

# Peripheral Blood Omics and Other Multiplex-based Systems in Pulmonary and Critical Care Medicine

Joseph Balnis<sup>1,2</sup>, Eitel J. M. Lauria<sup>3</sup>, Recai Yucel<sup>4</sup>, Harold A. Singer<sup>2</sup>, Reid S. Alisch<sup>5\*</sup>, and Ariel Jaitovich<sup>1,2\*</sup>

<sup>1</sup>Division of Pulmonary and Critical Care Medicine and <sup>2</sup>Department of Molecular and Cellular Physiology, Albany Medical College, Albany, New York; <sup>3</sup>School of Computer Science and Mathematics, Marist College, Poughkeepsie, New York; <sup>4</sup>Department of Epidemiology and Biostatistics, Temple University, Philadelphia, Pennsylvania; and <sup>5</sup>Department of Neurological Surgery, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin

## Abstract

Over the last years, the use of peripheral blood–derived big datasets in combination with machine learning technology has accelerated the understanding, prediction, and management of pulmonary and critical care conditions. The goal of this article is to provide readers with an introduction to the methods and applications of blood omics and other multiplex-based technologies in the pulmonary and critical care medicine setting to better appreciate the current literature in the field. To accomplish that, we provide essential concepts needed to

rationalize this approach and introduce readers to the types of molecules that can be obtained from the circulating blood to generate big datasets; elaborate on the differences between bulk, sorted, and single-cell approaches; and the basic analytical pipelines required for clinical interpretation. Examples of peripheral blood–derived big datasets used in recent literature are presented, and limitations of that technology are highlighted to qualify both the current and future value of these methodologies.

**Keywords:** precision medicine; big data; omics; peripheral blood; sequencing

Precision medicine, which is the development of diagnostic and therapeutic strategies that account for interindividual variability, has recently been expanded by the emergence of combined omic technologies; powerful methods to characterize an individual's molecular landscape, such as proteomics, metabolomics, and transcriptomics; and the creation of bioinformatic tools to analyze large datasets (1). The field of pulmonary and critical care medicine has seen substantial progress using nucleic acid and protein sequencing and other multiplex-based platforms to interrogate various patients' samples, and we focus in this perspective article on the use of peripheral blood samples to apply these tools. Although the peripheral blood provides indirect information that covaries with surrogates of healthy and diseased organs, it also entails

a powerful and relatively easy way to gain access to processes superseding a single organ or taking place at an inaccessible or otherwise distant tissue.

The goal of this article is to provide the reader with an introduction to the methods and applications of peripheral blood omic- and other multiplex-based technologies in the setting of pulmonary and critical care medicine to better appreciate the current literature in the field. As we elaborate later in the text, these applications currently exist at various technology readiness levels (2). Indeed, although some of them involve readily available clinical data in combination with relatively accessible cytokine concentration determinations, some others require less accessible technologies, such as RNA-sequencing analyses. The recent expansion of machine learning systems

anticipates an acceleration in these tools' development, and we hope that readers will improve their familiarity with an area that is likely to gain more relevance in coming years.

It is important to clarify that peripheral blood omic- and other multiplex-based technologies involve the identification of multiple, sometimes in the order of thousands, different molecules per patient. The clinical relevance of these molecules is typically unknown at the moment of their initial determination, and their association with a specific outcome requires bioinformatic tools, as we explain later in the article. For these reasons, these approaches should not be confused with the "liquid biopsy," which is the use of next-generation sequencing to detect tissue-specific molecules previously known to be associated with a diagnosis or an

(Received in original form April 27, 2023; accepted in final form June 28, 2023)

\*R.S.A. and A.J. are lead authors.

Supported in part by the National Heart, Lung, and Blood Institute through awards K01HL130704 and R01HL160661 (A.J.) and R01HL049426 (H.A.S.); by the National Institute of Allergy and Infectious Diseases through award 1R01AI173035 (A.J. and R.S.A.), and by the National Institute on Aging through award R01AG066179 (R.S.A.).

Correspondence and requests for reprints should be addressed to Ariel Jaitovich, M.D., Albany Medical College, 47 New Scotland Avenue, MC91, Albany, NY 12208. E-mail: jaitova@amc.edu.

Am J Respir Cell Mol Biol Vol 69, Iss 4, pp 383–390, October 2023

Copyright © 2023 by the American Thoracic Society

Originally Published in Press as DOI: 10.1165/rcmb.2023-0153PS on June 28, 2023

Internet address: www.atsjournals.org

outcome, such as in the case of tumor-specific circulating cell-free DNA (cfDNA) sequences that can be obtained from the plasma of patients with lung cancer (3). These circulating cell-free DNA fragments correspond to cancer-specific sequences coding for somatic mutations not expressed by any other tissue except for the transformed one.

We introduce big data essential concepts needed to understand the rationale of this approach, provide general examples of its applications, describe the types of molecules that are obtained from the circulating blood to generate big datasets, introduce the analytical pipeline essentials, and feature recent articles that illustrate the topic. The featured articles only provide examples of peripheral blood-derived big datasets used in pulmonary and critical care medicine and are not meant to discuss their content or implications exhaustively. Finally, we outline the limitations of these methodologies to provide readers with a balanced understanding of their current value, and we suggest future directions that we believe could accelerate the massive and efficient use of them.

