Microbiomics-focused Data Integration
A Fresh Solve for the Rubik’s Cube of Endophenotyping?

Traditionally, clinicians recognize that patients, even with the same disease, may demonstrate vastly differing clinical phenotypes: observable features and symptoms at the individual level that account for disease heterogeneity. Although clinical phenotyping has advanced respiratory medicine, it alone provides an incomplete picture in relation to disease-specific functional and/or pathobiological mechanisms (endotypes) that vary among individuals. An individual presenting with a particular clinical phenotype can demonstrate multiple endotypes, while a single endotype may be present across a range of clinical phenotypes. Respiratory diseases including severe asthma, chronic obstructive pulmonary disease (COPD), and bronchiectasis all demonstrate inherent disease heterogeneity accounted for by their various endophenotypes, which in some instances has precluded the success of randomized clinical trials (1–4).

One key approach to resolving endophenotyping in chronic respiratory disease is the use of multiomics. Our quest to realize precision medicine necessitates a combined view of clinical features and pathobiological assessment, including genomics, epigenomics, transcriptomics, metabolomics, lipidomics, proteomics, and most recently microbiomics. Large-scale assessment of multiomics, including the NHLBI’s Trans-Omics for Precision Medicine program (5) and the UK Biobank (6), provide vast data for clinicians and scientists to work with, but the key challenge is developing, selecting, and applying appropriate methodology to holistically view such “big data” at the personal level and allow bedside clinical translation.

Thus far, most endophenotyping efforts have focused on either a single omics or microbial kingdom, and therefore multiomics and/or microbiome assessment in respiratory disease remains limited. A multiomics approach integrating clinical parameters with the bacteriome, virome, transcriptome, and metabome in infants with respiratory syncytial virus identifies endotypes relating to risk of subsequent childhood asthma, while in bronchiectasis, microbiome (bacteria, viruses, and fungi) integration resolves a “high-risk” group characterized by increased exacerbations and intermicrobial cross-talk (7, 8). Importantly, microbiomics remains a field in rapid ascent, in which all microorganisms within a given community are assessed collectively. Considering strong emerging evidence for the importance of host–pathogen relationships in the pathogenesis and progression of chronic respiratory disease, studies assessing microbiomics in tandem with other host-derived omics are required.

In this issue of the Journal, two studies are presented, both using microbiomics-focused data assessment together with other omics to resolve endophenotypes in bronchiectasis and COPD (9, 10). First, Huang and colleagues (pp. 417–426) reveal that individuals with COPD and COPD associated with bronchiectasis present with variable lung microbiome and host response profiles and that the underlying pathophysiology of the latter is most closely related to bronchiectasis. Combining sputum bacteriomes with label-free proteomics, they propose five endotypes, according to features with known prognostic implications in COPD and bronchiectasis, respectively, and differing by microbiome/proteome profiles (9). Each endotype reveals varied mechanisms between patients (with the same disease): some with features leaning toward COPD and others toward bronchiectasis, with each speculated to be “targetable” by different treatment approaches (indicated in parentheses). These include 1) diverse protective (no treatment), 2) diverse T-helper cell type 2 (inhaled corticosteroids), 3) infected epithelial response (macrolides), 4) proteobacteria neutrophilic (macrolides or dipeptidyl peptidase-1 inhibitors), and 5) Haemophilus proteolytic (antifibrotic or antiinflammatory) groups (9). Critically, key findings related to the association of COPD with bronchiectasis were independently validated (9). Using a different approach, Madapoosi and colleagues (pp. 427–439) evaluate paired bacteriome and metabolomic data from BAL in ever-smokers in the SPIROMICS (Subpopulations and Intermediate Outcome Measures in COPD Study) cohort comparing individuals with or without mild to moderate COPD (i.e., Global Initiative for Chronic Obstructive Lung Disease groups 0–2) (10). Airway infection is linked to progressive COPD, but the role of airway microbes in milder disease has been unclear. SPIROMICS previously demonstrated, in independent studies, links between the lung microbiome and metabolome and clinical outcomes (11, 12). Here, presenting a combined, integrated analysis incorporating bacteriome sequencing, untargeted metabolomics (hydrophobic fraction of mainly lipids), and targeted metabolomics (mucointerfermatory panel of hydrophilic compounds), they reveal that specific microorganisms relate to COPD status, lung function, symptoms, and/or exacerbations, implying that lung bacteriomes and metabolomes, to different extents, influence pathogenesis in mild to moderate COPD (10). For instance, members of the Prevotella genus appear protective, demonstrating lower risks of COPD, higher lung function, and minimal symptoms. These bacteria are associated with metabolic signatures including adenosine, 5'-methylthioadenosine, sialic acid, tyrosine, and glutathione. Conversely, other bacterial groups, including Streptococcus, Neisseria, and Veillonella, accompanied by elevated glycosphingolipids, glycerophospholipids,
polyamines, and xanthine are linked to lower lung function, poorer COPD status, higher COPD assessment test scores, and chronic bronchitis. Taken together, these data, holistically integrated for the first time in COPD, offer a fresh view on pathogenesis in milder disease, suggesting complex interrelated relationships between airway microbes and their metabolites.

