Metabolomics is the field of “-omics” research concerned with the comprehensive characterization of the small low-molecular-weight metabolites in biological samples. In epidemiology, it represents an emerging technology and an unprecedented opportunity to measure environmental and other exposures with improved precision and far less measurement error than with standard epidemiologic methods. Advances in the application of metabolomics in large-scale epidemiologic research are now being realized through a combination of improved sample preparation and handling, automated laboratory and processing methods, and reduction in costs. The number of epidemiologic studies that use metabolic profiling is still limited, but it is fast gaining popularity in this area. In the present article, we present a roadmap for metabolomic analyses in epidemiologic studies and discuss the various challenges these data pose to large-scale studies. We discuss the steps of data preprocessing, univariate and multivariate data analysis, correction for multiplicity of comparisons with correlated data, and finally the steps of cross-validation and external validation. As data from metabolomic studies accumulate in epidemiology, there is a need for large-scale replication and synthesis of findings, increased availability of raw data, and a focus on good study design, all of which will highlight the potential clinical impact of metabolomics in this field.

Metabolomics is the field of “-omics” research concerned with the comprehensive characterization of the small low-molecular-weight metabolites in biological samples (urine, blood, saliva, tissues). Metabolites (e.g., peptides, oligonucleotides, sugars, organic acids, ketones, aldehydes, amines, amino acids, lipids, steroids, and alkaloids) are produced in the body as the result of chemical processes and also come from exogenous sources (e.g., diet, drugs, xenobiotics, or gut microbe-host co-metabolism). Therefore, metabolomics may provide a comprehensive assessment of gene actions, intrinsic metabolism, and environmental exposures ranging from dietary, gut microbial, and xenobiotic sources and their interplay (1–3).

Analogous to the genome or proteome, the metabolome refers to the complete complement of metabolites found in or produced by an organism. There are approximately 500 different histological cell types in the human body, and each one has its own gene expression, proteome, and metabolome. Therefore, humans have more than 500 dynamic cellular metabolomes (4). Currently, analyses of metabolites can detect several hundreds or thousands of small molecules, depending on the analytical platform, but no single technology can measure the complete metabolome. Nonetheless, the total number of human metabolites identified is relatively modest (the Human Metabolome Database (5) currently has 41,511 metabolite entries) compared with transcriptomics (∼85,000) and proteomics (>10,000,000), where more targets can be identified and quantified (6).

The metabolome may provide a high-resolution, multifactorial phenotypic signature of complex diseases. This profiling can provide important insights into the molecular pathways underlying complex biological processes, measure response and adherence to therapy, monitor disease progression, provide real-time information in surgical procedures and disease diagnostics, and provide a window on gut health.
and the metabolic consequences of infections, among many other possible applications (4). Until recently, the field of metabolomics was not well suited for large-scale epidemiologic data, reflecting the limited capacity of the analytical platforms and the processing requirements of the enormous amount of data generated. Advances in the application of metabolomics in large-scale epidemiologic research are now being realized through a combination of improved sample preparation and handling, automated data processing, improved quality control, and innovative use of multivariate statistical and chemometrics tools (1, 4, 7). The associated costs of metabolomic analysis are decreasing, but it is still prohibitive for large-scale epidemiologic data analyses (untargeted metabolomics epidemiologic studies to date have evaluated <6,000 individuals). Here, we review the state of the art of metabolomic profiling in epidemiologic studies along with several of challenges that the field is currently facing.

ANALYTICAL METHODS

Metabolomic profiling is often performed using either nuclear magnetic resonance (NMR) spectroscopy or mass spectrometry (MS). These methods separate individual signals or analytes by their resonance frequencies in a magnetic field or by their mass-to-charge ratio, resulting in a spectral profile of separation (6, 8). Both NMR and MS can identify and quantify a wide range of small molecules with good analytical precision, and both require a small amount of sample (10–400 μL). Their relative advantages and disadvantages have been discussed previously (1, 4). In summary, NMR is highly reproducible, requires little sample preparation, is nondestructive (no interference with sample), and is good for structure elucidation. On the other hand, NMR has low sensitivity, as it is orders of magnitude less sensitive than MS. MS meanwhile is less reproducible and more platform-dependent, destructive to the sample, and time consuming, and the data generated are more complex for data analysis than with NMR (8).

NMR and MS metabolic profiling can run in either targeted or untargeted mode. Targeted profiling separates a limited number of specific metabolites of known identity and is a more hypothesis-driven approach. Untargeted profiling involves the use of multiple analytical assays (NMR, MS) to measure as many metabolites as possible in a biological sample. In the latter approach, the chemical identity of the peaks may not be known a priori and chemical/spectral analysis must be performed post hoc to identify the molecular species. Once the profiling spectra have been obtained and the metabolites identified, pathway and network analyses can provide important information on underlying biochemical pathways and connectivities, thus enriching the biological context and clinical relevance. Untargeted analysis does not require an a priori hypothesis and can be used to discover novel metabolic associations and disease pathways. However, data density is high, and because analysis is not optimized for specific metabolites, metabolite identification and quantification may be difficult. Conversely, targeted mode is based on an a priori hypothesis and can be optimized to quantify metabolites of interest. Semi-targeted analysis is often used to describe measurement of a metabolite class (e.g., bile acids). This approach allows quantification of metabolites of interest and has the scope to discover new closely related metabolites associated with a disease phenotype. In an epidemiologic setting, the selection of the analysis mode should be driven by the hypothesis of the study; for example, when there are no candidate metabolites to be examined, an untargeted mode should be preferred in order to discover novel associations between biomarkers and disease phenotypes. A combination of sequential untargeted and targeted analysis may enhance metabolite selection and identification in epidemiologic studies. Initially, a range of features can be selected from the untargeted analysis and identified through assay-specific databases. Subsequently, those can be validated with targeted triple-quadrupole MS using an authentic stable isotope–labelled standard, which gives absolute quantitation of the metabolite level. This approach involves validation of the metabolite chemical identification but also a quantification to perform dose-response analysis.