## Big Data Essential Concepts

Samples from the circulatory compartment can be processed to generate large, complex datasets that need special software and bioinformatic expertise for analysis. “Big data” defines these large and complex

datasets, whereas the term “omics” combined with a prefix descriptor defines the molecular source of big data: genomics for DNA-sequencing data, transcriptomics for RNA transcript data, and so forth (Figure 1). Although the term “omics” generally applies to sequencing or otherwise unbiased large-scale studies, for simplicity, in this article, we describe omics together with other targeted data collection methods such as cytokine multiplex panels and microarrays, which are likewise used to generate datasets from blood samples and investigate pulmonary and critical care medicine processes (Figure 2).

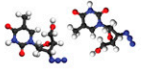



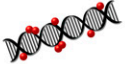
Individual datapoints, such as nucleotides or transcripts, are defined as *features*, and features from multiple sources, such as RNA and protein sequencing datasets, can be correlated to produce multiomic feature associations, which sometimes may suggest causality (4–6). For example, the upregulation of both the RNA transcript and its corresponding protein product and their correlation with a clinical outcome may suggest a relatively higher functional relevance of these features for that outcome than the elevation of the RNA transcript alone. Most of the time, given the inaccessibility of target organs with these technologies, causality cannot be corroborated using blood-derived features. An exception to that rule is the potential use of circulating leukocyte genomic DNA to conduct a genome-wide association study (GWAS) that fulfills the following criteria: 1) identifies a gene abnormality coding

for an aberrant protein; 2) that protein executes its function in the circulation; and 3) that malfunction leads to an adverse pulmonary outcome. An illustrative example of that exceptional case is the factor V Leiden, which consists of a genetic polymorphism in the coagulation factor V that increases its clotting activity because of deficient binding to the anticoagulant protein C, which is both present in the circulation and is causally related to venous thromboembolism development (7). It is important to clarify that a *polymorphism* is a variation in the DNA nucleotide sequence that is present in all the cells of an organism including the germ cells, and thus can be identified by analysis of *any cell's DNA*, such as circulating leukocytes. By contrast, a *somatic mutation* associated with a disease process such as lung cancer is an abnormal nucleotide sequence only present in the transformed *cell's DNA*, and thus requires for its identification the sequencing of circulating *cell-free DNA* fragments produced by the abnormal tissue, or liquid biopsy, as mentioned before.

Although medicine has traditionally operated with a rule-based system in which the predictions are made from previously established knowledge (8), big datasets are frequently processed without knowing *a priori* how the constituent features will associate with the investigated outcome. In other words, because of the large number of features in big datasets, their interactions and effects on clinical outcomes are complex and unpredictable. Eventually, complex feature combinations can generate predictable

Term	Description
DNA	Deoxyribonucleic acid, cellular molecule that stores genetic information
RNA	Ribonucleic acid, molecular transcripts unstream of protein synthesis
gDNA	Genomic DNA, chromosomal DNA from cell nucleus
cDNA	Complementary DNA, synthesized from single-stranded RNA transcripts
Library	Collection of fragmented and cloned DNA, including cDNA (from RNA)
NGS	Next-generation sequencing, used for RNA or DNA sequencing
CpG	5'—C—phosphate—G—3', cytosine and guanine separated by phosphate
DMR	Differentially methylated region (comprised of several CpGs)
Targeted	Analysis of a predesignated set of target biomolecules
Untargeted	Discovery-based interrogation of biomolecules, unbiased
Omics	Large-scale investigation of biomolecules (RNA, DNA, Lipids, etc.)
Alignment	Method of comparing related biomolecule sequences
Features	Individual datapoints such as peptides, transcripts, or cytokines

**Figure 1.** Common terms used in the big data and omics literature.

Molecule	Omic technology	Platforms	Assay
 Metabolites	Metabolomics	Mass spectrometry <ul style="list-style-type: none"> <li>• Targeted panels</li> <li>• Untargeted/Discovery-based</li> </ul>	Profile and quantify relative abundance of metabolites
 Lipids	Lipidomics	Mass spectrometry <ul style="list-style-type: none"> <li>• Targeted panels</li> <li>• Untargeted/Discovery-based</li> </ul>	Quantify soluble and insoluble lipids
 Proteins	Proteomics	Mass spectrometry <ul style="list-style-type: none"> <li>• Targeted panels               <ul style="list-style-type: none"> <li>• Luminex<sup>®</sup> cytokine panel</li> </ul> </li> <li>• Untargeted/Discovery-based</li> </ul>	Characterize signature of cellular protein expression
 RNA	Transcriptomics	Ion Torrent <sup>™</sup> <ul style="list-style-type: none"> <li>• AmpliSeq<sup>™</sup> panel</li> </ul> Illumina <sup>®</sup> <ul style="list-style-type: none"> <li>• De novo sequencing</li> </ul>	Determine expression levels of cellular transcripts
 DNA sequencing	Genomics	Methylomics <ul style="list-style-type: none"> <li>• Infinium MethyEPIC</li> <li>• Whole-genome sequencing</li> </ul> DNA Sequencing GWAS	Measurement of various genomic features

