Review Article

A bioinformatics approach towards bronchopulmonary dysplasia

Charles Taylor Valadie[1#], Shreyas Arya[2#], Tanima Arora[1], Nisha Reddy Pandillapalli[1], Alvaro Moreira[1]

[1]Department of Pediatrics, University of Texas Health San Antonio, San Antonio, TX, USA; [2]Department of Pediatrics, Dayton Children's Hospital, Dayton, OH, USA

Contributions: (I) Conception and design: CT Valadie, S Arya, T Arora, A Moreira; (II) Administrative support: None; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: T Arora, A Moreira; (V) Data analysis and interpretation: All authors; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Alvaro Moreira, MD, MSc. Department of Pediatrics, University of Texas Health San Antonio, 7703 Floyd Curl Drive, San Antonio, TX 78229, USA. Email: MoreiraA@uthscsa.edu.

Background and Objective: Bronchopulmonary dysplasia (BPD) is the most common morbidity associated with prematurity and remains a significant clinical challenge. Bioinformatic approaches, such as genomics, transcriptomics, and proteomics, have emerged as novel methods for studying the underlying mechanisms driving BPD pathogenesis. These methods can be used alongside clinical data to develop a better understanding of BPD and potentially identify the most at risk neonates within the first few weeks of neonatal life. The objective of this review is to provide an overview of the current state-of-the-art in bioinformatics for BPD research.

Methods: We conducted a literature review of bioinformatics approaches for BPD using PubMed. The following keywords were used: "biomedical informatics", "bioinformatics", "bronchopulmonary dysplasia", and "omics".

Key Content and Findings: This review highlighted the importance of omic-approaches to better understand BPD and potential avenues for future research. We described the use of machine learning (ML) and the need for systems biology methods for integrating large-scale data from multiple tissues. We summarized a handful of studies that utilized bioinformatics for BPD in order to better provide a view of where things currently stand, identify areas of ongoing research, and concluded with challenges that remain in the field.

Conclusions: Bioinformatics has the potential to enable a more comprehensive understanding of BPD pathogenesis, facilitating a personalized and precise approach to neonatal care. As we continue to push the boundaries of biomedical research, biomedical informatics (BMI) will undoubtedly play a key role in unraveling new frontiers in disease understanding, prevention, and treatment.

Keywords: Biomedical informatics (BMI); bioinformatics; bronchopulmonary dysplasia (BPD); computational biology; genomics

Submitted Mar 02, 2023. Accepted for publication May 31, 2023. Published online Jun 19, 2023.
doi: 10.21037/tp-23-133
View this article at: https://dx.doi.org/10.21037/tp-23-133

Introduction

What is biomedical informatics (BMI)?

Informatics seeks to take large amounts of data and determine how to best give it meaning (1). BMI is a field that combines computer science, mathematics, statistics, epidemiology, and engineering to solve problems in biology and medicine. It involves the application of computational methods to manage, analyze, and interpret biological and biomedical data, such as genomic data, clinical data, or imaging data. *Figure 1* illustrates a few of the subfields that encompass BMI (2).

BMI has become increasingly popular over the past few decades and has already made substantial contributions

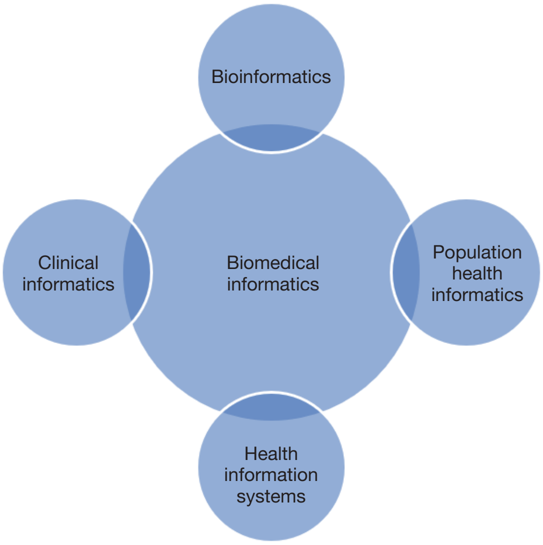

Figure 1 Overview of disciplines in biomedical informatics.

to healthcare. Recent advances in BMI have contributed significantly to various healthcare processes. For example, the development of clinical decision support tools within electronic health records (EHRs) has improved patient safety. These techniques have also enabled more accurate analyses of hospital costs and identification of inequities in healthcare (3). BMI has also become prominent in clinical research and the understanding of diseases at the molecular level. This reach notably includes a number of "omics" fields that allow a more comprehensive analysis of biologic systems such as genomics (studying DNA), transcriptomics (transcripts), proteomics (proteins) and metabolomics (metabolites). These also can undergo further specialization into more specialized fields such as epigenomics (studying epigenetic modifications of DNA) and metagenomics (combining study of the human genome with other organisms such as bacteria, viruses, etc.), among others (4). New exciting areas of medicine are evidenced by the Human Genome Project, predicting protein folding, and in a better understanding of complex genetic diseases such as cancer and Alzheimer's. These individual "omic" disciplines are often also combined for further, more in depth analysis, under the umbrella terms "multiomics" or "panomics".

In this review, we will examine the novel role of bioinformatics in understanding the pathophysiology of bronchopulmonary dysplasia (BPD) in neonates. By analyzing large-scale molecular and clinical data, BMI has the potential to uncover underlying mechanisms contributing to this chronic lung disease. Furthermore, these findings may facilitate the discovery of targeted therapies that could prevent or mitigate BPD, thereby improving outcomes in a vulnerable population. We present this article in accordance with the Narrative Review reporting checklist (available at https://tp.amegroups.com/article/view/10.21037/tp-23-133/rc).