Metabolomics can be performed on a wide variety of biological samples commonly collected in epidemiologic studies. Appropriate collection, handling, and storage of the samples is critical to metabolomics analyses, as the methods are sensitive to small changes in the metabolite profile that may be introduced through poor sample handling procedures (9–12). This is particularly important in relation to epidemiologic studies because samples regularly go under freeze-thaw cycles that may unpredictably affect the analytical results. Sample preservation and avoiding contamination are also crucial, especially as different studies have followed different sample handling protocols and used different preservatives and conditions to store samples (Table 1). Blood is the most commonly collected and stored biological fluid in epidemiologic studies and has been the most often used sample in metabolomic analyses to date. As blood components are under tight homeostatic regulation, the extent of variation in blood metabolite concentration is limited. Urine samples represent a good alternative to blood and have greater capture of exogenous compounds such as microbiota, drugs, and diet, and urine composition can vary a lot, especially in disease states (1). A 24-hour collection is preferred over spot urine collection because it provides the complete picture of cumulative metabolite excretion over the 24-hour period. However, 24-hour samples are rarely available for epidemiologic studies. Urine is easier to deal in term of analytical chemistry than is blood, as blood needs more sample preparation and is more prone to experimental error. Other specimens, such as saliva, are gaining popularity but still have not been used in large-scale epidemiologic studies for metabolomics analyses, and their analytical properties have been less studied (13, 14).

A number of commercial companies (e.g., Metabolon, Biocrates, Chenomx, Metabometrix), in addition to well-established research laboratories (e.g., the MRC-NIHR Pheno- neme Centre at Imperial College London, United Kingdom, and The Broad Institute in the United States) are offering metabolomics analyses for large-scale epidemiologic studies. In contrast to genomics, where the measurement and interpretation are standardized across different companies, in metabolomics each company uses different approaches in relation to assays, quality control, data preprocessing, and metabolite quantification and identification, which leads to
Table 1. Challenges in Metabolomics Research Associated With Different Stages of Research

<table>
<thead>
<tr>
<th>Component</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study design</td>
<td>Lack of detailed sample collection protocol</td>
</tr>
<tr>
<td></td>
<td>Sample contamination</td>
</tr>
<tr>
<td></td>
<td>Number of freeze-thaw cycles</td>
</tr>
<tr>
<td></td>
<td>Application of multiple analytical platforms (nuclear magnetic resonance, liquid chromatography mass spectrometry, ultra-performance liquid chromatography mass spectrometry) in untargeted mode to increase the coverage of detected metabolites</td>
</tr>
<tr>
<td></td>
<td>Power calculations when number of features is unknown (untargeted analysis)</td>
</tr>
<tr>
<td>Study reporting and analysis</td>
<td>Reporting preprocessing methods</td>
</tr>
<tr>
<td></td>
<td>Reporting measures of quality control</td>
</tr>
<tr>
<td></td>
<td>Reporting method for outlier detection and number of outliers detected</td>
</tr>
<tr>
<td></td>
<td>Reporting methods for handling missing data</td>
</tr>
<tr>
<td></td>
<td>Data transformation</td>
</tr>
<tr>
<td></td>
<td>Estimation of metabolome-wide significance level, taking into account data dependence</td>
</tr>
<tr>
<td></td>
<td>Appropriate univariate and multivariate analysis</td>
</tr>
<tr>
<td></td>
<td>Cross-validation of results (account for over-fitting)</td>
</tr>
<tr>
<td></td>
<td>External independent validation of results</td>
</tr>
<tr>
<td></td>
<td>Reporting method for metabolite identification</td>
</tr>
<tr>
<td>Replication and data synthesis</td>
<td>Relative rather than absolute quantification of metabolites</td>
</tr>
<tr>
<td></td>
<td>Different methods for identification/quantification of metabolites between companies with a lack of information on how metabolites are identified/quantified</td>
</tr>
<tr>
<td></td>
<td>Inability to pool results between different companies</td>
</tr>
<tr>
<td></td>
<td>Need for access to the full data, protocols, and analysis codes for replication</td>
</tr>
<tr>
<td></td>
<td>Need for development of methods for meta-analysis</td>
</tr>
</tbody>
</table>

results that are difficult to harmonize and pool between different studies. Biocrates (Innsbruck, Austria) and Metabolon (Durham, North Carolina) apply tandem MS, Chenomx (Alberta, Canada) specializes in NMR, and Metabometrix includes both analyses. Metabometrix and Metabolon take an untargeted approach, whereas Biocrates applies targeted metabolomics. Suhre et al. (15) examined the number of metabolites common to Metabolon, Biocrates, and Chenomx; 482 distinct values of metabolite concentrations were available for analysis, but only 50 metabolites were quantified on more than 1 platform.

