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

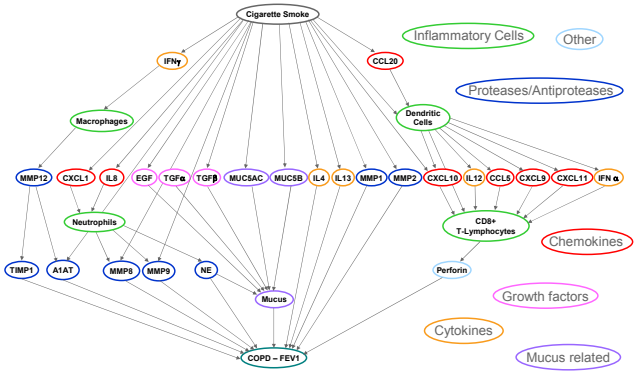
Carole Mathis, Pascal Cosandier, Grégory Vuillaume, Peter Sperisen, Gerd Kallischnigg, Natasa Forte, