. Van Hul M, Cani PD. The gut microbiota in obesity and weight management: microbes as friends or foe? *Nat Rev Endocrinol*. 2023;19(5):258-271. doi:10.1038/s41574-022-00794-0
37. Gill SR, Pop M, DeBoy RT, et al. Metagenomic analysis of the human distal gut microbiome. *Science*. 2006;312(5778):1355-1359. doi:10.1126/science.1124234
38. Lundberg DS, Yourstone S, Mieczkowski P, Jones CD, Dangl JL. Practical innovations for high-throughput amplicon sequencing. *Nat Methods*. 2013;10(10):999-1002. doi:10.1038/nmeth.2634
39. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol*. 2017;35(9):833-844. doi:10.1038/nbt.3935
40. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: an overview. *Hum Immunol*. 2021;82(11):801-811. doi:10.1016/j.humimm.2021.02.012
41. Wang Y, Zhao Y, Bolas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol*. 2021;39(11):1348-1365. doi:10.1038/s41587-021-01108-x
42. Travers A, Muskhelishvili G. DNA structure and function. *FEBS J*. 2015;282(12):2279-2295. doi:10.1111/febs.13307
43. Cornish-Bowden A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res*. 1985;13(9):3021-3030. doi:10.1093/nar/13.9.3021
44. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*. 2009;38(6):1767-1771. doi:10.1093/nar/gkp1137
45. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884-i890. doi:10.1093/bioinformatics/bty560
46. Fullwood MJ, Wei CL, Liu ET, Ruan Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res*. 2009;19(4):521-532. doi:10.1101/gr.074906.107
47. Fišer Pečnikar Ž, Buzan EV. 20 years since the introduction of DNA barcoding: from theory to application. *J Appl Genet*. 2014;55(1):43-52. doi:10.1007/s13353-013-0180-y
48. Knight R, Vrbancac A, Taylor BC, et al. Best practices for analysing microbiomes. *Nat Rev Microbiol*. 2018;16(7):410-422. doi:10.1038/s41579-018-0029-9
49. Navas-Molina JA, Peralta-Sánchez JM, González A, et al. Advancing our understanding of the human microbiome using QIIME. *Methods Enzymol*. 2013;531:371-444.
50. Kuczynski J, Stombaugh J, Walters WA, González A, Caporaso JG, Knight R. Using QIIME to analyze 16S rRNA gene sequences from microbial communities. *Curr Protoc Microbiol*. 2012;Chapter 1:Unit 1E.5. doi:10.1002/9780471729259.mc01e05s27
51. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, et al. Evaluating the information content of shallow shotgun metagenomics. *mSystems*. 2018;3:e00069-18. doi:10.1128/msystems.00069-18
52. Karst SM, Ziels RM, Kirkegaard RH, et al. High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat Methods*. 2021;18(2):165-169. doi:10.1038/s41592-020-01041-y
53. Armstrong G, Martino C, Morris J, et al. Swapping metagenomics preprocessing pipeline components offers speed and sensitivity increases. *mSystems*. 2022;7(2):e01378-21. doi:10.1128/msystems.01378-21
54. Douglas GM, Beiko RG, Langille MGI. Predicting the functional potential of the microbiome from marker genes using PICRUSt. *Methods Mol Biol*. 2018;1849:169-177.
55. Zhou J, Hoen AG, Mcritchie S, et al. Information enhanced model selection for Gaussian graphical model with application to metabolomic data. *Biostatistics*. 2021;23(3):926-948. doi:10.1093/biostatistics/kxab006
56. Cirulli ET, Guo L, Leon Swisher C, et al. Profound perturbation of the metabolome in obesity is associated with health risk. *Cell Metab*. 2019;29(2):488-500.e2. doi:10.1016/j.cmet.2018.09.022
57. Sumner SCJ, McRitchie S, Pathmasiri W. Chapter 10 – metabolomics for biomarker discovery and to derive genetic links to disease. In: Caterina RDE, Martinez JA, Kohlmeier M, eds. *Principles of Nutrigenetics and Nutrigenomics*. Academic Press; 2020:75-79.
