Journal of Data Science 0 (0), 1–15
??? 2026
DOI: 10.6339/26-JDS1250
Computing in Data Science

A Tabular Presentation of Variable Development

Lydia R. Lucchesi[1,*] and Roger D. Peng[1]

[1]*Department of Statistics and Data Science, University of Texas at Austin, USA*

Abstract

Data scientists regularly incorporate many different datasets into data analyses. Yet, communication tools for organizing important information about the different datasets used and how they relate to the key variables of the analysis are lacking. Therefore, in this paper we propose a new communication tool: the variTable. The variTable is a seven-column table that consolidates, organizes, and summarizes high-level variable development information for inclusion in original research articles. Specifically, the variTable shows for the key variables in an analysis what data sources were used, how they were used, and whether they were altered for the analysis. Additionally, to make building variTables easier, we present the `variTable` R package. We illustrate use of the variTable structure and R package for a real-world observational study that used nine different datasets to examine the association between air pollution and asthma in children. Ultimately, we hope that the variTable can serve as a practical communication tool for sharing and understanding important information about the variable development process in data analysis.

Keywords *analytic design theory; data preprocessing; data science; scientific publishing*

1 Introduction

In quantitative research, scientific publications serve as the primary means of disseminating new findings, alongside information about how those findings were produced. Publishing this combination of information—results *and* methods—is meant to ensure that others can comprehend, evaluate, replicate, act on, and/or build on the research (Gastel and Day, 2022). For a simple data analysis, authors might detail the data source, the analytical method, and the output of that method and, in turn, successfully support the end-user actions listed above. However, as the complexity and scale of a data analysis increases, communicating the details of the analysis in a way that is both comprehensive and comprehensible can become challenging.

Consider, for example, an observational study on the association between air pollution and asthma. This type of study often involves gathering environmental, health, and demographic data from a variety of sources. The data in those sources may then require considerable preprocessing before any statistical modeling can happen. In some analyses, developing the air quality variable alone might require sourcing air quality, satellite, and weather data to build a prediction model that can then produce air quality "data" for geographies where there is none. For analyses of this nature, producing an organized narrative of the *variable development*—the sourcing and preprocessing of data to instantiate the variables of interest—becomes a nontrivial task.

*Corresponding author. Email: lydia.lucchesi@austin.utexas.edu.

© 2026 The Author(s). Published by the School of Statistics and the Center for Applied Statistics, Renmin University of China. Open access article under the CC BY license.
Received December 16, 2025; Accepted October 25, 2026

For analyses such as the one above, there are many different pieces of variable development information to communicate. There is basic information about the different datasets, how they were used, whether they had to be altered from their original state, and which datasets relate to which variables in the analysis. As we will illustrate in Section 4, this information may get dispersed across different parts/pages of an original research article. Additionally, the information may get shared alongside an original research article in files of data or code, with the aim of the making the analysis reproducible (National Academies of Sciences, Engineering, and Medicine, 2019; Peng, 2011). However, practical communication methods that consolidate important high-level variable development details within an original research article are lacking.

Therefore, we propose an *additional* form of communication, designed to consolidate, organize, and summarize high-level details about variable development, within the main text of an original research paper. Specifically, we propose the *variable development table*—or *variTable*—a standardized tabular presentation of data source and preprocessing information for variables in an analysis. The proposed table schema consists of two columns of variable information on the left, three columns of data source information in the center, and two columns of data preprocessing information on the right. Four columns are categorical; three are text fields. We also develop the `variTable` R software package for variTable construction. It provides a user interface within an RStudio session to build variTables and export them as LaTeX or HTML code.

In this work, we suggest a tabular presentation for this communication task, for the following reasons. In scientific publications, tables are a widely accepted and utilized format for communicating information that would otherwise be tedious to read as paragraphs of text (but that is important to have access to nonetheless) (Buehler, 1977). Also, by consolidating and organizing information, tables support readers in locating specific pieces of information and identifying information patterns (Arnold, 1962). Consider "Table 1" in epidemiological and clinical research; its structured and largely standardized statistical description of the study population helps readers to evaluate the internal and external validity of a study (Hayes-Larson et al., 2019). Ultimately, we argue that a standardized table has logical applications in the communication of variable development information, especially for analyses with variables developed from multiple data sources and preprocessing decisions.

The remainder of this paper is organized as follows. In Section 2, we discuss related work in variable development communication. We also present the three variTable design goals, which relate to current gaps in variable development communication. Then in Section 3, we present the variTable design. We discuss the six design principles (Section 3.1) that shaped the variTable and the seven-part table schema (Section 3.2), which includes novel categorizations of data source and preprocessing information. In Section 4, we demonstrate use of the variTable and document variable development for a real-world observational study on the association between air pollution and asthma in children. This is followed by a description of the `variTable` R package in Section 5. We end in Section 6 noting limitations of this work as well as directions for future work.

2 Related Work and Design Goals

In this section, we discuss related work in variable development communication, before presenting the three main design goals for the variTable. The discussion on related work focuses on communication within and supplementary to original research articles. As mentioned in Section 1, we use the term *variable development* to refer to the task of sourcing and preprocessing

data, to instantiate the variables of interest in a data analysis. Although the variTable is designed primarily for variable development involving multiple existing datasets, the discussion on related work remains broad. We begin this section considering how variable development through primary data collection is communicated in original research articles.

