

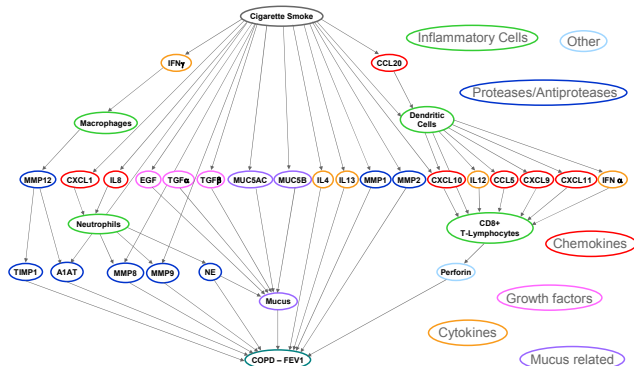
# New Method to Capture Context-Dependent Quantitative Data to Build a Data Warehouse for Chronic Obstructive Pulmonary Disease

Carole Mathis, Pascal Cosandier, Grégory Vuillaume, Peter Sperisen, Gerd Kallschnigg, Natasa Forte, Andrea Wohlsen, Christelle Haziza, Michael Peck, Rolf Weitkunat, and Zheng Sponsiello-Wang.  
Philip Morris International Research and Development, Philip Morris Products SA, Applied Science Department, Biostatistics & Epidemiology, CH-2000, Neuchâtel, Switzerland.

## Introduction

Chronic Obstructive Pulmonary Disease (COPD) is characterized by progressive airflow limitation associated with abnormal inflammatory response of the lung to noxious particles and gases. It is the fourth leading cause of death in the world. Comprehensive information related to COPD is available, but little is known about the disease mechanisms. In order to investigate disparate pieces of information, data must be collected, stored, and evaluated in a systematic way. Here, we present a **new method to capture context-dependent quantitative data** to build a data warehouse for COPD.

## Current COPD Biological Network



Network composed of the various links used so far for the literature search.

## Literature Search: Three Phases

**Query phase:** generation of a comprehensive list of articles from queries including all synonyms for all specified links of the defined biological pathways (see above) to search in PubMed. **Primary screening phase:** at the abstract level to determine the relevance of the articles to the specified links. **Secondary screening phase:** at the full-text article level and based on **inclusion/exclusion** criteria to select the most relevant publications. (Screening phases performed by biologists.)

### INCLUSION CRITERIA

- English peer-reviewed articles relevant for the different biological links defined in the COPD Bayesian model (direct acyclic graph with probabilistic distribution).
- Human, mouse, and rat studies.
- COPD patients with or without co-morbidities.
- Articles dealing with cigarette smoke (CS) & CS constituents inducing mechanistic modifications with quantitative data.

### EXCLUSION CRITERIA

- Articles dealing ONLY with immortalized cell lines.
- Articles dealing ONLY with cell-free assays.
- Case reports, reviews, editorials, comments, and letters.
- Epidemiological data and meta-analysis.
- Articles dealing ONLY with COPD patients with exacerbations AND/OR alpha-1 antitrypsin deficiency.

## Conclusion

We have set up a process to build a COPD data warehouse. The process **captures quantitative data and associated context information**, such as type of experiment (in vivo, in vitro), species used (human, mice, rat), characteristics of the study group (smoking status, demographic data, disease phenotype, etc.) and a scoring system (the Measurement Certainty Index) to evaluate the certainty of the data. Data from 600 research articles have been recorded to date. This mine of information is easily retrieved and provides us with a tool to identify data gaps and putative biomarkers. Based on these data, an in silico COPD Bayesian network model for disease risk prediction is being built. This approach can be also adapted to other research areas.