MAIN STEPS IN STATISTICAL ANALYSIS FOR EPIDEMIOLOGIC STUDIES

Metabonomic data have unique characteristics that pose special challenges in epidemiologic data analysis, including high dimensionality, use of multiple analytical methods (NMR, liquid chromatography MS, etc.), high degree of collinearity, some degree of missing data, nonlinearity, and non-normality. In addition, the number and the identity of most metabolites are unknown in untargeted mode, which is in contrast to other omic technologies, such as gene expression or sequence data. This also makes the power calculations difficult because the number of features in unknown and the number of test that will be performed cannot be estimated. Comprehensive reviews exist on statistical methods for metabolomics (16–22). In the present article, we will summarize the main steps in the process and the main statistical challenges involved in each step. Figure 1 gives an overview of the steps taken to analyze metabolomics data.

The first step is preprocessing, which is challenging and critical to final results and does not often receive adequate attention in epidemiologic studies (Table 1). In MS, preprocessing includes steps for peak detection, peak matching, retention time alignment, peak integration, and peak filling. Quality control analysis is also undertaken at this stage, usually by use of repeatability filters (e.g., filtering out all features with coefficient of variation >30% in quality control samples), and procedures should be clearly reported in epidemiologic analyses. Drift correction also needs to be done to correct feature-specific drift in measured intensities within a batch. NMR preprocessing needs to account for peak overlap and peak shift to aid peak interpretation (22, 23). Binning of metabolite signals based on correlation structure is commonly used to account for peak shift but does not overcome the problem of peak overlap (24). Peak fitting addresses both peak shift and overlap but is time consuming and does not take into account any unknown peaks, which results in information loss. As an alternative or addition to binning, high-resolution spectra can be used, although they are susceptible to peak shift and overlap (25). They retain all of the data and may be helpful for metabolite discovery and in interpretation of the data. Alignment involves shifting peaks to match common features in different spectra, a difficult task that often requires manual validation and can produce artifacts. Alignment is especially important when data from different cohorts analyzed in different batches are going to be synthesized. Normalization is common to NMR and MS and is used to account for uncontrolled metabolome-wide effects (e.g., dilution). Most instrument manufacturers provide their own software that covers the basic preprocessing (e.g., Fourier transformation in NMR or peak picking in MS), and some also provide tools for further downstream preprocessing described above. In NMR, examples of commercial software are the TopSpin and Amix packages from Bruker Biospin (Coventry, United Kingdom), manufacturers of NMR instruments. There are also third-party commercial and freeware applications such as iNMR (www.inmr.net), Chenomx (www.chenomx.com), and ProMetab (26). In MS, commercial examples include MassLynx and TransOminics from Waters (Milford, Massachusetts) (manufacturers of MS instruments). Freeware equivalents include XCMS (27), mzMine (28) and mzmMatch (29), which are all widely used. For both NMR and MS, preprocessing is a crucial procedure that significantly affects results and cannot be ignored. There is no standard workflow, largely because instrument technology continually evolves, requiring new approaches to be developed. An in-depth analysis of this is beyond the scope of this review; we merely emphasis that care should be taken to understand each step to ensure the validity of results. Identification of
outliers and other sources of variation should also be examined and clearly described and reported. Explanatory, unsupervised analysis is often used to allow a global overview of the data and the identification of outliers and examination of main sources of variation. Typical approaches include principal components analyses, which can produce low dimensional summaries of these complex data sets in which extreme outliers can be inspected visually or by statistical tests (e.g., Hotelling $T^2$) (30). Moderate outliers may not be shown in principal components analyses score plots but can be found in residual plots.

The next steps involve univariate and multivariate analyses to uncover metabolic signals that are associated with the phenotype under study. Linear or generalized linear models can be used for univariate analysis with adjustments for potential confounders. Partial least squares (PLS) regression (31) and its orthogonal filtered variants (referred to as O-PLS and O2-PLS) (32, 33) are commonly used in metabolomic analyses because of their ability to model highly collinear spectral data with appreciable amounts of both random and systematic noise. PLS models maximize the covariance between scores in $X$ (predictor) and $Y$ (outcome) spaces and in this way are able to account for both systematic variation in the metabolic profiles and correlations between the metabolic data and the outcome. PLS is often used in “discriminant analysis” mode to classify metabolic profiles according to a discrete set of categories (e.g., disease cases and controls) (34). However, the large sample sizes and presence of uncontrolled variation in epidemiologic studies often lead to complex PLS models that can be very hard to interpret. O-PLS and O2-PLS can help here by separating the variation into 2 parts (34). The first part, the “predictive” or “joint” variation, models variation that is common to both the $X$ (predictor) and $Y$ (outcome) blocks. For example, this could be variation in the metabolic profile that is linearly predictive of the disease class of an individual. The second part, the “orthogonal” or “unique” variation, models influences that are not linearly related to $Y$ (the outcome). This captures any extraneous factors, whether technical or biological, such as batch effects, sex, etc., that are not related to the outcome. Separation of the model into these 2 parts allows a more parsimonious explanation of the data because only the predictive part needs to be interpreted to understand the metabolic factors influencing the outcome. Nonetheless, confounding factors may have a component that is strongly collinear with the outcome, and this part will not be removed by orthogonal PLS methods. The main difference between O-PLS and O2-PLS is that in the latter, orthogonal components can be incorporated in the model of the $Y$ block (outcome). In this way, O2-PLS is an ideal method for integrating metabolomic data with other omic data, highlighting parts that are highly related between the different assays. PLS-based methods have found widespread application in metabolomics. However, apart from PLS, other multivariate and variable selection methods might have application in