In quantitative research, variable development based on primary data collection is supported by a well-established reporting protocol. For original research articles, authors are expected to detail the survey design, the subjects, the measurement procedures, and the variables measured (Gastel and Day, 2022; Nolan and Stoudt, 2021). Authors may also discuss missing data and data exclusion criteria (Nolan and Stoudt, 2021). These important data collection details can typically be found in the *Methods* section of the IMRAD (*Introduction*, *Methods*, *Results*, and *Discussion*) paper structure. Notably, this reporting protocol is not necessarily applicable to variable development involving multiple existing datasets, as the work involved is significantly different. Instead of planning and implementing a data collection protocol, the focus is on finding, transforming, and combining datasets.

Also of note are reporting checklists for different study types and fields. These usually contain items related to variable development. For randomized controlled trials (RCTs), there is the 30-item CONSORT checklist (Hopewell et al., 2025), with items like *Trial design* and *Outcomes*. For observational studies in epidemiology, there is the 22-item STROBE checklist (von Elm et al., 2008), with items like *Variables* and *Data sources/measurement*. The 32-item REFORMS checklist (Kapoor et al., 2024), designed for machine learning (ML) research, has sections titled *Data quality* and *Data preprocessing*. These checklists contain guidance on *what* to report for various aspects of variable development. However, there is less guidance on *how* to report it neatly, when dealing with a significant number of datasets. We develop the variTable to help with this *how* part.

Information on variable development can also increasingly be found in supplementary materials or repositories of data/code. Concerns about reproducibility have resulted in recommendations to make data and code publicly available where possible (National Academies of Sciences, Engineering, and Medicine, 2019; Peng, 2011). Concerns about data quality in the ML domain have motivated new dataset documentation frameworks. For instance, the *datasheets for datasets* framework (Gebru et al., 2021) prompts dataset creators to answer up to 57 questions about different aspects of a dataset, such as its creation, uses, and maintenance. Similarly, the Dataset Nutrition Label (or Label) (Chmielinski et al., 2022; Holland et al., 2018) is a modular framework for documenting the "ingredients" of a dataset, with the most recent version of the Label being shared through a web-based GUI. These forms of documentation will typically not be found in the main text of an original research article. In this work, we focus on an additional form that embeds high-level details about the variable development process (involving *multiple* datasets) within the main text, to support readers aiming to understand and evaluate the research, more broadly.

Specifically, we develop the variTable with three design goals (DG1, DG2, and DG3) in mind. The following goals relate to the benefits—that we envision—of having access to and consulting a variTable in an original research paper.

- **DG1**: The variTable consolidates high-level variable development information, such that the reader does not have to search for it.
- **DG2**: The variTable indicates—through its appearance and content—the complexity of the variable development process, without overwhelming the reader.
- **DG3**: The variTable reduces ambiguity around whether or not data preprocessing occurred, by standardizing a high-level reporting of it.

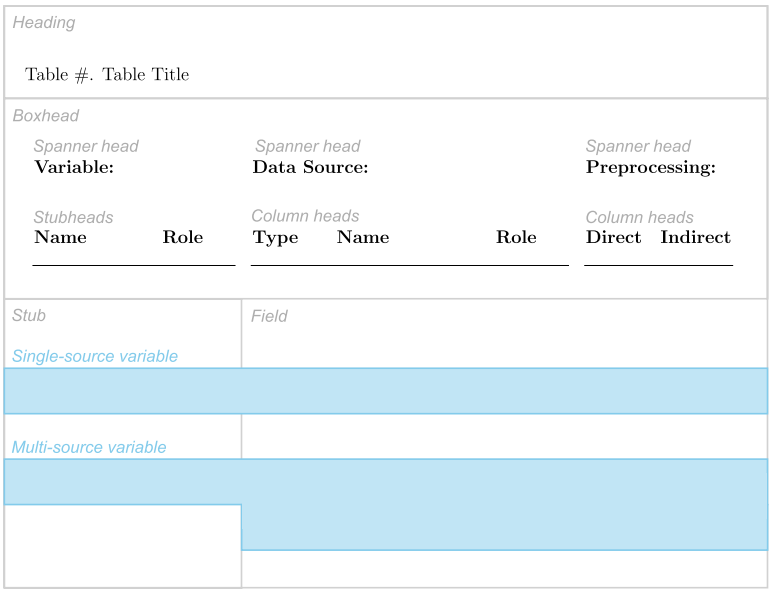


Figure 1: The variTable schema and structure shown in black, with annotations for the standard parts of an information table outlined and labeled in gray. Blue regions indicate the general shape of single-source and multi-source variable entries.

Notably, these design goals center around the experience of a paper *reader*, as opposed to a paper *author*. We focus on paper authors in Section 5, when we introduce a software tool for making variTables: the `variTable` R package.

3 Variable Development Tables

In this section, we discuss in detail the main contribution of this work: the variable development table, or variTable. Specifically, we discuss the design principles that shaped the variTable (Section 3.1) and describe each column of the seven-part table schema (Section 3.2).

3.1 Design Principles

There are six variTable design principles that fall under three broad categories: table type, table layout, and table content. We present the principles by category.

Table type. "Table" is a broad term encompassing both information tables and data tables. Therefore, for the variTable, we identified two design principles that clarify the table type and its general purpose: 1) *information table* and 2) *formal communication table*. Namely, the variTable is an information table for formal research communications, as opposed to a data table for data storage. These principles informed the general direction of the variTable design. We focused on creating a table that was human-readable (rather than machine-readable) and ready for inclusion in original research articles. Therefore, we composed the variTable structure to include a heading, boxhead, stub, and field—standard parts of an information table in formal research communications (Buehler, 1977; Jenkinson, 1949). The variTable structure is visualized and annotated in Figure 1.