58. Everett JR, Holmes E, Veselkov KA, Lindon JC, Nicholson JK. A unified conceptual framework for metabolic phenotyping in diagnosis and prognosis. *Trends Pharmacol Sci*. 2019;40:763-773.
59. Astarita G, Langridge J. An emerging role for metabolomics in nutrition science. *J Nutrigenet Nutrigenomics*. 2013;6(4-5):181-200. doi:10.1159/000354403
60. Roberts LD, Souza AL, Gerszten RE, Clish CB. Targeted metabolomics. *Curr Protoc Mol Biol*. 2012;Chapter 30:Unit 30.2.1-24. doi:10.1002/0471142727.mb3002s98
61. Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA. Untargeted metabolomics strategies-challenges and emerging directions. *J Am Soc Mass Spectrom*. 2016;27(12):1897-1905. doi:10.1007/s13361-016-1469-y
62. González-Peña D, Brennan L. Recent advances in the application of metabolomics for nutrition and health. *Annu Rev Food Sci Technol*. 2019;10:479-519.
63. Sumner S. Precision nutrition and the environment. Presented at: Exploring the Exosome: Mayo Clinic Individualizing Medicine

- Conference; Nov 2-3, 2022; Rochester, Minnesota. <https://individualizedmedicineblog.mayoclinic.org/2022/10/10/researchers-investigate-precision-nutrition-to-improve-health-prevent-diseases/>
64. Amberg A, Riefke B, Schlotterbeck G, et al. NMR and MS methods for metabolomics. *Methods Mol Biol*. 2017;1641:229-258. doi:10.1007/978-1-4939-7172-5_13
 65. Wishart DS, Cheng LL, Copié V, et al. NMR and metabolomics-a roadmap for the future. *Metabolites*. 2022;12(8):678. doi:10.3390/metabo12080678
 66. Reinhold D, Pielke-Lombardo H, Jacobson S, Ghosh D, Kechris K. Pre-analytic considerations for mass spectrometry-based untargeted metabolomics data. *Methods Mol Biol*. 2019;1978:323-340. doi:10.1007/978-1-4939-9236-2_20
 67. Fahy E, Subramaniam S. RefMet: a reference nomenclature for metabolomics. *Nat Methods*. 2020;17(12):1173-1174. doi:10.1038/s41592-020-01009-y
 68. Jiang W, Qiu Y, Ni Y, Su M, Jia W, Du X. An automated data analysis pipeline for GC-TOF-MS metabolomics studies. *J Proteome Res*. 2010;9(11):5974-5981. doi:10.1021/pr1007703
 69. Smirnov A, Liao Y, Du X. Memory-efficient searching of gas-chromatography mass spectra accelerated by prescreening. *Metabolites*. 2022;12(6):491. doi:10.3390/metabo12060491
 70. Smirnov A, Liao Y, Fahy E, Subramaniam S, Du X. ADAP-KDB: a spectral knowledgebase for tracking and prioritizing unknown GC-MS spectra in the NIH's metabolomics data repository. *Anal Chem*. 2021;93(36):12213-12220. doi:10.1021/acs.analchem.1c00355
 71. Montoye HJ, Washburn R, Servais S, Ertl A, Webster JG, Nagle FJ. Estimation of energy expenditure by a portable accelerometer. *Med Sci Sports Exerc*. 1983;15(5):403-407.
 72. Troiano RP, McClain JJ, Brychta RJ, Chen KY. Evolution of accelerometer methods for physical activity research. *Br J Sports Med*. 2014;48(13):1019-1023. doi:10.1136/bjsports-2014-093546
 73. Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. *Med Sci Sports Exerc*. 2008;40(1):181-188. doi:10.1249/mss.0b013e31815a51b3
 74. Belcher BR, Wolff-Hughes DL, Dooley EE, et al. US population-referenced percentiles for wrist-worn accelerometer-derived activity. *Med Sci Sports Exerc*. 2021;53(11):2455-2464. doi:10.1249/mss.0000000000002726