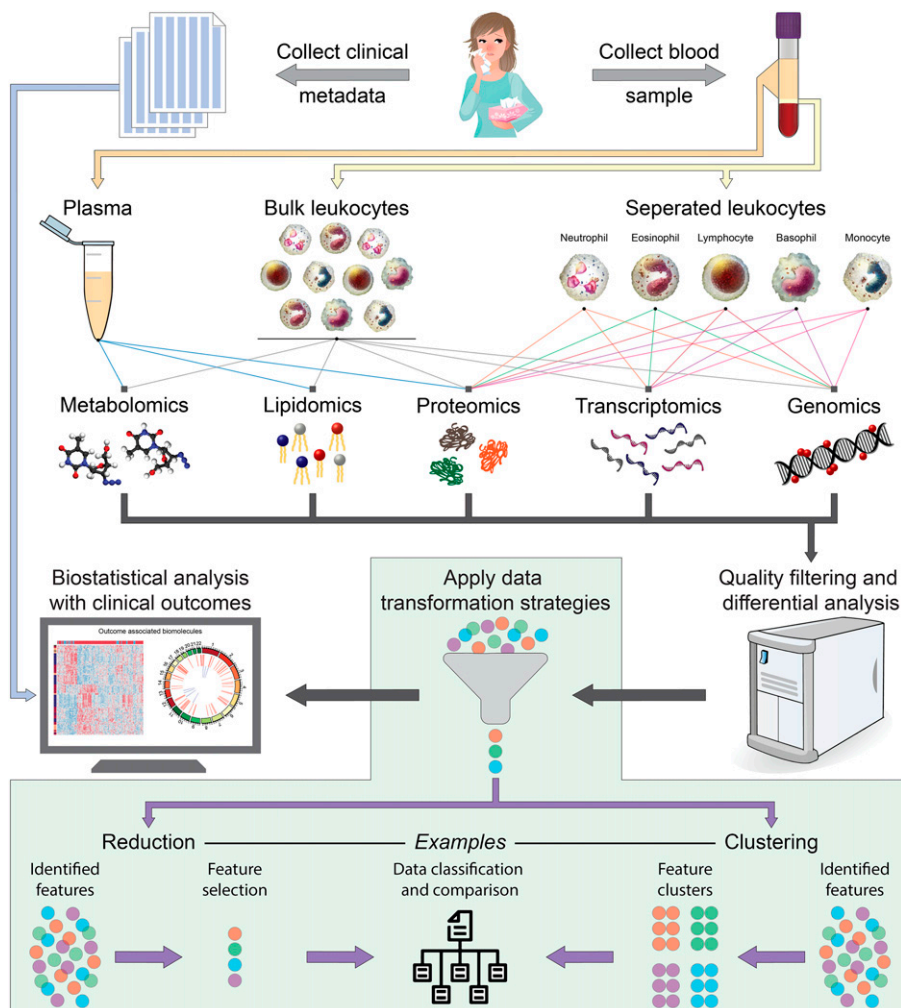
**Figure 2.** Description of the most popular platforms to generate datasets with molecular features from peripheral blood. GWAS = genome-wide association study.

signatures that are replicated by independent cohorts showing consistent correlations with an outcome of interest. However, even if a collection of features forms a signature that demonstrates a consistent association with an outcome, a constituent feature of that signature could be combined with other, different features and be associated with an opposite outcome, and thus the predictive effects of multiple feature combinations are impossible to anticipate with the use of preestablished rules (8). To maximize the power of this complexity, big datasets are typically, although not always, processed with machine learning technology, which uses software that learns how to make predictions from features input into the system and not from previously known rules (8). These individualized predictions often require a training process that the software conducts with a subset of the input data. Importantly, although in classic biostatistics it is accepted that more sample datapoints or features typically correlate with higher statistical power and prediction accuracy (9), this concept does not apply to machine learning–driven big data processing, because the predictive power of machine learning models is sharply reduced beyond a certain number of input features, which is known as the curse of dimensionality or the Hughes phenomenon (10). To deal

with this caveat, machine learning technology typically starts with a first step of data simplification, which can be done in multiple ways that broadly include the following: 1) clustering, in which features that share a relatively similar effect on the investigated outcome are lumped together; and 2) reduction, resulting from an initial filtering step, that defines which features are more impactful on the outcome of interest, followed by either feature elimination of the less relevant entities or feature selection of the most relevant entities. These filtering and reduction processes leave a final dataset that retains a subgroup of the original features that are more strongly correlated with the selected outcome (Figure 3). Examples of data clustering that we use later in the text are data hierarchical clustering (11) and latent class analysis (LCA) (12), and an example of data reduction is recursive feature elimination (RFE) (13). Often, data aggregation and reduction are combined in a single analytical strategy. For example, multiple features can be clustered, resulting in a reduced number of variables, and then, among those variables, the ones that provide the most impactful information on a selected outcome are retained, and the rest are eliminated. A typical use of that approach is to conduct principal component analysis on a big dataset and

retain only the components most impactful to the outcome of interest. These components can be later added as a covariable to adjust for an effect on said outcome using classical multivariable statistical analysis (14).