Table layout. According to Wainer (1992), the first rule of table construction is to "[o]rder the rows and columns in a way that makes sense." We identify the following two principles, related to the column and row layout of the variTable: *variable-centric columns* and *model-ordered rows*. We focus on the column ordering first, as the row ordering is dependent on it.

For the ordering of the columns, we chose a variable-centric approach, meaning the variables are the primary object of focus in the variTable. This choice aims to ensure the variTable supports readers in their interpretation of the analytical outputs, which typically center around the variables of interest. To achieve a variable-centric approach, we place variable information (e.g., the variable names) in the table stub, i.e., the left-most column(s) (see Figure 1). As a result, the table rows become about the variables. In particular, all subsequent columns to the right of the stub are interpreted as providing information about the variables.

One can then apply an ordering to the rows, i.e., variables. Where applicable, we recommend the variables be *model-ordered*. Namely, the variables are ordered in a way that resembles how the analytical model used in the analysis is generally written/presented. For example, if a linear regression model was fit, the outcome variable would be first, followed by the main explanatory variables. This combination of *variable-centric columns* and *model-ordered rows* is meant to place the most important variables in the top left corner of the table, where they will be seen first.

One potential downside to using *variable-centric* columns with *model-ordered* rows is repetition of dataset information in the table. If variables use the same dataset, the dataset information has to be listed separately for each variable. To avoid this repetition, we considered *dataset-centric* columns (putting datasets rather than variables in the table stub) or *dataset-ordered* rows (placing variables that use the same dataset next to each other, vertically), which would allow for row spans based on datasets. However, with dataset-centric columns, the information for a variable would get disjointed in the table, if the variable was developed using more than one data source. Multi-source variables could also pose problems for the dataset-ordered rows approach. Note that in general achieving both dataset-ordered and model-ordered rows would be difficult/unlikely. Ultimately, we chose to put the main variables of interest at the top of the table and keep all variable information together, at the cost of having to repeat dataset information (see Figure 1).

Table content. The last two design principles—*minimal* and *categorical*—relate to the content in the table cells. *Minimal* refers to the amount of content in each table cell, and *categorical* refers to the type of content. The majority of the variTable cells are designed to contain a single word or symbol that represents a category. This level of minimalism can make the table less informative, and the use of specific categories can make the table less applicable/adaptable to different analyses. However, there are certain benefits to the approach as well. Firstly, having minimal text in the cells means that the text that *is* there can be more legible (Rutter, 2017). Secondly, the columns remain narrow, allowing for more columns—that cover a range of variable development information—to be included in the table. Lastly, having pre-defined categories for certain columns means that contents can be compared systematically within and across tables. It also makes for a simpler point-and-click user interface for building variTables (see Section 5).

It is important to note that we use these principles under the assumption that the variTable will not be the only source of variable development information in an original research article. The limited content in the variTable makes it insufficient for that. We envision it as a supplement to more descriptive text throughout the article as well as data and code shared elsewhere.

3.2 Table Schema

As shown in Figure 1, the proposed variTable schema has seven columns. On the left side, two columns constitute the stub and contain variable information. They fall under the spanner head of *Variable*. Next, three columns in the table field, under the spanner head of *Data Source*, contain data source information. Finally, two columns about data preprocessing fall under the spanner head of *Preprocessing*. Each column is detailed below, in the order it appears in the table.

Variable: Name. This is the left-most column. It is a text field that contains a short name for a variable in the analysis. For a variable involving more than one data source, the cell will be blank for all rows after the first. For example, if a variable is developed with three data sources, there will be two blank cells in this column, under the initial listing of the variable name. Visually, this creates a grouping of the related rows for a variable, indicating a multi-source variable.

Variable: Role. The next column identifies the role of the variable in the analysis. Common options might be *outcome*, *predictor*, *treatment*, *covariate*, *confounder*, etc. As discussed in Section 2, the entries in this column should follow a logical ordering, i.e., be *model-ordered* (e.g., outcome first, predictor second, etc.). In other words, the table rows are ordered based on this column. The column is a free text field, as we do not define a set number of categories, given that the ways in which the variable "role" can be referred to are broad and domain-specific. Regardless, a one-word entry is preferred. Note that, where the adjacent **Variable: Name** cell is blank, this cell will also be blank. That is, for each variable, the name and role are listed only once.

Data Source: Type. This is the first column of the table field. It is categorical and has three levels referring to different types of data: *Primary*, *Secondary*, and *Tertiary*. Two levels are frequently used labels for data; *Primary* and *Secondary* refer to data collected firsthand and data collected by others, respectively (Blaikie, 2003). The third label is less common but provides an important point of distinction; *Tertiary* refers to analytical outputs (e.g., model predictions) repurposed as "data" in a new analysis (Blaikie, 2003).

Data Source: Name. This column is a text field for the name of a data source. Data source names can be long, so it will likely be the widest column in the table. However, one may choose to use acronyms where appropriate (e.g., ACS for American Community Survey).

Data Source: Role. This categorical column has two levels: *Main* and *Support*. The levels refer to the role that a data source plays in the variable development process. Data from a *Main* source provide (or are preprocessed into) the values of the variable. Data from a *Support* source provide auxiliary information used in preprocessing. For example, to crosswalk data in a *Main* source from one format to another, one may need to use a *Support* source (with data about both formats) to implement the crosswalk.