Figure 1. Flow chart of the steps in data analysis of metabolomic data for epidemiologic studies. MS, mass spectrometry; NMR, nuclear magnetic resonance.
metabolomics, including penalized regression and Bayesian variable selection approaches (16, 35). Most multivariable methods can be run through R packages in CRAN (R Foundation for Statistical Computing, Vienna, Austria), whereas SIMCA-P software (Umetrix, Malmö, Sweden) is commonly used for participant component analyses and PLS-based approaches in metabolomics (36, 37).

Multivariate approaches suffer from over-fitting, and validation is an obligatory component of any analysis strategy using these methods (Table 1). Typically, cross-validation approaches are used in which a proportion of the data (e.g., 10%, the “test set”) are removed, and the model is built with the remaining “training set.” This procedure is repeated many times until each sample has been in the test set exactly once. The accuracy of the model on these left out samples gives an estimate of the predictive power for unseen samples and also the robustness of the model to perturbations of the data (38, 39). However, cross-validation methods do not guarantee good performance across different populations and often overestimate discriminatory performance between classes, probably because biases are introduced in the process. Independent external validation usually yields more conservative results, but even external validation may also be inflated because of optimism, selective reporting, and other biases (39, 40).

MULTIPLE TESTING CORRECTIONS

A key goal of statistical modeling in epidemiologic metabolomics studies is to interpret models in terms of metabolites associated with the outcome of interest. In univariate models, multiple testing corrections are essential. Bonferroni and false discovery rates procedures are commonly used; however, these often produce overly conservative thresholds because of the very high degree of correlation present in the metabolic data. Resampling methods such as the Metabolome Wide Significance Level (41) have been proposed to estimate the $P$ value distribution under the null hypothesis and estimate corrections appropriate to the sample type and assay method used. For example, in a study with 500 cases and equal number of controls, assuming 7,100 spectral variables, the metabolome-wide significance level was estimated at $P = 2 \times 10^{-5}$ ($\alpha = 5\%$), resulting in a 60% reduction in the effective number of tests compared with Bonferroni correction (41). For multivariate PLS-based approaches, a number of methods exist for selecting influential variables, including the variable importance in projection, bootstrap, jackknife, and permutation approaches, though no method currently dominates the field in this area (16).

RESULTS SYNTHESIS AND REPLICATION

Systematic reviews, meta-analyses, and field synopses of epidemiologic studies are currently lacking in the metabolomics field, in which replication efforts are still rare. The reason why large replication efforts are still rare has to do the nature of metabolomic data, which are not yet as standardized and homogenized to follow the example of large-scale genome-wide association study replication efforts in which data from different studies were able to be combined successfully (42). Quantification of the metabolites is nonstandardized between studies (relative rather than absolute quantification), which prohibits the synthesis of the effect sizes between different studies. In theory, one could perform meta-analyses that would combine $P$ values or ranks if the peaks could be aligned across studies, but this has not been successfully tried to date and it would require careful pre-alignment. In addition, procedures from different commercial companies or research laboratories follow different preprocessing strategies and lack information about how the metabolites are identified or quantified, which further introduces heterogeneity among results. Finally, in untargeted metabolomics, it is difficult to recognize the same features/signals among different studies based on the chemical shift or the mass-to-charge ratio and retention time, as those are dependent on analytical properties and will vary between studies unless results have been aligned in the preprocessing stage. Meta-analyses methods for metabolomics in epidemiologic research need to be developed. A database with results that focus on replication efforts, summary effect sizes, and quality appraisal, as in field synopsis for genomics, is not currently available. In the meantime, epidemiologic studies that aim to combine results or be used for discovery and subsequent replication need to follow the exact analytical procedures and the same preprocessing methods, including alignment between cohorts, to be able to cross-examine results between studies before structural identification of the metabolites.

WAYS TO IDENTIFY METABOLITES, SOLUTIONS TO “UNKNOWNs,” AND PATHWAY ANALYSIS

The aforementioned statistical analysis will eventually lead to the selection of a combination of features that are associated with the outcome of interest. When the analysis is run in untargeted mode, the structural identification of the metabolites that correspond to the selected peaks is essential. Assignment of individual metabolites to those peaks can be time consuming and challenging (Table 1), but online databases of annotated metabolites are continuously growing, thereby enabling faster and more reliable annotation. A number of catalogues of online databases are available that contain information on metabolite chemical characteristics and molecular data (Table 2).