## Literature Data Transfer Template

General Information Sheet: Literature Data Transfer Sheet 4.2

<b>Reference Information</b>	Assessor: [Redacted]
<b>Information on article assessment</b>	Assessor Initials: [Redacted]
	Date of Assessment: [Redacted]
<b>Title, authors, PubMed ID</b>	Title: Plasma albumin-derived peptide levels in normal adults, children, and emphysematous subjects. Physiologic and pathologic implications.
	Authors: Olson TA, Vlahos R, Sankaranarayanan R, Edam E, Clark ES, Malmgren S.
	PubMed ID: 176363
<b>Objectives, study design, results, key words</b>	Objectives: The aim of the study was to measure the ECP levels in plasma by ELISA to confirm that pulmonary emphysema is characterized by lower elcristin breakdown.
	Study Design: An in-vivo observational study has been done in humans by considering normal non-smokers (n=39), normal smokers (n=46), normal informed smokers (n=41), emphysematous (n=45) and children below 10y of age (n=24) study subjects.
	Results: From COPD we were determined 4 subjects under study. A significant correlation was observed between elcristin breakdown measurements, CT scan percent emphysema score and plasma ECP levels.
	Key Words: Elcristin, Elcristin-derived peptide, ELISA, CT scan.
<b>Study design &amp; quality MCI score for the study as a whole</b>	Assessment Status: [Redacted]
	MCI: Study Design [0] 2   Experimental [0] (based on study design)   Statistical [0] (based on parameters)   Study Quality [0] 2   Study quality conditions, some major quality criteria failed.
<b>Abstract</b>	Abstract: Pulmonary emphysema is likely to be the result of elastic tissue degradation by proteolytic enzymes activity in the lung. Elastic breakdown by elastase results in the release of soluble elastic fragments (ECP). While the mechanism of action of ECP is unclear, plasma ECP levels measured using an ELISA were determined in the following group: disease-free children (n=24), 30-40 (n=46) light, disease-free adult non-smokers (n=39), 17-44 (n=41) light, normal smokers (n=46), 25-44 (n=45) light, informed smokers (n=41), 17-44 (n=41) light. Adults with emphysematous pulmonary emphysema (n=45) or children below 10 years of age (n=24) were included in the study. The aim of the study was to measure the plasma ECP levels in humans by ELISA to confirm that pulmonary emphysema is characterized by lower elcristin breakdown measurements. CT scan percent emphysema score. In addition, we measured the relationship of plasma ECP to these other indicators of pulmonary emphysema in a separate group of 20 subjects using elcristin breakdown measurements (ELISA) and another group of 10 subjects with CT scan percent emphysema score. A significant correlation (p < 0.05) was shown for plasma ECP levels and

**Dialog Box to Generate Tables:**

Tables Generation Dialog

Information Source: [Table 1] Species: [Human]

Main information describing the table:

- Information source
- Table name
- Species used
- Type of experiment

Header:

- Smoking/exposure status
- Groups within the smoking/exposure status
- Variables within the groups

Body Choices:

- Type of table to create
- Corresponding statistics options

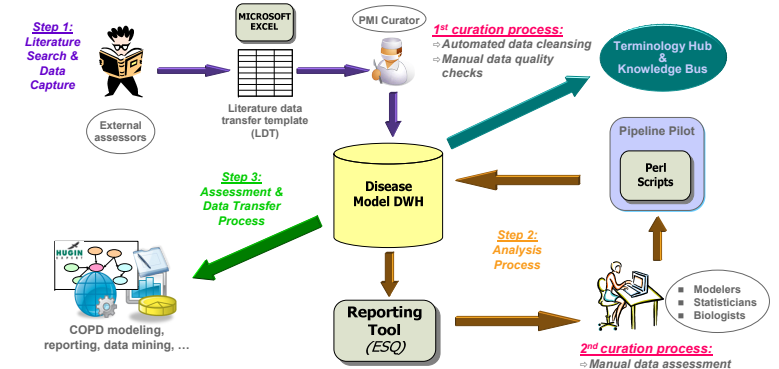
Group Statistics: [Row totals] [Aggregated data] [Regression] [Correlations]

OK Cancel

**Example of Raw Data Table:**

Information	Header	Raw Data	Body	MCI	Comments
Information	Header	Raw Data	Body	MCI	Comments
Header	Raw Data	Body	MCI	Comments	
Raw Data	Body	MCI	Comments		
Body	MCI	Comments			
MCI	Comments				
Comments					

## Data Curation & General Process

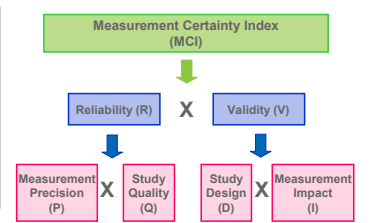


## Types of Quantitative Data

- Individual measurements** for individual subjects / experimental units and for one or several variables
- Aggregated measurements**
  - Group statistics data describing the groups or used as inclusion/exclusion criteria
  - Aggregated data for a group of individual subjects / experimental units and for one or several variables
- Regression equations** involving two or more variables  

$$y = ax + b; y = a_1x_1 + a_2x_2 + \dots + a_nx_n + b; y = a \cdot e^{bx}$$
- Correlation coefficients** between two variables

## Measurement Certainty



## COPD DWH in PMI Data Integration Platform