Bioinformatics: subfield of BMI

As discussed above, bioinformatics is one of many fields under the umbrella of BMI. As the name implies it is more specifically involved in the application of biological data. Bioinformatics is essential for management of data in modern biology and medicine (5). In addition to the analysis of genomic data, bioinformatics is now being used for more complex topics such as analyzing molecular pathways in order to understand gene-disease interactions, querying, biological databases, and drug design (6). Bioinformatics and gene sequencing have allowed for collection and organization of large amounts of molecular data that would be nearly impossible to use in a meaningful way otherwise (7). These collections have been integral for the study and comparison of genetic information, and furthermore are frequently collected into databases and online-tools for public use (8).

Bioinformatics continues to feature prominently across medical research. Given the heterogeneity and complexity of the field, it is unsurprising that it has been used extensively in the study of cancer. It has been used in gene sequencing to identify pathogenic variants and predisposing genes and in risk stratification (9). Once disease has been identified, it has also been used in drug development and the development of targeted chemotherapy. Even after treatment is ongoing, it has a place in monitoring for drug response and patterns of resistance (10). Whether in cancer or any of a vast number of diseases in which it is currently being used, the end is a more personalized or "precision medicine"-based approach to tailor diagnosis and treatment down to the individual patient level (11).

Although bioinformatics is a relatively new approach in neonatology, its utilization is rapidly growing. In the pediatric literature it has been used in several diseases such as asthma, attention deficit hyperactivity disorder (ADHD), inflammatory bowel disease, and obesity (12). As we will demonstrate later in this review, bioinformatics has proven to be a valuable tool in neonatal research, particularly in BPD.

Table 1 Search strategy

Items	Specification
Date of search	October 19, 2022
Databases and other sources searched	PubMed
Search terms used	"Biomedical informatics", "bioinformatics", "bronchopulmonary dysplasia", "computational biology", and "omics"
Timeframe	Inception till 10/19/2022
Inclusion and exclusion criteria	English articles
Selection process	T Arora, A Moreira

Table 2 Development of BPD definitions

Definition by	Year	Definition
National Institute of Child Health & Human Development & National Heart, Lung, Blood Institute	2001	Use of supplemental oxygen for >28 days (16)
Abman *et al.* (BPD collaborative)	2016	BPD if 28 days or greater of oxygen. Mild, moderate, severe type 1 or severe type 2 based on FiO_2 and mode of respiratory support (17)
Higgins *et al.*	2018	Need of oxygen supplementation in addition to respiratory support and early death due to parenchymal lung disease (18)
Jensen *et al.*	2019	Need of respiratory support at 36 weeks of PMA, irrespective of previous oxygen requirements (19)

BPD, bronchopulmonary dysplasia; FiO_2, inspiratory fraction of oxygen; PMA, postmenstrual age.

Methods

Literature search strategy

Articles were selected through PubMed and searched on October 19, 2022. Only articles in English were considered. The following keywords were used: "biomedical informatics", "bioinformatics", "bronchopulmonary dysplasia", "computational biology", and "omics". Retrieved studies were evaluated based on their titles and abstracts, and those considered relevant for the review. Two investigators conducted the research and included human and animal studies pertinent to the topic. *Table 1* summarizes our search strategy.

Bioinformatics for BPD

Despite significant advances in the management of preterm infants, BPD remains the most common morbidity associated with preterm birth, even half a century after it was first described (13,14). While part of the rising BPD rates can be attributed to the survival of more extremely preterm neonates, there are still significant gaps in our understanding of this condition and how we define it (15). The definition of BPD has undergone numerous changes over the past 20 years and now includes the degree of positive pressure ventilation at 36 weeks postmenstrual age, rather than supplemental oxygen alone, as a measure of its severity (*Table 2*) (16-19). However, the challenge with the current and all previous definitions of BPD is that they are clinical descriptors of an infant's lung function at a time they are expected to achieve lung maturity and are largely poor predictors of long-term respiratory outcomes (20). Additionally, there is significant phenotypic heterogeneity in the presentation of BPD and the complex interplay between lung injury and lung plasticity, making BPD a continuously evolving condition (21). With the advent of bioinformatic analysis and the numerous "omic" fields, backed by modern-day computational abilities, it is now possible to identify some of the pathobiological pathways that are likely involved in the development of BPD and identify some of the genetic risk factors that might predispose to it.

Applications of bioinformatics in BPD

The pursuit of the genetic basis of BPD has evolved from hypothesis-driven analysis of a few candidate genes to the use of high-throughput bioinformatic tools (22-24). A few studies have used the approach of genome wide association to elucidate the pathways associated with BPD.

Hadchouel *et al.* in 2011 used fine mapping techniques to identify *SPOCK2* as a susceptibility gene for the development of BPD. They also showed a significant increase in the mRNA level of *SPOCK2* in a rat model, during the alveolar stage of lung development (25). However, a genome-wide association study (GWAS) conducted by Wang *et al.* in 2013 did not identify a genomic loci or pathway that predicted the risk of BPD in very low birth weight (VLBW) infants (26). Another study in 2015 by Ambalavanan *et al.* performed a genome-wide scan to identify single nucleotide polymorphisms (SNPs) and pathways associated with BPD, finding an upregulation in the CD44 and miR-219 genes in the lungs of BPD patients and in association with hyperoxia. They also found differences in the pathways associated with mild/moderate and severe BPD and found racial/ethnic differences in these pathways (27). In 2017, Mahlman conducted a GWAS on preterm infants (24–30 weeks gestation) and identified SNPs near the C-reactive protein (*CRP*) gene as risk factors for BPD, independent of antenatal risks (28). Furthermore, Torgerson *et al.* in 2018 performed ancestry studies and GWAS to identify variants, genes, and pathways associated with survival in BPD infants treated with inhaled nitric oxide, identifying genetic variants related to lung development, drug metabolism, and immune response that contributed to individual and racial/ethnic differences in the respiratory outcomes of this high-risk population (29).