It is important to clarify that machine learning is not always combined with omics-level data. For example, recent studies that have defined subphenotypes in critical illness used LCA and machine learning technology on readily available laboratory and clinical records in combination with cytokine and chemokine analyses detected with multiplex targeted (nonomics) systems. In contrast, some omics data (e.g., DNA-sequencing data) are not always processed with machine learning technology. For example, GWASs conducted to associate genetic polymorphisms with lung phenotypes, such as between the *MUC5B* promoter polymorphism and idiopathic pulmonary fibrosis (IPF) (15), do not use machine learning technology to analyze the datasets. Because all the cells in a single organism have identical genetic information, these studies often interrogate DNA from an easily accessible cell type, such as circulating leukocyte DNA. Nevertheless, GWASs can still be combined with machine learning technology to refine analyses and suggest possible causal associations between polymorphisms and outcomes (16, 17).



**Figure 3.** Cartoon illustrating the pipeline of data generation involving collection of blood samples and clinical data from participant individuals, followed by blood constituent processing, omic processing, data simplification, and correlation with clinical outcomes.

## Applications of Big Datasets Obtained from the Peripheral Blood in Pulmonary and Critical Care Medicine

Blood-derived big datasets helped improve understanding and management of multiple conditions in pulmonary and critical care medicine. The following are typical uses of peripheral blood-derived data sources: 1) biomarker discovery, allowing the identification of one or more features that are useful to improve disease prediction and monitor treatment response, such as the association between IL-6 elevation and worse outcomes in acute respiratory distress syndrome (ARDS) (18); 2) identification of targetable mechanisms, which can potentially be “druggable” with outcome-modifying intent, such as the use of IL-6 pathway

antagonists to improve outcomes in severe COVID-19–induced ARDS (19); 3) identification of novel subphenotypes via clustering of signatures, which could refine treatments that do not work for the overall population but are effective on a unique phenotype, such as the beneficial use of corticosteroids on a subgroup of patients with sepsis (20); and 4) information about potential pathophysiological mechanisms, including the discovery of an adverse association of genetic polymorphisms, such as *MUC5B* with IPF (15, 21, 22).

## Specific Sources of Big Datasets Obtained from the Peripheral Blood

Peripheral blood contributes cellular (i.e., leukocytes) and soluble fractions. The

cellular fraction can be further used in bulk analysis, in which the sample is processed for the mean (average) value of a given feature. For example, in an RNA-sequencing analysis, the transcripts’ relative expression in each patient will be reported as the average value expressed by the whole, or bulk, cellular fraction from that patient used to conduct the analysis. Because different cell types express alternative genes and thus produce diverse proteins, a bulk RNA-sequencing dataset that aggregates transcripts produced by all the blood cells used for a given analysis may not capture the transcript coding for an outcome-impactful protein. For example, although IL-6 and IL-8 are associated with worse outcomes in ARDS (18, 23), bulk RNA-sequencing analysis does not identify their upregulation in sicker patients (5, 23, 24). Indeed, although circulating monocytes produce IL-6 (25),



these cells are relatively scarce in the peripheral blood count, and thus bulk RNA sequencing of unsorted cells will not capture these cytokine transcripts as differentially expressed between patients with better and worse outcomes. It is also possible that other noncirculating cells, such as tissue macrophages, produce a fraction of the cytokine RNA and protein products, in which case, even cell sorting of peripheral blood cells will fail to capture the RNA transcript coding for this outcome-relevant protein. Some other times, the bulk RNA-sequencing analysis can be very useful to inform on actionable measures, including the use of corticosteroids to treat patients demonstrating a specific transcriptomic bulk signature (20). An exception to the bulk-average rule is DNA-sequencing analysis, given that the data obtained in bulk analysis is identical in all the cells present in a single organism, which means that no data are obscured or underrepresented because of this aggregated type of sample collection.

The cellular fraction can also be used in cell sorting, where cellular components are further separated with the use of Fluorescence Activated Cell Sorting (FACS) that takes advantage of specific surface antigens expressed by certain cell types and fluorescent antibodies bound to those antigens. These separated cell fractions can be further processed with omic analysis (26). In addition, the cellular fraction can be used in single-cell sequencing, in which omic data are all generated from individual cells. As opposed to FACS-sorted cells, single-cell data require that cells be individually separated, even those expressing the same surface antigens, which means that each cell will contribute a unique dataset (26). Single-cell technology is relatively recent, and its applications in the field of peripheral blood omics are therefore more limited; yet, these approaches have already been used in multiple studies to define cell-specific mechanisms of disease heterogeneity, as we present later in the text. Samples from the cellular fraction contribute DNA, RNA, proteins, lipids, metabolites, and other molecules. Although the soluble fraction is not typically used for DNA or RNA analysis but instead contributes large amounts of proteins, lipids, and metabolites (5), circulating cfDNA has been used to stratify mortality risk and identify sources of tissue injury in patients with severe COVID-19 and pulmonary hypertension (27, 28). Moreover, circulating microRNA has recently been used to facilitate early detection of lung cancer (29) (Figure 3).

## Essential Concepts of Pipelines Used for Big Data Generation and Analysis

After the clinical data and blood samples are obtained from study participants, blood samples are processed to separate the fraction of interest. Then, for DNA or RNA, a collection of nucleotide fragments, or libraries, needs to be prepared for sequencing (Figure 1). The sequencing step produces a large amount of data, which is later bioinformatically processed, including data simplification (e.g., clustering and/or reduction), and correlated with selected clinical outcomes (Figure 3). Appropriate statistical models are employed to establish phenotype associations; specific model descriptions and uses can be found elsewhere (5).