Still, in any given epidemiologic study, the chemical identity of some detected metabolites will remain elusive; these are often described as “unknown metabolites.” Many recent metabolomic epidemiologic studies have reported a good proportion of their high-ranking hits to be unknown metabolites (e.g., 225 unknown compounds out of 517 measured compounds with ultra-performance liquid chromatography-mass spectrometry (MS) in the Cooperative Gesundheitsforschung in der Region Augsburg (KORA) Study; 75% unknown compounds in a diabetes case-control study), which hampers the interpretation and application of results (43, 44). Beyond the wide array of chemical analytic tools for identifying unknowns de novo, various computational methods have been developed. These include statistical correlation approaches, such as statistical total correlation spectroscopy (45), computerized prediction algorithms, and fragmentation trees, but these methods are still not efficient enough for high-throughput datasets, especially when commercial companies have carried out the
Table 2. Selection of Online Resources and Databases of Metabolites

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Metabolome Database</td>
<td>A freely available electronic database containing detailed information about small molecule metabolites found in the human body. The database is designed to contain or link 3 kinds of data: 1) chemical data, 2) clinical data, and 3) molecular biology/biochemistry data. The database currently contains 41,511 metabolite entries, including both water-soluble and lipid soluble metabolites, as well as metabolites that would be regarded as either abundant (&gt;1 μM) or relatively rare (&lt;1 nM).</td>
<td><a href="http://www.hmdb.ca/">www.hmdb.ca/</a></td>
</tr>
<tr>
<td>Human Metabolome Library</td>
<td>A chemical resource with the power to help confirm, validate, or quantify suspected metabolites in tissues or biofluids. Each of the 1,030 in-stock compounds were acquired through purchase, custom synthesis, or purification. Researchers may obtain metabolites from the Human Metabolome Library for basic research purposes.</td>
<td><a href="http://www.hmdb.ca/hml">www.hmdb.ca/hml</a></td>
</tr>
<tr>
<td>BiGG</td>
<td>A metabolic reconstruction of human metabolism designed for systems biology simulation and metabolic flux balance modeling. It is a comprehensive literature-based genome-scale metabolic reconstruction that accounts for the functions of 2,766 metabolites and 3,311 metabolic and transport reactions.</td>
<td><a href="http://bigg.ucsd.edu/">http://bigg.ucsd.edu/</a></td>
</tr>
<tr>
<td>MetaboLights</td>
<td>A database for metabolomic experiments and derived information. The database is cross-species and cross-technique and covers metabolite structures and their reference spectra, as well as their biological roles, locations, concentrations, and experimental data from metabolomic experiments.</td>
<td><a href="http://www.ebi.ac.uk/metabolights/index">http://www.ebi.ac.uk/metabolights/index</a></td>
</tr>
<tr>
<td>SetupX and BinBase</td>
<td>The SetupX system enables entering biological metadata for steering laboratory workflows by generating &quot;classes&quot; that reflect experimental designs. After data acquisition, a relational database system (BinBase) is used for automated metabolite annotation.</td>
<td><a href="http://fiehnlab.ucdavis.edu/projects/binbase_setupx">http://fiehnlab.ucdavis.edu/projects/binbase_setupx</a></td>
</tr>
<tr>
<td>Biological Magnetic Resonance Data Bank</td>
<td>A repository for data from nuclear magnetic resonance spectroscopy on proteins, peptides, nucleic acids, and other biomolecules.</td>
<td><a href="http://www.bmrb.wisc.edu/">http://www.bmrb.wisc.edu/</a></td>
</tr>
<tr>
<td>METLIN: Metabolite and Tandem MS Database</td>
<td>A metabolite database containing over 64,000 structures, with a data management system to assist in a broad array of metabolite research and metabolite identification by providing public access to its repository of current and comprehensive mass spectrometry/mass spectrometry metabolite data.</td>
<td><a href="http://metlin.scripps.edu/index.php">http://metlin.scripps.edu/index.php</a></td>
</tr>
<tr>
<td>KEGG: Kyoto Encyclopedia of Genes and Genomes</td>
<td>A resource for understanding high-level functions and utilities of the biological systems with large-scale molecular data sets generated by different high-throughput experimental technologies.</td>
<td><a href="http://www.genome.jp/kegg/">http://www.genome.jp/kegg/</a></td>
</tr>
<tr>
<td>Chemspider</td>
<td>A chemical structure database providing text and structure search access to over 29 million structures from hundreds of data sources.</td>
<td><a href="http://www.chemspider.com/">http://www.chemspider.com/</a></td>
</tr>
<tr>
<td>Metabolomics Workbench</td>
<td>An extensible informatics infrastructure which will serve as a US national metabolomics resource. The specific objectives of this infrastructure are to 1) develop a metabolomics data repository, tools, and interfaces, 2) develop a cloud computing infrastructure for metabolomics, 3) coordinate Regional Comprehensive Metabolomics Resource Cores and other metabolomics initiatives, and 4) provide the resources for advancing metabolomics education and research to enable new biomedicine.</td>
<td><a href="http://www.metabolomicsworkbench.org/">www.metabolomicsworkbench.org/</a></td>
</tr>
</tbody>
</table>

assays and the raw spectra are not available. Krumsiek et al. (43) have also provided an alternative strategy that integrates genome-wide association data, Gaussian graphical modeling, and metabolomics to predict the identity of unknown signals.