While GWAS have been able to identify risk loci in association with many human diseases, their success with BPD has been mixed. The reason may be because the GWAS methodology targets common variants in the population but since BPD and other respiratory morbidities in preterm infants are associated with a high mortality rate, the potential genetic risk alleles are expected to have a very low frequency (30). However, whole-exome sequencing allows for opportunities to study these rare variants. In 2015, Li *et al.* performed exome sequencing on 50 BPD twin pairs using DNA from neonatal blood spots and identified 258 genes associated with rare nonsynonymous mutations (31). Another study by Carrera *et al.* in the same year used whole exome sequencing in 26 unrelated infants with severe BPD to identify potential candidate genes implicated in the development of BPD (32).

Most recently, Wang *et al.* [2022] conducted an epigenome wide association study of BPD in preterm infants using cord blood DNA and DNA methylation techniques, revealing insights into the biological pathways involved in BPD pathogenesis (33). Most of these studies have utilized data from publicly available biorepositories or collected their own data and contributed to these repositories.

We have included the above studies to give a summary of the work that has been done by multiple groups in an effort to uncover the complexity of this disease. Taken as a whole, they highlight the excitement for the field but also many of the difficulties. The identification of new genes and pathways gives the opportunity for better understanding and insight as well as the opportunity to ultimately alter the way we attempt to prevent and treat BPD. At the same time, similar to research from other fields, loci and pathways found in animal models do not always translate into similar findings in human models. Human models also do not always identify similar genes/pathways as being the most significant and depend heavily on the patient population represented by an individual database, the type of tissue being collected, the methods used to obtain gene expression or pathway data and the models used to analyze them. Multiple of these studies involve systems that are almost certainly involved in BPD including epigenomic changes, inflammation (more specifically macrophage function and cytokine response) and airway septation. However most individual studies find their "most significant" changes at different loci from the ones before it. As research and databases grow, they will hopefully be accompanied by refinement and consistency in their analysis and more overlap of patient populations. BMI gives us the tools to explore disease in new ways, but also bring about new challenges.

Bioinformatics and data sharing

As the amount of data, we are able to collect continues to grow at an exponential rate, it has become equally important that we are able to maintain a system for keeping the data useable. Fortunately, the recent advances in data collection have been complemented by the development of sophisticated systems for the storage of biological data (5). This has had a number of benefits including international collaborations and database (or biorepository) creation, improvements in data quality, and helping to ensure traceability and transparency (34). It would be impossible to create a comprehensive list of these tools and data collections, but a few terms and examples are helpful in illustrating their use (*Table 3*).

Table 3 Example databases in bioinformatics

Database collections	Functions
GWAS catalog, https://www.ebi.ac.uk/gwas/	Human genome to find genetic changes associated with phenotypes of interest. The sharing of GWAS allows researched access to individual level data including pathway analysis/correlation, risk prediction and heritability estimation (26)
The KEGG Japanese database, https://www.genome.jp/kegg/	Functions by taking genome sequence data and associating it with the corresponding cellular/biologic pathway (35,36)
GO, http://geneontology.org/	Used both in human and animal research to link specific genes to their biological function (37)
Reactome, https://reactome.org/	Open-source, peer-reviewed pathway database whose goal is to provide bioinformatics tools for the visualization, pathway interpretation and analysis, genome analysis, modeling, systems biology, and education (38)

Other databases include SNPedia, Promethease, and MAGMA (39-42). This information is often then used as the basis for new projects and questions. GWAS, genome-wide association study; KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene Ontology; MAGMA, Multi-Marker Analysis of GenoMic Annotation.

These databases are often combined to help researchers make the leap from identifying genetic changes to discovering what biological pathways the genes in question belong to. Once identified, these pathways provide context to the results and can be used to help explain underlying physiology.

For applications in neonatology, a biorepository that houses data on BPD is the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO). Developed in the early 2000s, it is a publicly available database that contains a vast array of bioinformatic data (43). Similar to the previously discussed databases, it also has built-in tools to help download and interpret results.

The above examples are only a few of the many tools that exist to help obtain, analyze, and interpret biologic data. These systems are continually becoming more sophisticated, as are the projects that utilize them. Collaborations such as these have significant potential to rapidly expand our knowledge of disease much more quickly and clearly than individual studies could accomplish. As we continue to push the boundaries of biomedical research, BMI will undoubtedly play a key role in unraveling new frontiers in disease understanding, prevention, and treatment.

Future directions for use of bioinformatics in BPD

Recently RNA sequencing techniques have been developed that can evaluate messenger RNA (mRNA), long non-coding RNA (lncRNA) and microRNA (miRNA, miR) (44). In 2016, Kho *et al.* analyzed the fetal whole-lung transcriptomic profile at 54–127 days post conception and found that postconceptional age had a more dominant effect on the fetal lung transcriptome than the sex of the fetus at this early stage of lung development (45). While initial transcriptomic research primarily focused on mRNA, studies are now starting to evaluate the non-coding RNAs-lncRNA, miRNA, and single-cell transcriptomics. Several miRNAs have been shown to play a critical role in lung development and a 2013 study by Syed *et al.* elucidated the role of pharmacological inhibition of miR-34a to prevent and treat hyperoxia induced lung injury in a BPD mouse model (46-48).