Significantly, both studies apply multiomics analysis to a single disease population and retain a microbiomics focus. This leverages the complementary information provided by the microbial (microbiome) and host (omic) components of their respective data sets to resolve endophenotypes, but each study uses a different approach. Huang and colleagues take a classical approach, using principal-component analysis (PCA), whereas Madapoosi and colleagues take a modern approach, exploiting data integration techniques (9, 10, 13). In doing so, both studies exercise a two-tiered analytical approach: first, Huang and colleagues perform an unsupervised dimensionality reduction by PCA on concatenated microbiome and proteomes to better understand the distribution of their study groups. Next, they implement partial least squares discriminant analysis (PLS-DA), a supervised dimensionality reduction technique that allows further dimensionality reduction of the concatenated data while maximizing difference between the PCA-chosen study groups. This approach offers several advantages, including easy-to-interpret results (on the basis of scores from PCA) and metabolome data sets. Subsequently, the selected features are used as predictors in an elastic net regression model for the prediction of clinical outcomes. Use of DIABLO (in contrast to PLS-DA) “integrates” various omics while concurrently performing discriminant analysis to identify correlated omics features explaining the dichotomized clinical measures (13). The use of predictive models such as elastic net regression allows confounder adjustment such as patient demographics and the use of balanced accuracy (to account for class imbalances), a common issue observed with clinical and/or biological data. However, a key limitation of this approach is the dilution of discriminatory effects (because of dichotomization of clinical measures) in DIABLO analysis. Furthermore, neither study fully accounted for the inherent compositionality structure of microbiome data, and importantly, the use of centered log ratios before computing correlations, as performed by Madapoosi and colleagues, is largely insufficient to achieve this (10). Other important limitations to consider in both studies include their cross-sectional nature, lack of patient “matching” between groups, and need for further external (independent) validation of the findings. There are established limitations to the use of targeted amplicon sequencing, and neither study presents data that may be extrapolated to the

Figure 1. Conceptualization of multiomics data integration using a Rubik’s cube analogy. An individual patient is represented as a single Rubik’s cube, with (A) a solved cube indicating “stable” disease (i.e., homogeneous coloration on each respective cube’s face) and (B) a scrambled cube indicating “active, exacerbated, or progressive” disease (i.e., nonhomogeneous coloration on each respective cube’s face). Independent of underlying disease and/or its state of activity or progression, several views can be taken of a single cube (or an individual patient), reflected by each respective view representing a single omics platform (i.e. view 1, view 2, view 3, etc.). No single view appreciates the overall cube in its entirety; hence an approach integrating multiomics data leveraging interrelationships among the various cube faces (or multiomics) is necessary.
exacerbation state of disease. Despite limitations, these works collectively advance our understanding of microbiomics-focused multiomics data integration, best understood by visualizing a Rubik’s cube (Figure 1).

Consider an individual cube a single patient, where a solved cube is that patient in a “stable” disease state (Figure 1A). A scrambled cube represents active, exacerbated, or even progressive disease in the same patient (Figure 1B). Therefore, irrespective of disease type and/or underlying state, there are several views one can take of an individual patient (akin to the variable faces of the cube), each represented by a single omics. Most available studies draw conclusions about patients using a single view (i.e., single-omics analysis). A Rubik’s cube cannot be solved in its entirety by focusing solely on one face; each move affects the other faces, confounding the overall strategy of solving the puzzle if not wholly considered. Similarly, individual variation and disease pathogenesis cannot be fully expounded by a single-omics approach. Therefore, a holistic approach integrating multiomics data and leveraging their interrelationships is necessary, one that provides greater relevance and resolution for endophenotyping. Various advantages of data integration are reported, including better precision and accuracy for patient stratification, increased statistical power to detect rarer patient subgroups, and an ability to assess information flow among various omics, capturing dynamic relationships among them (8, 14). Taking an integrated data approach exemplifies the established notion that the “whole” is always greater than the sum of each individual part and provides a fresh perspective on solving the Rubik’s cube of endophenotyping respiratory disease.

An increasing number of data integration techniques are in use for studying respiratory disease, including similarity network fusion, multiomics factor analysis, and DIABLO, and the field continues to grow (13–16). In recognition of the importance of translational science in this space, multiomics cohorts, including U-BIOPRED (Unbiased Biomarkers for the Prediction of Respiratory Disease Outcomes) for severe asthma, have produced results that clearly demonstrate the valuable insights gained from a multiomics approach (17). Strong study design, considering confounders, temporality, missing data, and normalization processes to allow comparability and optimize data quality, does pose significant challenges to multiomics studies, and to tackle the emerging need for data integration, we must overcome specific challenges. Most data integration tools are currently available only as source code and require computer programming expertise to understand and apply them in clinical settings. This may be overcome by iterative software tool development with a focus on accessibility to clinicians and that incorporates preprocessing techniques such as normalization, filtering, and feature selection (8). A conceptual shift is in the making, one whereby optimal endophenotyping of respiratory disease requires a combined approach of prospective clinical study coupled to integrated multiomics analysis. To realize precision medicine at the bedside, we will need to incorporate this fresh perspective on solving the Rubik’s cube of endophenotyping.