The list of metabolites associated with the outcome of interest can also be used to identify and interrogate the underlying biochemical pathways through automated analyses. For example, the Metabolite Set Enrichment Analysis (MSEA) follows the principles of gene enrichment analysis to help researchers identify and interpret patterns of human metabolite concentration changes (46). The Metscape database can be used to visualize and interpret metabolomic data in the context of human metabolic networks. The database allows users to trace the connections between metabolites and genes, visualize compound networks, and display compound structures, but it is only useful when metabolites have been identified (47). Other agnostic approaches, which deal with unknown features, have been proposed but have not yet been tested in epidemiologic settings (48).

CURRENT STATUS OF METABOLOMIC PROFILING IN EPIDEMIOLOGY

Metabolomics was initially used as a term in the biomedical literature in 1998, and since then, the term has been increasingly used, with the number of publications rising...
rapidly after 2006 (Figure 2). Still, the field is relatively small compared with other omics approaches ("genom*" retrieves 41,684 publications in 2013 alone). Cancer has received relatively more attention than cardiovascular disease among human studies of metabolomic profiling, but for both outcomes, the papers restricted to humans are relatively rare and do not exceed 1,200 as of November 2013, which is when our search was performed (see Figure 2 for search terms and source). Articles with primary data are only a small fraction of this number. The number of epidemiologic studies that use metabolic profiling is still limited and smaller than that for other omic technologies, but its use is gaining popularity quickly. In order to map the phenotypes already examined in metabolomics studies in diverse epidemiologic settings, we extracted the phenotypes described in the title of the metabolomics studies with epidemiologic study designs (n = 379) (Appendix Table 1). This crude analysis shows that the range of phenotypes already explored is remarkable, including (in addition to cardiovascular disease and cancer) type 1 and 2 diabetes, autoimmune diseases, pregnancy-related outcomes, respiratory disorders, diet and lifestyle characteristics, mood disorders, and many others (Appendix Table 1). Here, we provide a summary of the current status of metabolomics profiling in type 2 diabetes as an example. Type 2 diabetes is one of the most studied phenotypes within the cardiovascular group and includes large studies with both targeted and untargeted metabolomics measurements.

**Metabolomics and type 2 diabetes**

Metabolic profiling may offer information that can help in early detection, disease prediction, guided treatment, and understanding of molecular pathways in type-2 diabetes. A number of epidemiologic studies have already used targeted metabolomic approaches to examined associations between metabolomics and type 2 diabetes or related cardiometabolic traits, such as insulin resistance and obesity. In the Framingham Offspring Study, metabolic profiling was used to examine predictors of type 2 diabetes onset among 189 individuals who developed diabetes over 12 years of follow-up and

---

**Figure 2.** Metabolomics in medical literature. A) Number of publications in PubMed until November 2013 with terms metabolomic* OR metabonomic* OR metabolome by year of publication. B) Number of publications in PubMed with terms metabolomic* OR metabonomic* OR metabolome limited to humans and further limited to cardiovascular disease or cancer or diabetes or epidemiologic study designs (case-control OR cohort OR cross-sectional OR prospective OR epidemiologic study).
189 matched controls. The study profiled 48 amino acids, biogenic amines, and other polar plasma metabolites by liquid chromatography-MS. Five branched-chain and aromatic amino acids (isoleucine, leucine, valine, tyrosine, and phenylalanine) had statistically significant associations with future diabetes risk after Bonferroni correction, and the results were replicated in an independent, prospective cohort study (49). Another study examined a metabolomics approach to identify candidate biomarkers of pre-diabetes using a targeted liquid chromatography-MS method that quantified 188 metabolites. Three metabolites (glycine, lysophosphatidylcholine (18:2), and acetyl carnitine) had significantly altered levels in persons with impaired glucose tolerance as compared with those with normal glucose tolerance after Bonferroni correction, and again the results were replicated externally (50). Several other investigations on other epidemiologic studies have also identified the branched-chain and aromatic amino acids isoleucine, leucine, valine, tyrosine, and phenylalanine as being associated with the risk of future diabetes (49, 51–55). Nonetheless, these results have been based on so-called targeted approaches, which analyze a panel of preselected metabolites and thus pose no problems for biomarker identification. In addition, findings may underscore the potential key role of amino acid metabolism early in the pathogenesis of diabetes; however, such amino acids could have been measured by conventional means, and the added value of the metabolomics approach remains to be shown. As in other fields, mechanistic studies and integration of other omic data are still needed to suggest biological links between these biomarkers to the onset of type 2 diabetes. Studies have also developed risk models for type 2 diabetes with integration of metabolomics data and other risk factors (53). Despite the fact that these have shown high discrimination and very good calibration (area under the curve > 0.90), the models still require external independent validation and comparison with existing validated diabetes models (56–60).