Similar to the evolution of the field of genomics, the future of proteomic research is also shifting from establishing associations between individual proteins and BPD to unbiased proteomic analysis. The study of the modulation of protein function as the disease evolves may provide insights into the “real-time” status of the diseases. Magagnotti *et al.* collected tracheal aspirates from preterm infants and performed proteomic analysis to find that there were clear differences in the proteomic profiles of 23–25 and 26–29 weeks’ gestation infants and between infants with mild and severe BPD (49). Recently in a 2022 study, Ahmed *et al.* conducted proteomic analysis on the urine of infants with BPD and validated multiple proteins previously found in serum samples and tracheal aspirates that have been implicated in the pathogenesis of BPD, thereby opening the doors for noninvasive longitudinal monitoring of disease progression (50).

These studies highlight the current state of the field in simultaneously evaluating changes at a molecular level while also focusing on applying those findings to the current clinical status of the patients they come from. Before they could possibly be used bedside for BPD prediction and management, more work needs to be done to contextualize many of these more isolated findings. For instance, which of these are specifically related to BPD rather than prematurity (or illness, inflammation, etc.) as a whole, how many of these findings are specific to one patient population, what effects do the type of analysis have on our findings, and many other considerations.

The use of state-of-the-art computational methods in the future will hopefully enable a more comprehensive and systems-level understanding of the underlying molecular mechanisms driving BPD pathogenesis. By leveraging computational biology to integrate multi-omic data from several tissue sources (e.g., blood, urine, umbilical cord, tracheal aspirates, stool), we can gain a more complete view of BPD. Another future direction includes the integration of machine learning (ML) and artificial intelligence (AI) given the increasingly complex bioinformatic data that will be necessary to accurately understand BPD. For instance, Moreira *et al.* recently derived a five-gene peripheral blood transcriptomic signature using prediction modeling and AI that accurately predicts BPD in the first week of life. They also found that pathways related to T-cell development were associated with BPD (51). The studies in this and the previous sections concerning the use of bioinformatics in BPD are briefly summarized in *Table 4*.

AI and ML are becoming increasingly popular topics in the world of medicine and scientific research. While both can be intimidating on their surface (as can the statistics that underlie them), advances in both fields have been instrumental in helping to make them more accessible to the non-statistician. AI has been described over the past few decades as the advancement of technology, namely computers, to mimic human behavior and understanding (52). This pursuit has been present across multiple different disciplines, medicine among them. The goal of medical AI is to take the vast array of data that is becoming increasingly available and determine how to best use it in clinical decision making. ML, as a subset of AI, seeks to use that same computational power to improve the ability of the computer's ability to "think" and help make complex decisions (53). In other words, how can a model (or machine) collect data on the mistakes it makes and find patterns or strategies to help reduce them.

The advent of chat generative pre-trained transformer (ChatGPT), an online tool developed by open AI, has shown the vast utility of AI (54). This program is not only capable of answering questions and having conversations, but also collects information from these interactions to continue to "teach itself" and make improvements over time. As an example, the paragraph below was written by ChatGPT using the following prompt 'write a paragraph about bioinformatics and bronchopulmonary dysplasia'.

BPD is a complex and multifactorial chronic lung disease that affects premature neonates. BMI has emerged as a powerful tool for understanding the underlying molecular mechanisms driving BPD pathogenesis and developing targeted therapies. By integrating various types of data from genomic, proteomic, and clinical sources, BMI can provide novel insights into the disease's pathogenesis, identify potential biomarkers, and facilitate the development of precision medicine approaches for improved patient outcomes. For example, the use of state-of-the-art computational methods can allow for the integration of miRNAs and other transcriptomic factors into clinically relevant biological pathways. Additionally, data-driven approaches such as ML and network analysis can help identify complex relationships between genetic and environmental factors that contribute to the development of BPD. As we continue to advance our understanding of BPD and other neonatal diseases, BMI will undoubtedly play a critical role in developing more personalized and effective approaches to patient care, ultimately improving the long-term health outcomes of premature neonates affected by this devastating condition.

ChatGPT is one of many AI models currently being designed and evaluated to "learn" from feedback and numerous interactions to improve its function. ChatGPT itself is undergoing many changes, evidenced by the presence of multiple versions (version 3.5 which is free and trained on more limited information through June 2021, compared to version 4 available via subscription that is trained on 5+ times the amount of prompts and is still undergoing updating), highlighting how quickly the field is changing and adapting to the challenges that accompany it (55). While none of these AI models are currently able to successfully answer clinical questions or fully analyze bioinformatic data, they will continue to impact research and clinical discussions going forward.