 75. Marvasti FA, ed. *Nonuniform sampling: theory and practice*. Kluwer Academic/Plenum Publishers; 2001.
 76. Doherty A, Jackson D, Hammerla N, et al. Large scale population assessment of physical activity using wrist worn accelerometers: the UK biobank study. *PLoS One*. 2017;12(2):e0169649. doi:10.1371/journal.pone.0169649
 77. Centers for Disease Control and Prevention. National Health and Nutrition Examination Survey 2013-2014 Data Documentation. Codebook, and Frequencies. Physical activity monitor - hour. Published November 2020. https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/PAXHR_H.htm
 78. Montgomery-Downs HE, Insana SP, Bond JA. Movement toward a novel activity monitoring device. *Sleep Breath*. 2012;16(3):913-917. doi:10.1007/s11325-011-0585-y
 79. Fuller D, Colwell E, Low J, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. *JMIR Mhealth Uhealth*. 2020;8(9):e18694. doi:10.2196/18694
 80. Castaneda D, Esparza A, Ghamari M, Soltanpur C, Nazeran H. A review on wearable photoplethysmography sensors and their potential future applications in health care. *Int J Biosens Bioelectron*. 2018;4(4):195-202. doi:10.15406/ijbsbe.2018.04.00125
 81. ActiLife Software. <https://actilife.theactigraph.com/support/software/actilife/>
 82. Nunavath V, Johansen S, Johannessen TS, et al. Deep learning for classifying physical activities from accelerometer data. *Sensors (Basel)*. 2021;21(16):5564. doi:10.3390/s21165564
 83. Ayabe M, Kumahara H, Morimura K, Tanaka H. Epoch length and the physical activity bout analysis: an accelerometry research issue. *BMC Res Notes*. 2013;6:20. doi:10.1186/1756-0500-6-20
 84. Berger AM, Wielgus KK, Young-McCaughan S, Fischer P, Farr L, Lee KA. Methodological challenges when using actigraphy in research. *J Pain Symptom Manage*. 2008;36(2):191-199. doi:10.1016/j.jpainsymman.2007.10.008
 85. Esliger DW, Rowlands AV, Hurst TL, Catt M, Murray P, Eston RG. Validation of the GENEA accelerometer. *Med Sci Sports Exerc*. 2011;43(6):1085-1093. doi:10.1249/MSS.0b013e31820513be
 86. van Hees VT, Sabia S, Anderson KN, et al. A novel, open access method to assess sleep duration using a wrist-worn accelerometer. *PLoS One*. 2015;10(11):e0142533. doi:10.1371/journal.pone.0142533
 87. Vanhelst J, Béghin L, Turck D, Gottrand F. New validated thresholds for various intensities of physical activity in adolescents using the Actigraph accelerometer. *Int J Rehabil Res*. 2011;34(2):175-177. doi:10.1097/MRR.0b013e31823840129e
 88. Choi L, Ward SC, Schnelle JF, Buchowski MS. Assessment of wear/nonwear time classification algorithms for triaxial accelerometer. *Med Sci Sports Exerc*. 2012;44(10):2009-2016. doi:10.1249/MSS.0b013e318258cb36
 89. Giffen CA, Wagner EL, Adams JT, et al. Providing researchers with online access to NHLBI biospecimen collections: the results of the first six years of the NHLBI BioLINCC program. *PLoS One*. 2017;12(6):e0178141. doi:10.1371/journal.pone.0178141
 90. Zinkhan M, Berger K, Hense S, et al. Agreement of different methods for assessing sleep characteristics: a comparison of two actigraphs, wrist and hip placement, and self-report with polysomnography. *Sleep Med*. 2014;15(9):1107-1114. doi:10.1016/j.sleep.2014.04.015
 91. Knaier R, Höchsmann C, Infanger D, Hinrichs T, Schmidt-Trucksäss A. Validation of automatic wear-time detection algorithms in a free-living setting of wrist-worn and hip-worn ActiGraph GT3X+. *BMC Public Health*. 2019;19(1):244. doi:10.1186/s12889-019-6568-9
 92. Aguilar-Farias N, Peeters G, Brychta RJ, Chen KY, Brown WJ. Comparing ActiGraph equations for estimating energy expenditure in older adults. *J Sports Sci*. 2019;37(2):188-195. doi:10.1080/02640414.2018.1488437