Preprocessing: Direct. This column is binary. It either contains a check mark to indicate that "direct" preprocessing of the adjacent data source occurred or remains empty to indicate that it did not. In the variTable, direct preprocessing refers to data preprocessing that is directly related to the variable and its preparation for subsequent analytical steps. Examples of direct preprocessing could include binning, transforming, or imputing values in the variable of interest.

Preprocessing: Indirect. This is the final column in the table. It is binary. To indicate that "indirect" preprocessing of the adjacent data source occurred, the cell is populated with a check mark. To indicate that it did not, the cell is left empty. In the variTable, indirect preprocessing refers to data preprocessing that is *not* directly related to the variable of interest yet affects the values of it. An example of this would be filtering dataset rows in an unrelated

data preprocessing step for a different variable, that results in data loss in the variable of interest.

4 Application

This section illustrates use of the variTable for communicating variable development information. We compare two different variable development scenarios to highlight the primary application of the variTable (Section 4.1). We then use the variTable to document variable development for an observational study on the association between air pollution and asthma in children (Section 4.2).

4.1 Differences in Variable Development

Variable development can look substantially different across studies and analyses. For example, in a randomized controlled trial (RCT), we would generally expect to see the use of primary data collection with minimal data preprocessing. Alternatively, in some observational studies, we might expect to see the use of multiple existing datasets that require data preprocessing to transform what is available in the datasets into what is needed for the analysis. In Figure 2, we present two partially-populated example variTables for a hypothetical RCT and observational study. The RCT scenario in Figure 2(a) has two variables: outcome and treatment. The observational study scenario in Figure 2(b) has six variables: outcome, exposure, and four covariates.

As previously noted, in this paper we are primarily concerned with variable development involving multiple (mainly existing) datasets. Generally-speaking, variable development based on primary data collection alone is unlikely to necessitate a variTable. All variables will have the same data source, so all rows in the variTable will have exactly the same information for columns 3–5. This type of scenario is visualized in Figure 2(a).

Figure 2(b) visualizes a notably different scenario and one that the variTable is designed to document: the use of multiple existing datasets, not all immediately functional for the intended analysis and therefore needing preprocessing. This type of variable development results in more *unique* pieces of data source and preprocessing information to communicate, across all variables. In turn, the variTable becomes a way to consolidate and organize the pieces, such that they are not overlooked or challenging to piece together for paper readers. From the variTable in Figure 2(b), we can quickly see that variables were developed from different datasets and combinations of datasets, some of which required preprocessing. Section 4.2 considers a real-world example of variable development similar to the hypothetical scenario in Figure 2(b).

4.2 A Study on Air Pollution and Asthma

To illustrate use of the variTable in practice, we build an example variTable for a published observational study (for which the second author of this paper is a coauthor). We then visually demonstrate how this particular variTable consolidates variable development information spread across multiple pages in the original research article. We also discuss how the variTable can be viewed holistically, scanned horizontally, or scanned vertically to get different insights into and overviews of the variable development process.

The observational study of interest is by Keet et al. (2018). The study estimated the association between long-term exposure to coarse particulate matter (PM) and asthma prevalence in

a) Randomized controlled trial (RCT)

Variable:		Data Source:			Preprocessing:	
Name	**Role**	**Type**	**Name**	**Role**	**Direct**	**Indirect**
	Outcome	Primary		Main	✓	
	Treatment	Primary		Main		✓

b) Observational study with existing data sources

Variable:		Data Source:			Preprocessing:	
Name	**Role**	**Type**	**Name**	**Role**	**Direct**	**Indirect**
	Outcome	Secondary		Main	✓	✓
	Exposure	Secondary		Main	✓	
		Secondary		Support		
		Tertiary		Support		
	Covariate	Secondary		Main		✓
	Covariate	Secondary		Main	✓	✓
		Secondary		Support		
	Covariate	Secondary		Main	✓	✓
	Covariate	Secondary		Main	✓	

Figure 2: Two partially-populated variTables for the two hypothetical variable development scenarios discussed in Section 4.1. White outlined boxes in the first column indicate where a variable name would go. Gray filled boxes in the fourth column indicate where a source name would go, with shades representing different source names (for each variTable).

children enrolled in Medicaid. The authors first fit a semiparametric regression model to predict PM exposure for the contiguous United States. Then, using those predictions, they estimated the association between PM exposure and asthma prevalence with a log-linear regression model that also adjusted for several demographic and socioeconomic factors. In total, nine different datasets were used for the analysis.

For the variable development part of this study, we build a variTable, shared as Table 1. Table 1 brings together information communicated across five different sub-sections of the methods section of Keet et al. (2018). When looking at Table 1 as a whole, we immediately get a sense of the number of variables used and of the non-uniformity of the rows in the table field, indicating to us that a variety of variable development work was involved.

From the variTable, we can also learn specifically about each variable, by focusing on the rows in the table. Figure 3 shows how row 1 in the variTable consolidates variable development information—from across the article—to provide a concise high-level variable development summary for the outcome variable. Reading row 1 from left to right, we get the following information: *Prevalent asthma is the outcome variable. It was derived using data from one secondary data source: the Medicaid Analytic Extract, which required data preprocessing.*

Additionally, one benefit of a two-dimensional tabular presentation is that it can be read both horizontally and vertically to get different overviews of the information. This benefit applies to the variTable. Like the rows, the columns of the variTable consolidate related information spread across the article. As shown in Figure 4, we can vertically scan the *Data Source Name* column to see the list of all nine data sources used in variable development. Due to the structure

Table 1: Example variTable for the study by Keet et al. (2018) discussed in Section 4.2.