Optimally, an integrated and comprehensive approach that incorporates clinical informatics (e.g., demographics, morbidities, ventilator settings), laboratory markers (e.g., complete blood count, electrolytes, cultures), imaging informatics (e.g., chest X-rays, computed tomography, or magnetic resonance imaging), data extraction (e.g., natural

Table 4 Sample of literature on the use of GWAS and whole-exome sequencing for BPD

Name of study	Year	GWAS analysis	Cohort	Results	Conclusion
Identification of SPOCK2 as a susceptibility gene for bronchopulmonary dysplasia (25)	2011	DNA pooling strategy to create discovery sets	N=418	*SPOCK2* gene (with lung expression pattern) identified by both discovery tests	*SPOCK2* gene is a possible susceptibility gene for BPD
A genome-wide association study (GWAS) for bronchopulmonary dysplasia (26)	2013	Genotyping using genome DNA + pathways analysis	N=1,726	No SNPs associated with BPD were identified	No genomic loci or pathways were found to account BPD heritability
Integrated genomic analyses in bronchopulmonary dysplasia (27)	2015	Genome wide scan on SNPs	N=751 (with 428 BPD/death)	Association of miR-219 and phosphorous oxygen lyase activity in severe BPD/death/survivors and increased *CD44* and *ADARB2*	Confirmed involvement of known pathways (CD44 and POL activity) and new pathways (ADARB2 and miR-219 targets) in BPD
Genome wide association study of bronchopulmonary dysplasia (28)	2017	Genome-wide SNP genotyping	N=174 (GWAS done on Finnish 24–30-week preterm infants); N=943 (replication cohorts that underwent genotyping of SNPs associated with BPD)	SNP rs11265269 found to be a risk factor for BPD independent of antenatal risks	Variants near the *CRP* gene proposed to be a risk factor for BPD
Ancestry and genetic associations with bronchopulmonary dysplasia in preterm infants (29)	2018	Ancestry and GWAS + admixture mapping	N=387 BPD treated with inhaled nitric oxide	Identified top genetic variant within intron of NBL1. Upregulated genes associate with CCL18 cytokine	Genetic variation contributes to differences in respiratory outcomes after inhaled NO_2
Exome sequencing of neonatal blood spots and the identification of genes implicated in bronchopulmonary dysplasia (31)	2015	Exome sequencing using DNA from neonatal blood sports	N=50 BPD affected and unaffected twin pairs	258 rare genes found, enriched for processes involved in pulmonary structure and function	Rare and high confidence genes are implicated in BPD
Exome sequencing and pathway analysis for identification of genetic variability relevant for bronchopulmonary dysplasia (32)	2015	Exome sequencing	N=26 (with severe BPD)	Identified 3369 novel variants with top candidate genes being *NOS2*, *MMP1*, *CRP*, *LBP* and *TLR* family genes	Identified potential candidate genes for the development of severe BPD
Epigenome-wide association study of bronchopulmonary dysplasia in preterm infants: results from the discovery-BPD program (33)	2022	Illumina 450 K methylation arrays on cord blood DNA and epigenome wide association study	N=107	Total of 313 differentially methylated CpGs associated with BPD, with elevated stochastic epigenetic mutation burden at birth	Potential insights into biological pathways involved in BPD pathogenesis were identified

Table 4 (*continued*)

Table 4 (*continued*)

Name of study	Year	GWAS analysis	Cohort	Results	Conclusion
Age, sexual dimorphism, and disease associations in the developing human fetal lung transcriptome (45)	2015	Transcriptome and gene expression analysis from fetal lung tissue	N=139	Post-conceptual age was a more significant factor than gender differences. Enriched genes found associated with BPD and asthma	Potential insights into developing fetal lung, including involved genes, biological pathways, and association with long term outcomes
MicroRNA in late lung development and bronchopulmonary dysplasia: the need to demonstrate causality (46)	2015	MiRNA gene expression via mouse models	N/A	Only two studies attempted to determine causality or possible pathways, and will need to be validated in human models	Potential insight into genetic basis underlying lung development and eventual development of BPD
Regulation of alveolar septation by microRNA-489 (47)	2015	MiRNA expression	627 mouse miRNAs and 39 mouse viral miRNA	miRNA-489 appears to be associated with inhibiting alveolar septation. Disregulation of miR-489 and downstream genes is associated with hyperoxia associated lung injury and BPD	Identify potential candidate genes for abnormal lung development and possible target for future therapies
Hyperoxia exacerbates postnatal inflammation-induced lung injury in neonatal BRP-39 Null mutant mice promoting the M1 macrophage phenotype (48)	2013	Lung histology, cell count and cytokine analysis	N=12 per litter, otherwise unspecified	BRP-39 (associated with anti-inflammatory effects) more sensitive to hyperoxia and lipopolysaccharide administration, suggesting protective effect	Identified potential biologic pathways and candidate genes for targeted therapy
Calcium signaling-related proteins are associated with broncho-pulmonary dysplasia progression (49)	2013	Gel electrophoresis for proteins obtained via bronchoalveolar lavage	N=12	Changes in protein expression present across changes in gestational age along with BPD severity	Potential insights into biological pathways involved in BPD pathogenesis and molecular changes
Urine proteomics for noninvasive monitoring of biomarkers in bronchopulmonary dysplasia (50)	2022	Urine proteomics via mass spectrometry	N=42	Multiple targets (16 of which associated with FDA approved drugs) associated with BPD-associated changes in the urine	Identified potential biologic pathways in BPD and candidates for targeted therapy
Development of a peripheral blood transcriptomic gene signature to predict 386 bronchopulmonary dysplasia (51)	2022	Whole blood microarray data to develop transcriptomic signature. Development of ML model for BPD prediction and pathway analysis	N=97	4,523 significant genes (FDR <0.01) out of 33,252. Model using 5 genes at day of life 5 outperformed clinical model using birth weight or gestational age	Potential insight into biological pathways underlying BPD. Development of ML model for BPD prediction that performs as well or better than clinical model

GWAS, genome-wide association study; BPD, bronchopulmonary dysplasia; SNPs, single nucleotide polymorphisms; CD44, cluster of differentiation 44; POL, phosphorus oxygen lyase; ADARB2, adenosine deaminase, RNA specific, B2; CRP, C-reactive protein; NBL1, neuroblastoma suppressor of tumorigenicity 1; CCL18, chemokine ligand 18; NO_2, nitrogen dioxide; NOS2, nitric oxide synthase; MMP1, matrix metallopeptidase 1; LBP, lipopolysaccharide binding protein; TLR, toll-like receptor; CpGs, cytosine-guanine dinucleotides; miRNA, microRNA; N/A, not available; FDA, Food and Drug Administration; ML, machine learning; FDR, false discovery rate.

language processing to pull data from clinical documentation, EHR generated reports) and multi-omic bioinformatic data (e.g., metabolomic, transcriptomic, proteomic) intertwined with AI would be instrumental in advancing our understanding of BPD.