## Examples of Big Datasets Generated with Peripheral Blood in the Recent Pulmonary and Critical Care Medicine Literature

1. *Circulating RNA-sequencing analysis:* Research has shown that circulating bulk RNA sequencing followed by hierarchical clustering analysis can be used to define two distinct leukocyte transcriptomic groups in sepsis, sepsis response signature 1 (SRS1; immunosuppressed) and SRS2 (nonimmunosuppressed), which are associated with higher and lower mortality, respectively (30). The same group recently reported that SRS1 and SRS2 are partially replicated in patients with COVID-19 (11). These two phenotypes could respond differently to corticosteroid administration (20), potentially refining selection of patients to receive these drugs and improving clinical outcomes. Another example of this approach is the use of 52 transcripts in circulating leukocytes to predict mortality in IPF (31). Interestingly, expression levels of 50 of these 52 genes were found in one study to define risk of death in hospitalized patients with COVID-19, suggesting that severe IPF and COVID-19 may share a similar inflammatory profile in the peripheral

blood. These data could potentially inform drug development if these signatures are corroborated in other studies (32). Single-cell sequencing analysis obtained from a different cohort (33) indicates that among patients with higher mortality risk, transcripts are contributed by monocytes, dendritic cells, and neutrophils, whereas low-risk profile-expressing cells are predominantly lymphocytes (32).

2. *DNA-sequencing analysis:* GWASs identified an association of *MUC5B* promoter variants with IPF (15). Recently, that polymorphism was found also to be associated with the development of ARDS, suggesting that ARDS and undiagnosed interstitial lung disease or pulmonary fibrosis may share a potential pathogenic overlap in some cases (34). Moreover, associations of higher COVID-19 mortality with the ABO blood group (35) and human leukocyte antigen systems (36) have recently been described. Although genetic association studies do not typically require machine learning-mediated processing, recent research has used these algorithms to further analyze genomic datasets, allowing more refined identification of genetic variants, known as supervariants, that may be causally associated with COVID-19 mortality (37). It is important to emphasize that, except for genes that code for protein products that execute their function in the circulation and are associated with adverse pulmonary outcomes, big data analyses involving features present in the peripheral blood are largely inferential, and thus mechanistic causality cannot be substantiated using these methodologies.
3. *Plasma cytokines, chemokines, and other proteins:* Protein biomarkers in the circulation, including IL-6, IL-8, soluble TNF receptor 1, and protein C, can discriminate hyper- versus hypoinflammatory phenotypes in ARDS and predict patient response to positive end-expiratory pressure and other measures (38). Machine learning models can distinguish these phenotypes with regular clinical data, which could facilitate the rapid bedside identification of phenotypes leading to more refined fluid challenge, positive

end-expiratory pressure setting, and other measures (39). A recent study using LCA suggested a partially overlapping pattern in critically ill patients with COVID-19 and a differential response to corticosteroid administration between these patient classes (12).

4. **Plasma metabolomics:** Using mass spectrometry to identify circulating blood metabolites followed by feature selection and machine learning prediction, a recent study reported that plasma metabolites such as glycylproline and long-chain acylcarnitines could be associated with antibody fading in convalescent patients with COVID-19 and with higher susceptibility to reinfection. This finding could be instrumental to personalize the vaccination “booster” timing or to develop strategies to maintain levels of neutralizing antibody levels (40). Another study combining circulating macrophages with single-cell sequencing and metabolomics has identified multiple metabolism-related genes and substrates correlated with worse patient outcomes (4). A very recent analysis of patients with ARDS and sepsis found that metabolic LCA clustering defined classes that are independently associated with mortality and could inform personalized approaches as well (41).
5. **DNA methylation analysis:** DNA methylation is an epigenetic change that regulates gene accessibility and thus gene expression, thereby influencing the cellular phenotype (42). Blood DNA methylation can be determined with untargeted DNA sequencing, with targeted DNA microarrays, or with direct target sequencing of individual gene areas (42). To determine the DNA methylation status, a first step of either chemical or enzymatic treatment of the DNA is needed (42). The circulating DNA methylome has been used to predict long-term outcomes in acute critical illness, including the effects of early parenteral nutrition on neurocognitive development in pediatric critical illness (43). In chronic obstructive pulmonary disease (COPD), circulating leukocyte DNA methylation has recently been used to characterize potentially targetable loci in the genome for future

drug development (14, 44). DNA methylation sequencing has identified multiple positions and regions across the genome that are differentially methylated in association with COVID-19 diagnosis and severity (13). Recursive feature elimination identified a limited set of positions that, if methylated, predict the mortality of patients with COVID-19 (13). Because DNA methylation is a covalent and relatively stable chemical modification to the DNA, some of these regions could remain aberrantly dysregulated long after COVID-19 and provide a biological underpinning for postacute sequelae of SARS-CoV-2 infection (45, 46).

6. **cfDNA and cfRNA:** cfDNAs are circulating short DNA fragments (~165 base pairs) that represent cell injury or cell turnover. Elevated total cfDNA concentration has been associated with worse prognosis in heterogeneous conditions such as sepsis and trauma (47). DNA methylation sequencing facilitates the discrimination of cfDNA subsets based on cell type origin, allowing the detection of tissue-specific injury. Indeed, recent studies using methylated cfDNA sequencing followed by unsupervised clustering have identified the sources of tissue injury associated with worse outcomes in pulmonary hypertension and COVID-19 (27, 28). These data could potentially inform measures of organ support to improve outcomes in these conditions. Moreover, multiple cfRNA signatures have been used to identify patients with lung cancer and differentiate them from those with no cancer (29), offering promise to better surveil patients with elevated risk of lung cancer and to monitor relapse of those already treated with curative intent.