Variable:		Data Source:			Preprocessing:	
Name	Role	Type	Name	Role	Direct	Indirect
Prevalent asthma	Outcome	Secondary	Medicaid Analytic Extract	Main	✓	✓
Coarse PM	Exposure	Secondary	EPA AQS Database	Main	✓	
		Secondary	U.S. Census Bureau 2010 DP1	Support		
		Secondary	U.S. Census Bureau 2010 TIGER/Line Shapefiles	Support		
		Secondary	National Land Cover Database	Support		
		Secondary	2008 National Emission Inventory	Support		
Age	Covariate	Secondary	Medicaid Analytic Extract	Main		✓
Sex	Covariate	Secondary	Medicaid Analytic Extract	Main		✓
Race/ethnicity	Covariate	Secondary	Medicaid Analytic Extract	Main		✓
Urbanicity	Covariate	Secondary	National Center for Health Statistics	Main		
Poverty	Covariate	Secondary	U.S. Census Bureau 2011 S1702	Main		
Education	Covariate	Secondary	U.S. Census Bureau 2011 S1501	Main		

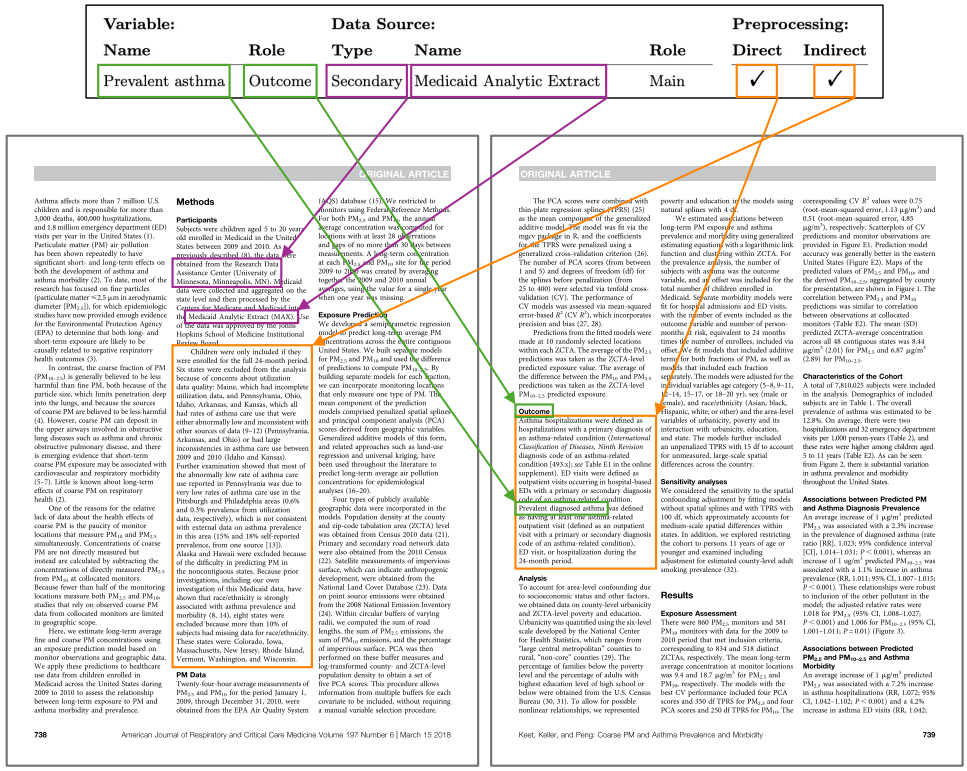
of the variTable, this vertical scan of the column also reveals how much each data source was used. For example, we can see that the Medicaid Analytic Extract dataset was used for four different variables. Moreover, based on the vertical spacing of the data source names, we can see which data sources were used in conjunction for the development of a single variable (as they sit closer together, vertically). For example, we can see a cluster of five sources toward the top of the table, indicating that one variable alone was derived using five different datasets.

In the next section, we present the `variTable` R package, which generated the LaTeX code for this example variTable (Table 1).

5 Software

In this section, we present the `variTable` R software package (Lucchesi and Peng, 2026). With this package, we explore how software can be designed to assist quantitative researchers (or paper authors) with variTable construction. The goal is to make variTable construction simple. In turn, the `variTable` R package creates a structured space for populating a variTable and automates formatting. Specifically, the package provides an interactive user interface for entering table inputs and generating the HTML or LaTeX code that renders a variTable. The package also provides several R functions for working with the table contents in R directly. These different features—and how they were implemented—are discussed next.

When the `variTable` R package is loaded in an RStudio session (Posit Team, 2025), the user interface for building variTables can be opened in either the RStudio viewer pane or a browser window. Users simply run the `variTable()` function to launch the interface. As shown