Challenges and ethical considerations in bioinformatics

The increasing cost effectiveness and accessibility of bioinformatic tools has introduced a new set of challenges. The most worthwhile among these is the challenge of translating these discoveries to clinical medicine. While there is more information being generated with each passing day, there is significant information that is still to be learned about how and when to best incorporate it. These findings carry serious implications when discussed with patients and it will be imperative that medical personnel using this information are comfortable interpreting results from "big data" and "omics" studies and placing it in appropriate context, especially as more results are automatically being shared with patients and their families. Given the heterogeneity in BPD phenotypes and the multifactorial nature of its onset, the approach of 'one SNP causing one phenotype' is insufficient and requires further study of complex gene-gene and gene-environment interactions (56).

While GWAS have been extensively utilized to establish statistical associations between SNPs and disease states, they do not inform us of the biological basis of the relationship between genetic variants and phenotypic traits (57). This is a field that is relatively new in its use in medicine. This has led to several institutions producing their own primary studies, but we have yet to see larger, pooled analyses comparing findings across populations or across different points in time. We also have not yet seen multi-institutional analyses of differentially expressed genes or pathways of which less is currently known, therefore not giving us the opportunity to learn more about how these variants might all contribute to an outcome of interest. A systems biology-based approach that integrates data from multiple biological levels including genome, transcriptome and proteome may be successful in elucidating these relationships, but further clinical and translational research is needed to validate these genetic associations (58).

Another barrier to bioinformatic research is that tools needed to characterize more precise phenotypic interpretation are either invasive or not accessible to most centers, leading to selection bias and insufficient sample size (59). There is also virtually non-existent training of physicians in these tools in standard medical education and competency in these fields requires a significant amount of time and resources. While bioinformaticians and data scientists are available in some places, access is extremely limited in many centers. Some of these challenges can be circumvented by combining data from multiple sources but significant heterogeneity in bioinformatic data sets may lead to loss of statistical power. While the error rate of current high-throughput sequencing techniques is very low, when extrapolated to the entire human genome, these errors can pose significant hurdles and contribute to high false discovery rates. This makes the accurate identification of novel variants difficult (60). These factors all contribute to a lack of consistent reproducibility across studies when evaluating changes at the molecular level or in the context of biological pathways. Efforts have been made using more strict criteria (such as false discovery rates or corrected P values <0.05 or correcting for multiple comparisons using Bonferroni, Šidák, or similar methods), however these are also not applied in a consistent manner and therefore do not solve the reproducibility problem across studies. Thus, there is a need for a standardization of design and accepted criteria (accepted databases or tools for molecular identification and pathway analysis, significance cutoffs and statistical corrections, among others) and large scale multicenter genomic trials to validate these findings.

Lastly, as with any emerging field of research involving human subjects, bioinformatic research and use of AI and ML algorithms must be done in an ethical, fair, and unbiased manner (61). Care must be taken to avoid data leaks of sensitive patient information thereby preserving privacy and confidentiality (62). Medico-legal risks and potential problems related to insurability if adverse long-term outcomes are predicted early also need to be addressed (61,63). Minority populations are also consistently underrepresented in research and in databases (genomics included) (64) which is not only a problem in current clinical care but will also be of utmost importance as these models start to influence research design and clinical decision making. This also highlights a significant need for transparency of models and databases to ensure this information remains available as applicability to different populations is considered. Future research must balance the ethical aspects of beneficence from these new discoveries with the increased uncertainty and anxiety it might bring for parents and caregivers.

Conclusions

We live in the age of data deluge, where emerging bioinformatic tools have made it possible to obtain massive amounts of genetic information in a timely and cost-effective manner. Modern-day computational analysis including ML algorithms and AI has given us the ability to draw meaningful inferences from this data, to bring the promise of precision medicine to fruition.

Going forward, our success is going to be defined by making this knowledge accessible, explainable, and interpretable for frontline clinicians. Future application of these technologies will hopefully allow us to characterize BPD endotypes, develop novel diagnostics and targeted therapeutics, and enable us to understand the biology of this menacing disease in all its complexity.

Acknowledgments

Funding: This work was supported by Parker B. Francis (to A Moreira), the National Institutes of Health (NIH) Eunice Kennedy Shriver National Institute of Child Health and Human Development (No. K23HD101701 to A Moreira), and NIH National Heart Lung and Blood Institute Subaward (No. 5R25HL126140-06 to A Moreira).

Footnote

Provenance and Peer Review: This article was commissioned by the Guest Editor (Antonio F. Corno) for the series "The Impact of the Progresses of Knowledge and Technologies in Pediatrics" published in *Translational Pediatrics*. The article has undergone external peer review.

Reporting Checklist: The authors have completed the Narrative Review reporting checklist. Available at https://tp.amegroups.com/article/view/10.21037/tp-23-133/rc

Peer Review File: Available at https://tp.amegroups.com/article/view/10.21037/tp-23-133/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tp.amegroups.com/article/view/10.21037/tp-23-133/coif). The series "The Impact of the Progresses of Knowledge and Technologies in Pediatrics" was commissioned by the editorial office without any funding or sponsorship. The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The views expressed in the article are the responsibility of the authors and do not reflect any of the views of the funding agencies.