Untargeted metabolomics approaches are still rare in epidemiologic research, but their application is increasing. In a proof-of-concept metabolome-wide association study, we demonstrated the potential of untargeted metabolic profiling (using NMR) in 4,630 twenty-four-hour urinary samples for the discovery of novel metabolites associated with blood pressure of individuals. The strongest association was observed for urinary formate excretion, which was inversely related to blood pressure of individuals (61). In relation to diabetes, Suhre et al. used a cross-platform (MS by 3 different commercial companies) approach in 40 diabetes cases and 60 controls and reported 32 associations after multiple testing correction (false discovery rate <5%) between 423 metabolites and type 2 diabetes (15). Menni et al. (62) used an untargeted metabolic approach that identified a range of 447 fasting plasma metabolites when searching for novel molecular markers that arise before and after hyperglycemia in 2,204 females from Twins UK. Overall, 42 metabolites were found to significantly correlate with type 2 diabetes after adjustment for multiple testing; of these, 17 had been previously reported as associated with either type 2 diabetes or insulin resistance. Among the metabolites identified, 3-methyl-2-oxovalerate was found to be the strongest predictive biomarker for impaired fasting glucose after glucose (odds ratio = 1.65, standard error, 0.14; \( P = 8.46 \times 10^{-6} \)) and was replicated in an independent population (\( n = 720 \), odds ratio = 1.68, 95% confidence interval: 1.34, 2.11; \( P = 6.52 \times 10^{-5} \)). Untargeted approaches showed a broad range of metabolites, revealing previously unknown metabolites and highlighting potential biomarkers for disease prediction and a deeper understanding of causal mechanisms in type 2 diabetes. However, these approaches also pose challenges as to metabolite selection and identification, and further studies are needed to explore replication, synthesis, and impact of current results (39).

CONCLUSIONS AND FUTURE PROSPECTS

Metabolomics represent an unprecedented opportunity in epidemiologic research that offers measurement of environmental and exogenous exposures with far less measurement error than with standard epidemiologic questionnaires. However, the aforementioned description of the metabolomics data handling highlighted numerous challenges. These challenges, summarized in Table 1, operate in all stages of research, from study design to analysis and reporting and finally to the replication of findings. An issue of special concern is the synthesis and large-scale replication of metabolicomic data from different studies. Studies often use different analytical platforms (between studies and within studies), and even within the same platforms commercial companies use different methods to identify and quantify metabolites, which make data harmonization difficult, though not impossible. Metabolomics in epidemiologic research is still in its early days, but many large-scale studies (\( n > 6,000 \)) with untargeted metabolic profiling using multiple platforms are now ongoing. At the same time, epidemiologic studies are becoming richer in other data, such as proteomics and epigenetics. In order for metabolicomic studies to provide results that are valid and reproducible, there is need for large coalition of teams and consortia with expertise in different fields to provide interdisciplinary solutions to complex data (63). At the same time, it is essential to make publicly available raw data, protocols, and analysis codes for these complex investigations to improve the transparency, reliability, and reproducibility of this research (7, 39, 40, 64). Finally, the discovery and replication of metabolicomic profiles in relation to disease outcomes may need to be followed by impact studies and randomized controlled trials to explore the potential clinical impact of metabolomic signatures.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology and Biostatistics, Imperial College London, London, United Kingdom (Ioanna Tzoulaki, Paul Elliott); MRC-PHE Centre for Environment and Health, Imperial College London, London, United Kingdom (Ioanna Tzoulaki, Paul Elliott); Computational and Systems Medicine, Department of Surgery and Cancer, Imperial College London, London, United Kingdom (Timothy M. D. Ebbels); Faculty of Medicine and Health Sciences, Nottingham City Hospital, Nottingham, United Kingdom.
Kingdom (Ana Valdes); Department of Medicine, Stanford Prevention Research Center, Stanford University School of Medicine, Stanford, California, (John P. A. Ioannidis); De-
partment of Health Research and Policy, University School of Medicine, Stanford, California, (John P. A. Ioannidis); Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, California, (John P. A. Ioannidis); and Meta-Research Innovation Center at Stanford, California, California, (John P. A. Ioannidis).

The Meta-Research Innovation Center is funded by a grant from the Laura and John Arnold Foundation.

Conflict of interest: none declared.