 93. Stańczyk U, Jain LC, eds. *Feature Selection for Data and Pattern Recognition*. Springer; 2014.
 94. Smalter A, Huan J, Lushington G. Feature selection in the tensor product feature space. *Proc IEEE Int Conf Data Min*. 2009;1004-1009. doi:10.1109/icdm.2009.101
 95. Agresti A. *Analysis of Ordinal Categorical Data*. 2nd ed. Wiley; 2010.
 96. Jolliffe IT. *Principal Component Analysis*. 2nd ed. Springer; 2002.
 97. Dorman KS, Maitra R. An efficient k-modes algorithm for clustering categorical datasets. *Stat Anal Data Min*. 2022;15(1):83-97. doi:10.1002/sam.11546
 98. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning with Applications in R*. Springer; 2021.
 99. Bennett TR, Gans N, Jafari R. Multi-sensor data-driven: synchronization using wearable sensors. *Proc Int Symp Wearable Comput*. 2015:113-116.
 100. National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey. <https://www.cdc.gov/nchs/nhanes/index.htm>
 101. Scisco JL, Muth ER, Hoover AW. Examining the utility of a bite-count-based measure of eating activity in free-living human beings. *J Acad Nutr Diet*. 2014;114(3):464-469. doi:10.1016/j.jand.2013.09.017
 102. Amft O, Stäger M, Lukowicz P, Tröster G. Analysis of chewing sounds for dietary monitoring. *Proc ACM Int Conf Ubiquitous Comput*. 2005:56-72.

103. Mirtchouk M, McGuire DL, Deierlein AL, Kleinberg S. Automated estimation of food type from body-worn audio and motion sensors in free-living environments. *Proc Mach Learn Res*. 2019;106:641-662.
104. Mirtchouk M, Merck C, Kleinberg S. Automated estimation of food type and amount consumed from body-worn audio and motion sensors. *Proc ACM Int Conf Ubiquitous Comput*. 2016:451-462.
105. Martin CK, Miller AC, Thomas DM, Champagne CM, Han H, Church T. Efficacy of SmartLoss, a smartphone-based weight loss intervention: results from a randomized controlled trial. *Obesity (Silver Spring)*. 2015;23(5):935-942. doi:[10.1002/oby.21063](https://doi.org/10.1002/oby.21063)
106. Harray AJ, Boushey CJ, Pollard CM, et al. A novel dietary assessment method to measure a healthy and sustainable diet using the mobile food record: protocol and methodology. *Nutrients*. 2015;7(7):5375-5395. doi:[10.3390/nu7075226](https://doi.org/10.3390/nu7075226)
107. Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press; 2000.
108. Kleinberg S. *Causality, Probability, and Time*. Cambridge University Press; 2013.
109. Zheng M, Kleinberg S. Using domain knowledge to overcome latent variables in causal inference from time series. *Proc Mach Learn Res*. 2019;106:474-489.
110. Woodward J. The problem of variable choice. *Synthese*. 2016;193(4): 1047-1072. doi:[10.1007/s11229-015-0810-5](https://doi.org/10.1007/s11229-015-0810-5)
111. Stauffer E, Dolan JA, Newman R. CHAPTER 8 - gas chromatography and gas chromatography—mass spectrometry. In: Stauffer E, Dolan JA, Newman R, eds. *Fire Debris Analysis*. Academic Press; 2008:235-293.

How to cite this article: Thomas DM, Knight R, Gilbert JA, et al. Transforming Big Data into AI-ready data for nutrition and obesity research. *Obesity (Silver Spring)*. 2024;32(5): 857-870. doi:[10.1002/oby.23989](https://doi.org/10.1002/oby.23989)