References

1. Bernstam EV, Smith JW, Johnson TR. What is biomedical informatics? J Biomed Inform 2010;43:104-10.
2. Wyatt JC, Liu JL. Basic concepts in medical informatics. J Epidemiol Community Health 2002;56:808-12.
3. Veinot TC, Ancker JS, Bakken S. Health informatics and health equity: improving our reach and impact. J Am Med Inform Assoc 2019;26:689-95.
4. Conesa A, Beck S. Making multi-omics data accessible to researchers. Sci Data 2019;6:251.
5. Bayat A. Science, medicine, and the future: Bioinformatics. BMJ 2002;324:1018-22.
6. Tsoka S, Ouzounis CA. Recent developments and future directions in computational genomics. FEBS Lett 2000;480:42-8.
7. Roy S. Principles and Validation of Bioinformatics Pipeline for Cancer Next-Generation Sequencing. Clin Lab Med 2022;42:409-21.
8. Anashkina AA, Leberfarb EY, Orlov YL. Recent Trends in Cancer Genomics and Bioinformatics Tools Development. Int J Mol Sci 2021;22:12146.
9. Hu C, Hart SN, Gnanaolivu R, et al. A Population-Based Study of Genes Previously Implicated in Breast Cancer. N Engl J Med 2021;384:440-51.
10. Rosati D, Giordano A. Single-cell RNA sequencing and bioinformatics as tools to decipher cancer heterogenicity and mechanisms of drug resistance. Biochem Pharmacol 2022;195:114811.

11. Graeber TG, Eisenberg D. Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. Nat Genet 2001;29:295-300.
12. Connolly JJ, Hakonarson H. The impact of genomics on pediatric research and medicine. Pediatrics 2012;129:1150-60.
13. Islam JY, Keller RL, Aschner JL, et al. Understanding the Short- and Long-Term Respiratory Outcomes of Prematurity and Bronchopulmonary Dysplasia. Am J Respir Crit Care Med 2015;192:134-56.
14. Northway WH Jr, Rosan RC, Porter DY. Pulmonary disease following respirator therapy of hyaline-membrane disease. Bronchopulmonary dysplasia. N Engl J Med 1967;276:357-68.
15. Stoll BJ, Hansen NI, Bell EF, et al. Neonatal outcomes of extremely preterm infants from the NICHD Neonatal Research Network. Pediatrics 2010;126:443-56.
16. Jobe AH, Bancalari E. Bronchopulmonary dysplasia. Am J Respir Crit Care Med 2001;163:1723-9.
17. Abman SH, Collaco JM, Shepherd EG, et al. Interdisciplinary Care of Children with Severe Bronchopulmonary Dysplasia. J Pediatr 2017;181:12-28.e1.
18. Higgins RD, Jobe AH, Koso-Thomas M, et al. Bronchopulmonary Dysplasia: Executive Summary of a Workshop. J Pediatr 2018;197:300-8.
19. Jensen EA, Dysart K, Gantz MG, et al. The Diagnosis of Bronchopulmonary Dysplasia in Very Preterm Infants. An Evidence-based Approach. Am J Respir Crit Care Med 2019;200:751-9.
20. Gage S, Kan P, Oehlert J, et al. Determinants of chronic lung disease severity in the first year of life; A population based study. Pediatr Pulmonol 2015;50:878-88.
21. Thébaud B, Goss KN, Laughon M, et al. Bronchopulmonary dysplasia. Nat Rev Dis Primers 2019;5:78.
22. Concolino P, Capoluongo E, Santonocito C, et al. Genetic analysis of the dystroglycan gene in bronchopulmonary dysplasia affected premature newborns. Clin Chim Acta 2007;378:164-7.
23. Hadchouel A, Decobert F, Franco-Montoya ML, et al. Matrix metalloproteinase gene polymorphisms and bronchopulmonary dysplasia: identification of MMP16 as a new player in lung development. PLoS One 2008;3:e3188.
24. Sampath V, Garland JS, Le M, et al. A TLR5 (g.1174C>T) variant that encodes a stop codon (R392X) is associated with bronchopulmonary dysplasia. Pediatr Pulmonol 2012;47:460-8.
25. Hadchouel A, Durrmeyer X, Bouzigon E, et al. Identification of SPOCK2 as a susceptibility gene for bronchopulmonary dysplasia. Am J Respir Crit Care Med 2011;184:1164-70.
26. Wang H, St Julien KR, Stevenson DK, et al. A genome-wide association study (GWAS) for bronchopulmonary dysplasia. Pediatrics 2013;132:290-7.
27. Ambalavanan N, Cotten CM, Page GP, et al. Integrated genomic analyses in bronchopulmonary dysplasia. J Pediatr 2015;166:531-7.e13.
28. Mahlman M, Karjalainen MK, Huusko JM, et al. Genome-wide association study of bronchopulmonary dysplasia: a potential role for variants near the CRP gene. Sci Rep 2017;7:9271.
29. Torgerson DG, Ballard PL, Keller RL, et al. Ancestry and genetic associations with bronchopulmonary dysplasia in preterm infants. Am J Physiol Lung Cell Mol Physiol 2018;315:L858-69.
30. Yu KH, Li J, Snyder M, et al. The genetic predisposition to bronchopulmonary dysplasia. Curr Opin Pediatr 2016;28:318-23.
31. Li J, Yu KH, Oehlert J, et al. Exome Sequencing of Neonatal Blood Spots and the Identification of Genes Implicated in Bronchopulmonary Dysplasia. Am J Respir Crit Care Med 2015;192:589-96.
32. Carrera P, Di Resta C, Volonteri C, et al. Exome sequencing and pathway analysis for identification of genetic variability relevant for bronchopulmonary dysplasia (BPD) in preterm newborns: A pilot study. Clin Chim Acta 2015;451:39-45.
33. Wang X, Cho HY, Campbell MR, et al. Epigenome-wide association study of bronchopulmonary dysplasia in preterm infants: results from the discovery-BPD program. Clin Epigenetics 2022;14:57.
34. Kumuthini J, Chimenti M, Nahnsen S, et al. Ten simple rules for providing effective bioinformatics research support. PLoS Comput Biol 2020;16:e1007531.
35. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27-30.
36. Kanehisa M, Furumichi M, Sato Y, et al. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res 2023;51:D587-92.
37. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25-9.
38. Gillespie M, Jassal B, Stephan R, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res 2022;50:D687-92.
39. Weng L, Macciardi F, Subramanian A, et al. SNP-based