Author disclosures are available with the text of this article at www.atjcjournals.org.

Jayanth Kumar Narayana, B.S., M.S.
Lee Kong Chian School of Medicine
Nanyang Technological University
Singapore, Singapore

Krasimira Tsaneva-Atanasova, Ph.D.
College of Engineering, Mathematics and Physical Sciences
University of Exeter
Exeter, United Kingdom

Sanjay H. Chotirmall, M.D., Ph.D.
Lee Kong Chian School of Medicine
Nanyang Technological University
Singapore, Singapore

and

Department of Respiratory and Critical Care Medicine
Tan Tock Seng Hospital
Singapore, Singapore

ORCID ID: 0000-0003-0417-7607 (S.H.C.).

References

More than 137 million Americans live in counties with unhealthy air pollution concentrations (1). Globally, 4.2 million deaths are attributed to air pollution every year (2). Since the beginning of the coronavirus disease (COVID-19) pandemic, researchers have postulated a link between air pollution and COVID-19 severity. Air pollution is associated with development of cardiometabolic and lung disease, which are well-established risk factors for COVID-19 morbidity and mortality. Air pollution may facilitate entry of severe acute respiratory syndrome coronavirus 2 (SARS CoV-2) into the body, impair immune responses and viral clearance, and promote tissue damage (3). Indeed, early analyses suggested that a 1 microgram per cubic meter (μg/m³) increase in long-term county-level fine particulate matter (PM2.5) was associated with an 11% higher county-level COVID-19 mortality rate in the United States (4). These ecological analyses are subject to bias as they cannot adjust for individual-level risk factors such as poverty and access to healthcare, for example, that often covary with air pollution. Emerging evidence using individual-level data supports an association between long-term air pollution and COVID-19 severity; however, more evidence is required, particularly with respect to associations with short-term exposures (5–7).

In this edition of the Journal, Chen and colleagues (pp. 440–448) leverage a cohort of 74,915 patients with COVID-19 from Southern California to examine associations between residential address short-term (1-mo) and long-term (1-yr) PM2.5, ozone (O3) and nitrogen dioxide (NO2) concentrations and COVID-19 severity (8). COVID-19 diagnosis was made by SARS CoV-2 PCR or medical record diagnostic coding; outcomes included 30-day COVID-19 hospitalization, ICU admission, and intensive respiratory support (IRS, defined as invasive mechanical ventilation, noninvasive ventilation, high flow nasal cannula, or high flow mask), and 60-day mortality. The authors found that long- and short-term PM2.5 and NO2 exposures were independently associated with worse COVID-19 outcomes. In multipollutant models, long-term PM2.5 and short-term NO2 associations persisted—a 1.5 μg/m³ increase in PM2.5 exposure in the year preceding diagnosis was associated with 24%, 33%, and 32% higher odds of hospitalization, IRS, and ICU admission, respectively, and 14% higher risk of 60-day mortality. Separately, a 3.3 parts per billion (ppb) increase in NO2 in the month preceding diagnosis was associated with a 12%, 18%, and 21% higher odds of hospitalization, IRS, and ICU admission, respectively; a trend toward 7% higher mortality risk was observed.

A striking finding in this study is the lack of a “safe” level of air pollution. Air pollution concentrations were on average below U.S. Environmental Protection Agency (EPA) annual standards of 12 μg/m³ and 53 ppb for PM2.5 and NO2, respectively. Chen and colleagues present generalized additive models between long-term PM2.5 and odds of COVID-19 hospitalization that find linearity at concentrations far below the US EPA annual PM2.5 standard. One-month exposures were also low, and linear associations at low NO2 concentrations also observed. These data contribute to the evidence base supporting the harmful effects air pollution at levels below current EPA standards (6, 9).

This study occurs at a critical time. The COVID-19 pandemic will continue to impact human health. Policy is driving reductions in air pollution, but progress has been stilted particularly in the western United States due to climate-related increases in wildland fires. In 2020, the year from which the study acquired data, 4.3 million acres were burned in California during wildland fires (10). Over 60% of the cohort was diagnosed with COVID-19 in June and July 2020, during which time more than 125 wildland fires, including 45 large-scale (>300 acres) fires, were reported across California (10). In this context, the findings by Chen and colleagues are timely and concerning.

Given the mounting evidence of air pollution—a modifiable exposure—increasing COVID-19 severity, where do we go from here? We offer three actions.

First, we must continue to build a policy-relevant evidence base for the harms of short-term exposures. Chen and colleagues begin to tackle this issue by examining associations with 1-month averages; however, more evidence is required to support policy changes. In considering changes to the current air quality standards, the EPA cited limited evidence evaluating alternative forms of the short-term standard (11). Policy gaps such as this should be addressed by researchers.

Second, we must improve air quality monitoring especially for environmental justice communities in “hot-spot” locations that, due