REFERENCES


34. Fonville JM, Richards SE, Barton RH, et al. The evolution of partial least squares models and related chemometric
objected-oriented algorithm for Bayesian stochastic search
cran.r-project.org/. Accessed May 1, 2014.
37. Umetrics. SIMCA: the standard in multivariate data analysis.
http://www.umetrics.com/products/simca. Accessed May 1,
2014.
38. Castaldi PJ, Dahabreh IJ, Ioannidis JP. An empirical assessment
of validation practices for molecular classifiers. Brief Bioinform.
39. Ioannidis JP, Khoury MJ. Improving validation practices in
Epidemiol. 2010;63(9):945–949.
profiling and the metabolome-wide association study: significance
42. Evangelou E, Ioannidis JP. Meta-analysis methods for
systems approach to metabolite identification combining genetic and
profiles reflective of glucose homeostasis in non-diabetic and
2010;5(12):e15824.
45. Cloarec O, Dumas ME, Craig A, et al. Statistical total
correlation spectroscopy: an exploratory approach for latent
biomarker identification from metabolic 1H NMR data sets.
46. Xia J, Wishart DS. MSEA: a web-based tool to identify
biologically meaningful patterns in quantitative metabolomic data.
plug-in for visualizing and interpreting metabolic data in the
26(7):971–973.
W481–W484.
49. Wang TJ, Larson MG, Vasan RS, et al. Metabolite profiles and
pre-diabetes identified by metabolomics. Mol Syst Biol. 2012;8:
615.
identifies pathways associated with metabolic risk in humans.
markers of the development of dysglycemia and type 2 diabetes
1730–1737.
metabolites associated with risk of type 2 diabetes using a
639–648.
54. Bain JR, Muehlbauer MJ. Metabolomics reveals unexpected
aromatic amino acids are predictors of insulin resistance in
56. Ioannidis JP, Tzoulaki I. What makes a good predictor? The
evidence applied to coronary artery calcium score. JAMA.
57. Siontis GC, Tzoulaki I, Ioannidis JP. Predicting death: an
58. Siontis GC, Tzoulaki I, Siontis KC, et al. Comparisons of
established risk prediction models for cardiovascular disease:
systematic review. BMJ. 2012;344:e3318.
59. Tzoulaki I, Liberopoulos G, Ioannidis JP. Assessment of claims
of improved prediction beyond the Framingham risk score.
60. Tzoulaki I, Liberopoulos G, Ioannidis JP. Use of
reclassification for assessment of improved prediction: an
phenotype diversity and its association with diet and blood
and impaired fasting glucose using a nontargeted metabolomics
epidemiology for 21st century medicine and public
508–516.
open-access general-purpose repository for metabolomics

(Appendix follows)
## Appendix Table 1. Phenotypes/Risk Factors as Described by the Studies on Metabolomics and Epidemiologic Study Designs

<table>
<thead>
<tr>
<th>Phenotype/Risk Factor by Category</th>
<th>Diet and lifestyle factors</th>
<th>Pregnancy-related factors and outcomes</th>
<th>Autoimmune disorders</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular disease</td>
<td>Smoking and smoking cessation</td>
<td>Intrauterine growth restriction</td>
<td>Systemic lupus erythematosus</td>
<td>Pediatric septic shock</td>
</tr>
<tr>
<td>Heart failure</td>
<td>Response to vitamin supplements</td>
<td>Gestational diabetes</td>
<td>Amyotrophic lateral sclerosis</td>
<td>Primary dysmenorrhea</td>
</tr>
<tr>
<td>Diabetes, types 1 and 2</td>
<td>Wine consumption</td>
<td>Pre-eclampsia</td>
<td>Rheumatoid arthritis</td>
<td>Autism spectrum disorders</td>
</tr>
<tr>
<td>Acute coronary syndrome</td>
<td>Physical activity level</td>
<td>Low birth weight</td>
<td>Inflammatory bowel disease</td>
<td>Bariatric surgery</td>
</tr>
<tr>
<td>Atherosclerosis</td>
<td></td>
<td>Preterm delivery</td>
<td></td>
<td>Tuberculosis</td>
</tr>
<tr>
<td>Lipids</td>
<td></td>
<td></td>
<td></td>
<td>Aging</td>
</tr>
<tr>
<td>Obesity</td>
<td></td>
<td></td>
<td></td>
<td>Polycystic ovary syndrome</td>
</tr>
<tr>
<td>Insulin resistance</td>
<td></td>
<td></td>
<td></td>
<td>Spinal muscular atrophy</td>
</tr>
<tr>
<td>Stent restenosis</td>
<td></td>
<td></td>
<td></td>
<td>Pancreatitis</td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td></td>
<td></td>
<td></td>
<td>Spinal muscular atrophy</td>
</tr>
<tr>
<td>Cancer</td>
<td></td>
<td></td>
<td></td>
<td>Nephropathy</td>
</tr>
<tr>
<td>Esophageal</td>
<td></td>
<td></td>
<td></td>
<td>Nonalcoholic fatty liver disease</td>
</tr>
<tr>
<td>Pancreatic</td>
<td></td>
<td></td>
<td></td>
<td>Cystic fibrosis</td>
</tr>
<tr>
<td>Colorectal</td>
<td></td>
<td></td>
<td></td>
<td>Periodontal disease</td>
</tr>
<tr>
<td>Breast</td>
<td></td>
<td></td>
<td></td>
<td>Human immunodeficiency virus infection</td>
</tr>
<tr>
<td>Endometrial</td>
<td></td>
<td></td>
<td></td>
<td>Sickle cell anemia</td>
</tr>
<tr>
<td>Prostate</td>
<td></td>
<td></td>
<td></td>
<td>Osteosarcoma</td>
</tr>
<tr>
<td>Bladder</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thyroid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gastric</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hepatocellular carcinoma</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malignant glioma</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acute myeloid leukemia</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respiratory illness</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asthma</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronic kidney disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mood disorders</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schizophrenia</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Major depressive disorder</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neurodegenerative diseases</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alzheimer’s disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parkinson’s disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a See Figure 1.