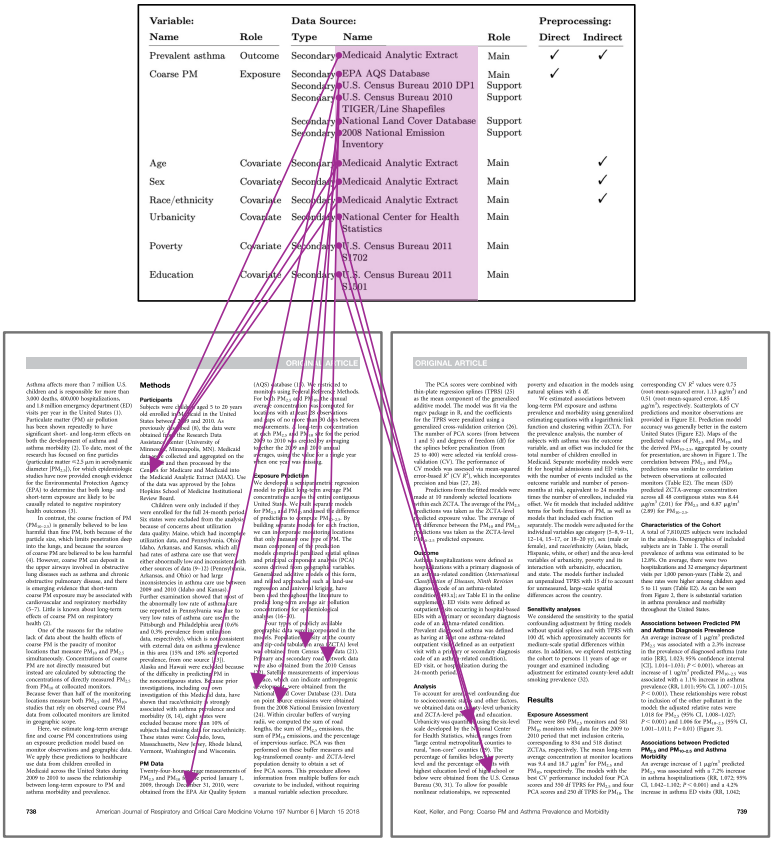


Reprinted with permission of the American Thoracic Society.
Copyright © 2025 American Thoracic Society. All rights reserved.
Cite: Keet, C. A., Keller, J. P., & Peng, R. D. (2018). Long-Term Coarse Particulate Matter Exposure Is Associated with Asthma among Children in Medicaid. *American Journal of Respiratory and Critical Care Medicine*, 197(6), 737–746.
The American Journal of Respiratory and Critical Care Medicine is an official journal of the American Thoracic Society.

Figure 3: Visualization showing the link between each cell in the first row of the example variTable from Section 4.2 (Table 1) and the text from the original research article it contains or represents. The colors of the boxes and arrows correspond with the three main sections of the variTable outlined in Section 3.2: *variable* (green), *data source* (purple), and *preprocessing* (orange).

in Figure 5, the user interface is divided into two sections: *Table Input* and *Table Preview*. As a user enters information into the *Table Input* section, a live preview of the variTable is provided in the *Table Preview* section. Users also have the option to enable advanced input options, including *Dynamic Inputs* and *Input Notes*. Dynamic inputs are URLs that result in hyperlinks in the variTable. Input notes are documentation about the variTable inputs and do not appear in the variTable.

Through the user interface, users can also save inputs in a JSON file and export variTables as HTML or LaTeX code. Saving inputs in a JSON file can be used to save progress on variTable construction and/or collaborate on variTable construction across different machines. Specifically, when launching the interface, a user can pass in a JSON file with saved inputs, so that the inputs are automatically reloaded into the *Table Input* section of the user interface. This action also causes the *Table Preview* section to update. From the *Table Preview* section, users can export variTables as HTML or LaTeX code. The variTable presented in Section 4.2 (Table 1) was created with the `variTable` R package and exported as LaTeX code for inclusion in this paper.



Variable: Name	Role	Data Source: Type	Name	Role	Preprocessing: Direct	Indirect
Prevalent asthma	Outcome	Secondary	Medicaid Analytic Extract	Main	✓	✓
Coarse PM	Exposure	Secondary	EPA AQS Database	Main	✓	
		Secondary	U.S. Census Bureau 2010 DP1	Support		
		Secondary	U.S. Census Bureau 2010 TIGER/Line Shapefiles	Support		
		Secondary	National Land Cover Database	Support		
		Secondary	2008 National Emission Inventory	Support		
Age	Covariate	Secondary	Medicaid Analytic Extract	Main		✓
Sex	Covariate	Secondary	Medicaid Analytic Extract	Main		✓
Race/ethnicity	Covariate	Secondary	Medicaid Analytic Extract	Main		✓
Urbanicity	Covariate	Secondary	National Center for Health Statistics	Main		
Poverty	Covariate	Secondary	U.S. Census Bureau 2011 S1702	Main		
Education	Covariate	Secondary	U.S. Census Bureau 2011 S1501	Main		

Reprinted with permission of the American Thoracic Society.
Copyright © 2025 American Thoracic Society. All rights reserved.
Cite: Keet, C. A., Keller, J. P., & Peng, R. D. (2018). Long-Term Coarse Particulate Matter Exposure Is Associated with Asthma among Children in Medicaid. *American Journal of Respiratory and Critical Care Medicine*, 197(6), 737–746.
The American Journal of Respiratory and Critical Care Medicine is an official journal of the American Thoracic Society.

Figure 4: Visualization showing where each data source name in the fourth column of the example variTable from Section 4.2 (Table 1) is located in the text of the original research article.

The `variTable` package also provides several R functions for working programmatically with inputs from the interactive user interface. Two of these functions provide a programmatic alternative to a button in the user interface. Namely, users can save inputs and export variTables with the `saveInputs()` and `exportTable()` functions, respectively. Another function—`tidyInputs()`—can be used to convert inputs from the user interface into a tibble data frame (Müller and Wickham, 2025) in the tidy data format (Wickham, 2014) in a user's R global environment. Use of this function is demonstrated in the console section of Figure 5. Furthermore, related to this last function, in Section 6 we discuss future work on the treatment of variTable contents as data, for exploring variable development visually and quantitatively.