### Limitations of This Approach

The use of peripheral blood omics and multiplex systems is associated with significant caveats that need to be carefully considered. First, peripheral blood features associated with a particular disease characterize not the primary organ driving the disease course but instead its peripheral blood molecular profile. For example, as

mentioned before, patients admitted to the ICU because of pneumonia and sepsis can be classified as immunosuppressed (SRS1, with relatively higher 14-d mortality) or nonimmunosuppressed (SRS2, with relatively lower 14-d mortality) (30). That classification is based on bulk transcriptomic analysis of circulating leukocytes and does not mean that the immune response taking place in the lungs is mirrored by the peripheral blood signature. As mentioned before, except for genomic DNA findings of genetic variants coding for aberrant proteins that execute their function in the circulation and impact lung phenotypes, the information provided by the peripheral blood should never be assumed as mechanistically representing the status of an inaccessible organ. Second, peripheral blood features are indirect surrogates of organ abnormalities, and, given that pulmonary and critical care conditions are often associated with multiple comorbidities, a set of features assumed to covary with a primary organ dysfunction could instead do so with a secondary organ. For example, features originally described as surrogates of COPD (48) have recently been shown to correlate with COPD-induced skeletal muscle and not pulmonary integrity (49). Third, a specific signature found in a patient’s population could not necessarily be replicated in another, unrelated group of individuals for multiple reasons, including racial and ancestral background, geographical factors, and others (50). For those reasons, the findings from a study conducted on a single cohort of patients should ideally be corroborated in a separate group of individuals serving as a validation cohort. Fourth, the sequencing technology used to identify features is costly and not easily implemented at a large scale in clinical practice, at least for now. The cost is variable, with bulk analysis being cheaper than single-cell sequencing; yet, the limitations associated with each technique mentioned before currently limit the universal use of these technologies. Similarly, even for readily available data, the use of an outcome-associated cluster might require consistent aggregation of many features, and missing data can undermine the models’ predictive power. The recent expansion and accessibility of artificial intelligence tools could help practicing physicians organize large-scale data collection and use, and

develop actions to facilitate patient care in the near future (51).

## Future Directions

The following are major areas that could lead to an acceleration of peripheral blood omics- and other multiplex-based technologies' development in the setting of pulmonary and critical care medicine:

1. Improving scalability, such as by the development of cheaper and efficient pipelines that can procure, at the bedside, high-quality samples for further omic processing in a reproducible way;
2. Generation of outcome-associated signatures that require fewer features to

establish accurate correlations with outcomes of interest and thus can be probed with less sophisticated technologies;

3. Wider availability of data-processing systems, including deep learning instruments, that are readily available to the general public;
4. Automatization of data curation, which could improve imputation of accurate subject data, including clinical covariables collected and aggregated in real time, severities, prehospitalization comorbidity scores, and other aspects that are relevant to capture population heterogeneity potentially undermining reproducibility; and
5. Development of collaborative platforms using cloud computing systems that can merge and analyze data from multiple centers in a way that protects patient

confidentiality and at the same time accelerates timely validation.

## Conclusions

The combination of peripheral blood big data generation with clinical variables and outcomes has contributed to a better understanding of pulmonary and critical care conditions, together with potentially actionable therapies. The recent acceleration of machine learning and other artificial intelligence technologies will likely facilitate larger-scale data generation and processing in the near future. ■