pathway enrichment analysis for genome-wide association studies. BMC Bioinformatics 2011;12:99.

40. Cariaso M, Lennon G. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. Nucleic Acids Res 2012;40:D1308-12.
41. Promethease. Available online: https://promethease.com/
42. de Leeuw CA, Mooij JM, Heskes T, et al. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol 2015;11:e1004219.
43. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207-10.
44. Lal CV, Bhandari V, Ambalavanan N. Genomics, microbiomics, proteomics, and metabolomics in bronchopulmonary dysplasia. Semin Perinatol 2018;42:425-31.
45. Kho AT, Chhabra D, Sharma S, et al. Age, Sexual Dimorphism, and Disease Associations in the Developing Human Fetal Lung Transcriptome. Am J Respir Cell Mol Biol 2016;54:814-21.
46. Nardiello C, Morty RE. MicroRNA in late lung development and bronchopulmonary dysplasia: the need to demonstrate causality. Mol Cell Pediatr 2016;3:19.
47. Olave N, Lal CV, Halloran B, et al. Regulation of alveolar septation by microRNA-489. Am J Physiol Lung Cell Mol Physiol 2016;310:L476-87.
48. Syed MA, Bhandari V. Hyperoxia exacerbates postnatal inflammation-induced lung injury in neonatal BRP-39 null mutant mice promoting the M1 macrophage phenotype. Mediators Inflamm 2013;2013:457189.
49. Magagnotti C, Matassa PG, Bachi A, et al. Calcium signaling-related proteins are associated with broncho-pulmonary dysplasia progression. J Proteomics 2013;94:401-12.
50. Ahmed S, Odumade OA, van Zalm P, et al. Urine Proteomics for Noninvasive Monitoring of Biomarkers in Bronchopulmonary Dysplasia. Neonatology 2022;119:193-203.
51. Moreira A, Tovar M, Smith AM, et al. Development of a peripheral blood transcriptomic gene signature to predict bronchopulmonary dysplasia. Am J Physiol Lung Cell Mol Physiol 2023;324:L76-87.
52. Ramesh AN, Kambhampati C, Monson JR, et al. Artificial intelligence in medicine. Ann R Coll Surg Engl 2004;86:334-8.
53. Choi RY, Coyner AS, Kalpathy-Cramer J, et al. Introduction to Machine Learning, Neural Networks, and Deep Learning. Transl Vis Sci Technol 2020;9:14.
54. ChatGPT. 2021. GPT-3.5 based language model. Available online: https://chat.openai.com
55. Martindale J. GPT-4 vs. GPT-3.5: how much difference is there? 2023. Available online: https://www.digitaltrends.com/computing/gpt-4-vs-gpt-35
56. Fernald GH, Capriotti E, Daneshjou R, et al. Bioinformatics challenges for personalized medicine. Bioinformatics 2011;27:1741-8.
57. Frazer KA, Murray SS, Schork NJ, et al. Human genetic variation and its contribution to complex traits. Nat Rev Genet 2009;10:241-51.
58. Kohl P, Crampin EJ, Quinn TA, et al. Systems biology: an approach. Clin Pharmacol Ther 2010;88:25-33.
59. Pierro M, Van Mechelen K, van Westering-Kroon E, et al. Endotypes of Prematurity and Phenotypes of Bronchopulmonary Dysplasia: Toward Personalized Neonatology. J Pers Med 2022;12:687.
60. Pammi M, Aghaeepour N, Neu J. Multiomics, artificial intelligence, and precision medicine in perinatology. Pediatr Res 2023;93:308-15.
61. Wang S, Jiang X, Singh S, et al. Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States. Ann N Y Acad Sci 2017;1387:73-83.
62. Oliva A, Grassi S, Vetrugno G, et al. Management of Medico-Legal Risks in Digital Health Era: A Scoping Review. Front Med (Lausanne) 2021;8:821756.
63. Richardson A, Ormond KE. Ethical considerations in prenatal testing: Genomic testing and medical uncertainty. Semin Fetal Neonatal Med 2018;23:1-6.
64. Konkel L. Racial and Ethnic Disparities in Research Studies: The Challenge of Creating More Diverse Cohorts. Environ Health Perspect 2015;123:A297-302.

Cite this article as: Valadie CT, Arya S, Arora T, Pandillapalli NR, Moreira A. A bioinformatics approach towards bronchopulmonary dysplasia. Transl Pediatr 2023;12(6):1213-1224. doi: 10.21037/tp-23-133