The package is comprised of R, HTML, CSS, and JavaScript code. We used the latter three languages to build the user interface (UI). The HTML and CSS code set the layout and style of the UI, while the JavaScript code adds the interactive and responsive elements. The UI is defined in an HTML file and loaded as a local web page in either the RStudio viewer pane or a browser window, depending on the preference of the user. Additionally, when the UI is loaded, a WebSocket connection is established with the local server to create connectivity between the UI and the RStudio session. Namely, it is through this WebSocket connection that a user's inputs

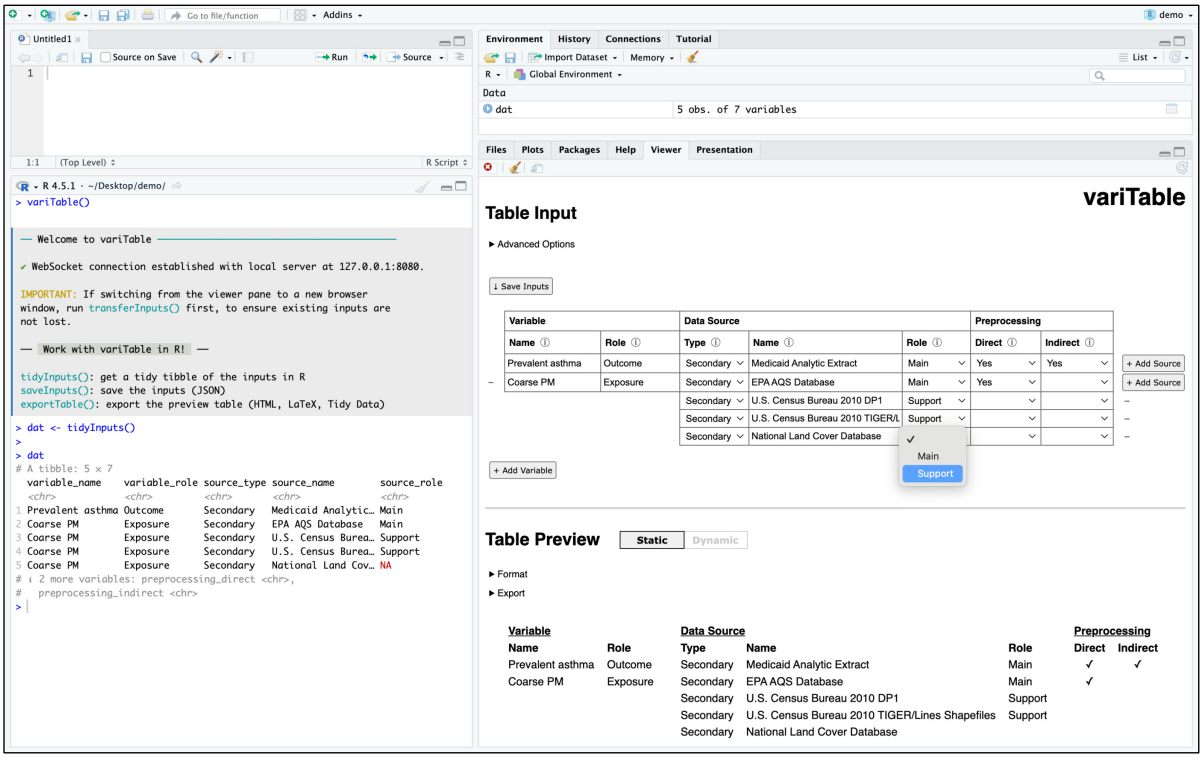


Figure 5: Screenshot of an RStudio session, in which the `variTable` user interface is displayed in the viewer pane (bottom-right pane). Example inputs to the user interface come from the real-world observational study discussed in Section 4.2. The console pane (bottom-left pane) demonstrates use of the `variTable()` function for launching the user interface and the `tidyInputs()` function for converting inputs to the user interface into a (tidy) tibble in R.

to the UI can simultaneously be accessed from the R console via the aforementioned auxiliary `variTable` R functions (e.g., `tidyInputs()`).

6 Discussion

When writing original research articles, the task of describing the variable development process—when it involves multiple existing datasets and various data preprocessing decisions—can be challenging. In this paper, we propose a standardized tabular presentation of variable development information: the variable development table, or variTable. The variTable consists of seven columns. These columns consolidate, organize, and summarize high-level variable development information, so that a reader does not have to search for and assemble this information on their own. We demonstrate this consolidation of information in Section 4.2, creating a variTable that brings together information spread across multiple pages of a published research article, to provide an overview of the variable development process (see Table 1).

In designing the variTable, we made a series of design choices that come with limitations. Firstly, we opted for a simple table that, wherever possible, limits column contents to a single word/symbol. This design choice makes the table more readable. One does not have to parse dense blocks of text within the table to get information. Instead, one can quickly scan the

variTable horizontally and vertically to learn, broadly, about the variable development process. However, the prioritization of simplicity and minimalism also means the depth and specificity of the information presented in the table is limited. This limitation is most notable in the preprocessing section of the table, where a check mark could denote either a single preprocessing operation or a complex preprocessing pipeline.

Secondly, in the variTable we introduce novel categorizations of data source and preprocessing information. We categorize the role of a data source—*main* versus *support*—as well as the type of data preprocessing—*direct* versus *indirect*. These categorizations are used in the last three columns of the table (see Section 3). Because their meaning may not be immediately familiar to readers, a portion of the table is not immediately interpretable. Rather, readers will have to first learn what these columns mean to fully use the table. To help address this issue, we recommend the use of footnotes below the variTable to explain the meaning of these columns.