**Author disclosures** are available with the text of this article at [www.atsjournals.org](http://www.atsjournals.org).

## References

1. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015;372:793–795.
2. Heder M. From NASA to EU: the evolution of the TRL scale in public sector innovation. *Innov J* 2017;22:3.
3. Paik PK, Felip E, Veillon R, Sakai H, Cortot AB, Garassino MC, et al. Tepotinib in non-small-cell lung cancer with MET exon 14 skipping mutations. *N Engl J Med* 2020;383:931–943.
4. Ambikan AT, Yang H, Krishnan S, Svensson Akusjärvi S, Gupta S, Lourda M, et al. Multi-omics personalized network analyses highlight progressive disruption of central metabolism associated with COVID-19 severity. *Cell Syst* 2022;13:665–681.e4.
5. Overmyer KA, Shishkova E, Miller IJ, Balnis J, Bernstein MN, Peters-Clarke TM, et al. Large-scale multi-omic analysis of COVID-19 severity. *Cell Syst* 2021;12:23–40.e7.
6. Chen J, Tang J, Nie M, Li Y, Wurfel MM, Meyer NJ, et al. WNT9A affects late-onset ARDS and 28-day survival: evidence from a three-step multi-omics study. *Am J Respir Cell Mol Biol* [online ahead of print] 24 Apr 2023; DOI: 10.1165/rcmb.2022-0416OC.
7. Bertina RM, Kooleman BP, Koster T, Rosendaal FR, Dirven RJ, de Ronde H, et al. Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* 1994;369:64–67.
8. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019;380:1347–1358.
9. Lin M, Lucas HC Jr, Shmueli G. Too big to fail: large samples and the *p*-value problem. *Inf Syst Res* 2013;24:906–917.
10. Hughes G. On the mean accuracy of statistical pattern recognizers. *IEEE Trans Inf Theory* 1968;14:55–63.
11. Cano-Gamez E, Burnham KL, Goh C, Allcock A, Malick ZH, Overend L, et al.; GAInS Investigators. An immune dysfunction score for stratification of patients with acute infection based on whole-blood gene expression. *Sci Transl Med* 2022;14:eabq4433.
12. Sinha P, Furfaro D, Cummings MJ, Abrams D, Delucchi K, Maddali MV, et al. Latent class analysis reveals COVID-19-related acute respiratory distress syndrome subgroups with differential responses to corticosteroids. *Am J Respir Crit Care Med* 2021;204:1274–1285.
13. Balnis J, Madrid A, Hogan KJ, Drake LA, Chieng HC, Tiwari A, et al. Blood DNA methylation and COVID-19 outcomes. *Clin Epigenetics* 2021;13:118.
14. Lee M, Huan T, McCartney DL, Chittoor G, de Vries M, Lahousse L, et al. Pulmonary function and blood DNA methylation: a multi-ancestry epigenome-wide association meta-analysis. *Am J Respir Crit Care Med* 2022;206:321–336.
15. Seibold MA, Wise AL, Speer MC, Steele MP, Brown KK, Loyd JE, et al. A common MUC5B promoter polymorphism and pulmonary fibrosis. *N Engl J Med* 2011;364:1503–1512.
16. Szymczak S, Biernacka JM, Cordell HJ, González-Recio O, König IR, Zhang H, et al. Machine learning in genome-wide association studies. *Genet Epidemiol* 2009;33:S51–S57.
17. van Hilten A, Kushner SA, Kayser M, Ikram MA, Adams HHH, Klaver CCW, et al. GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. *Commun Biol* 2021;4:1094.
18. Sinha P, Delucchi KL, Thompson BT, McAuley DF, Matthay MA, Calfee CS; NHLBI ARDS Network. Latent class analysis of ARDS subphenotypes: a secondary analysis of the statins for acutely injured lungs from sepsis (SAILS) study. *Intensive Care Med* 2018;44:1859–1869.
19. REMAP-CAP Investigators. Interleukin-6 receptor antagonists in critically ill patients with Covid-19. *N Engl J Med* 2021;384:1491–1502.
20. Antcliffe DB, Burnham KL, Al-Beidh F, Santhakumaran S, Brett SJ, Hinds CJ, et al. Transcriptomic signatures in sepsis and a differential response to steroids. From the VANISH randomized trial. *Am J Respir Crit Care Med* 2019;199:980–986.
21. Hancock LA, Hennessy CE, Solomon GM, Dobrinskikh E, Estrella A, Hara N, et al. Muc5b overexpression causes mucociliary dysfunction and enhances lung fibrosis in mice. *Nat Commun* 2018;9:5363.
22. Hunninghake GM, Hatabu H, Okajima Y, Gao W, Dupuis J, Latourelle JC, et al. MUC5B promoter polymorphism and interstitial lung abnormalities. *N Engl J Med* 2013;368:2192–2200.
23. Balnis J, Adam AP, Chopra A, Chieng HC, Drake LA, Martino N, et al. Unique inflammatory profile is associated with higher SARS-CoV-2 acute respiratory distress syndrome (ARDS) mortality. *Am J Physiol Regul Integr Comp Physiol* 2021;320:R250–R257.
24. Bos LDJ, Scicluna BP, Ong DSY, Cremer O, van der Poll T, Schultz MJ. Understanding heterogeneity in biologic phenotypes of acute respiratory distress syndrome by leukocyte expression profiles. *Am J Respir Crit Care Med* 2019;200:42–50.
25. Tosato G, Seamon KB, Goldman ND, Sehgal PB, May LT, Washington GC, et al. Monocyte-derived human B-cell growth factor identified as interferon-beta 2 (BSF-2, IL-6). *Science* 1988;239:502–504.
26. Grant RA, Morales-Nebreda L, Markov NS, Swaminathan S, Querrey M, Guzman ER, et al.; NU SCRIPT Study Investigators. Circuits between infected macrophages and T cells in SARS-CoV-2 pneumonia. *Nature* 2021;590:635–641.
27. Andargie TE, Tsuji N, Seifuddin F, Jang MK, Yuen PS, Kong H, et al. Cell-free DNA maps COVID-19 tissue injury and risk of death and can cause tissue injury. *JCI Insight* 2021;6:e147610.

28. Brusca SB, Elinoff JM, Zou Y, Jang MK, Kong H, Demirkale CY, *et al*. Plasma cell-free DNA predicts survival and maps specific sources of injury in pulmonary arterial hypertension. *Circulation* 2022;146:1033–1045.
29. Fehlmann T, Kahraman M, Ludwig N, Backes C, Galata V, Keller V, *et al*. Evaluating the use of circulating microRNA profiles for lung cancer detection in symptomatic patients. *JAMA Oncol* 2020;6:714–723.
30. Davenport EE, Burnham KL, Radhakrishnan J, Humburg P, Hutton P, Mills TC, *et al*. Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study. *Lancet Respir Med* 2016;4:259–271.
31. Herazo-Maya JD, Sun J, Molyneaux PL, Li Q, Villalba JA, Tzouvelelis A, *et al*. Validation of a 52-gene risk profile for outcome prediction in patients with idiopathic pulmonary fibrosis: an international, multicentre, cohort study. *Lancet Respir Med* 2017;5:857–868.
32. Juan Guardela BM, Sun J, Zhang T, Xu B, Balnis J, Huang Y, *et al*. 50-gene risk profiles in peripheral blood predict COVID-19 outcomes: a retrospective, multicenter cohort study. *EBioMedicine* 2021;69:103439.
33. Lee JS, Park S, Jeong HW, Ahn JY, Choi SJ, Lee H, *et al*. Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. *Sci Immunol* 2020;5:eabd1554.
34. Rogers AJ, Solus JF, Hunninghake GM, Baron RM, Meyer NJ, Janz DR, *et al*. MUC5B promoter polymorphism and development of acute respiratory distress syndrome. *Am J Respir Crit Care Med* 2018;198:1342–1345.
35. Ellinghaus D, Degenhardt F, Bujanda L, Buti M, Albillos A, Invernizzi P, *et al*.; Severe Covid-19 GWAS Group. Genomewide association study of severe Covid-19 with respiratory failure. *N Engl J Med* 2020;383:1522–1534.
36. Weiner J, Suwalski P, Holtgrewe M, Rakitko A, Thibeault C, Müller M, *et al*. Increased risk of severe clinical course of COVID-19 in carriers of HLA-C\*04:01. *EClinicalMedicine* 2021;40:101099.
37. Liu Z, Dai W, Wang S, Yao Y, Zhang H. Deep learning identified genetic variants for COVID-19-related mortality among 28,097 affected cases in UK Biobank. *Genet Epidemiol* 2023;47:215–230.
38. Calfee CS, Delucchi K, Parsons PE, Thompson BT, Ware LB, Matthay MA; NHLBI ARDS Network. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. *Lancet Respir Med* 2014;2:611–620.
39. Sinha P, Churpek MM, Calfee CS. Machine learning classifier models can identify acute respiratory distress syndrome phenotypes using readily available clinical data. *Am J Respir Crit Care Med* 2020;202:996–1004.
40. Yang Z, Wu D, Lu S, Qiu Y, Hua Z, Tan F, *et al*. Plasma metabolome and cytokine profile reveal glycolipid modulating antibody fading in convalescent COVID-19 patients. *Proc Natl Acad Sci USA* 2022;119:e2117089119.
41. Alipanah-Lechner N, Neyton L, Mick E, Willmore A, Leligdowicz A, Contrepolis K, *et al*. Plasma metabolic profiling implicates dysregulated lipid metabolism and glycolytic shift in hyperinflammatory ARDS. *Am J Physiol Lung Cell Mol Physiol* 2023;324:L297–L306.
42. Singer BD. A practical guide to the measurement and analysis of DNA methylation. *Am J Respir Cell Mol Biol* 2019;61:417–428.
43. Güiza F, Vanhorebeek I, Verstraete S, Verlinden I, Derese I, Ingels C, *et al*. Effect of early parenteral nutrition during paediatric critical illness on DNA methylation as a potential mediator of impaired neurocognitive development: a pre-planned secondary analysis of the PEPaNIC international randomised controlled trial. *Lancet Respir Med* 2020;8:288–303.
44. Morrow JD, Make B, Regan E, Han M, Hersh CP, Tal-Singer R, *et al*. DNA methylation is predictive of mortality in current and former smokers. *Am J Respir Crit Care Med* 2020;201:1099–1109.
45. Balnis J, Madrid A, Hogan KJ, Drake LA, Adhikari A, Vancavage R, *et al*. Persistent blood DNA methylation changes one year after SARS-CoV-2 infection. *Clin Epigenetics* 2022;14:94.
46. Balnis J, Madrid A, Hogan KJ, Drake LA, Adhikari A, Vancavage R, *et al*. Whole-genome methylation sequencing reveals that COVID-19-induced epigenetic dysregulation remains 1 year after hospital discharge. *Am J Respir Cell Mol Biol* 2023;68:594–597.
47. Gögenur M, Burcharth J, Gögenur I. The role of total cell-free DNA in predicting outcomes among trauma patients in the intensive care unit: a systematic review. *Crit Care* 2017;21:14.
48. Verrills NM, Irwin JA, He XY, Wood LG, Powell H, Simpson JL, *et al*. Identification of novel diagnostic biomarkers for asthma and chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2011;183:1633–1643.
49. Balnis J, Vincent CE, Jones AJ, Drake LA, Coon JJ, Lee CG, *et al*. Established biomarkers of chronic obstructive pulmonary disease reflect skeletal muscle integrity's response to exercise in an animal model of pulmonary emphysema. *Am J Respir Cell Mol Biol* 2020;63:266–269.
50. Khan AT, Gogarten SM, McHugh CP, Stilp AM, Sofer T, Bowers ML, *et al*. Recommendations on the use and reporting of race, ethnicity, and ancestry in genetic research: experiences from the NHLBI TOPMed program. *Cell Genom* 2022;2:100155.
51. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. *N Engl J Med* 2023;388:1201–1208.