Additionally, as discussed in Section 3.1, we designed the table columns to be *variable-centric* and the rows to be *model-ordered*. This design choice was made so that key variables in a data analysis appear in the top-left corner of the table, where they will be considered first. However, in quantitative research, it is not uncommon for researchers to run multiple related analyses (and report multiple results). For example, researchers may fit several different models with different variables. Furthermore, the variables may be prepared differently depending on the model. The variTable, at present, is not designed to capture these different analyses and variable preparations, all at once. Therefore, paper authors will need to either choose which analysis to feature in the variTable or build multiple variTables.

Moving forward, there are several directions for future work. One direction is to explore visual encodings of the information in the variTable. The visual encodings could be designed to assist with comprehension of the table contents, pattern identification, or comparisons between variTables. Another direction for future work is the treatment of variTable contents as data. Currently, one of the `variTable` R functions puts the variTable content into a tibble (Müller and Wickham, 2025) in the tidy data format (Wickham, 2014) (see Figure 5). Using that data format, we are interested in exploring ways to visualize and analyze the variTable data, for insights into variable development. Another related area of interest is creating a summary visualization or metric that indicates the complexity of the variable development process.

To conclude, in original research articles, the communication of how quantitative research has been done is crucial for understanding, evaluating, and acting on research findings. Today, this communication task often involves the challenge of describing how many different existing datasets were transformed and combined into what was needed for the analysis. We hope that the variTable can assist with this particular communication challenge and serve as a practical tool for sharing and understanding important aspects of the variable development process.

Supplementary Material

The supplementary material for this article is available on the official website of the Journal of Data Science at https://jds-online.org/. It may include additional results, appendices, tables, figures, code, and data.

References

Arnold CK (1962). The construction of statistical tables. *IRE Transactions on Engineering Writing and Speech*, 5(1): 9–14. https://doi.org/10.1109/TEWS.1962.4322756

Blaikie N (2003). Social research and data analysis: Demystifying basic concepts. In: *Analyzing Quantitative Data*, 10–36. SAGE Publications Ltd.

Buehler MF (1977). Report construction: Tables. *IEEE Transactions on Professional Communication*, PC-20(1): 29–32. https://doi.org/10.1109/TPC.1977.6594173

Chmielinski KS, Newman S, Taylor M, Joseph J, Thomas K, ..., Qiu YC (2022). The dataset nutrition label (2nd gen): Leveraging context to mitigate harms in artificial intelligence. arXiv preprint: https://arxiv.org/abs/2201.03954.

Gastel B, Day RA (2022). *How to Write and Publish a Scientific Paper*. Greenwood. ninth edition.

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, ..., Crawford K (2021). Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92. https://doi.org/10.1145/3458723

Hayes-Larson E, Kezios KL, Mooney SJ, Lovasi G (2019). Who is in this study, anyway? Guidelines for a useful Table 1. *Journal of Clinical Epidemiology*, 114: 125–132. https://doi.org/10.1016/j.jclinepi.2019.06.011

Holland S, Hosny A, Newman S, Joseph J, Chmielinski K (2018). The dataset nutrition label: A framework to drive higher data quality standards. https://doi.org/10.48550/arXiv.1805.03677

Hopewell S, Chan AW, Collins GS, Hróbjartsson A, Moher D, ..., Boutron I (2025). CONSORT 2025 statement: Updated guideline for reporting randomised trials. *The Lancet*, 405(10489): 1633–1640. https://doi.org/10.1016/S0140-6736(25)00672-5

Jenkinson BL (1949). *Bureau of the Census Manual of Tabular Presentation*. United States Government Printing Office.

Kapoor S, Cantrell EM, Peng K, Pham TH, Bail CA, ..., Narayanan A (2024). REFORMS: Consensus-based recommendations for machine-learning-based science. *Science Advances*, 10(18): eadk3452. https://doi.org/10.1126/sciadv.adk3452

Keet CA, Keller JP, Peng RD (2018). Long-term coarse particulate matter exposure is associated with asthma among children in Medicaid. *American Journal of Respiratory and Critical Care Medicine*, 197(6): 737–746. https://doi.org/10.1164/rccm.201706-1267OC

Lucchesi LR, Peng RD (2026). variTable: Variable Development Tables. R package version 0.1.0, https://github.com/lydialucchesi/variTable.

Müller K, Wickham H (2025). tibble: Simple Data Frames. R package version 3.3.0. https://CRAN.R-project.org/package=tibble.

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. National Academies Press.

Nolan D, Stoudt S (2021). *Communicating with Data: The Art of Writing for Data Science*. Oxford University Press.

Peng RD (2011). Reproducible research in computational science. *Science*, 334(6060): 1226–1227. https://doi.org/10.1126/science.1213847

Posit Team (2025). *RStudio: Integrated Development Environment for R*. Posit Software, PBC, Boston, MA.

Rutter R (2017). *Web Typography*. Ampersand Type.

von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP (2008). The strengthening the reporting of observational studies in epidemiology (strobe) state-

ment: Guidelines for reporting observational studies. *Journal of Clinical Epidemiology*, 61(4): 344–349. https://doi.org/10.1016/j.jclinepi.2007.11.008

Wainer H (1992). Understanding graphs and tables. *Educational Researcher*, 21(1): 14–23. https://doi.org/10.3102/0013189X021001014

Wickham H (2014). Tidy data. *Journal of Statistical Software*, 59(10): 1–23. https://doi.org/10.18637/jss.v059.i10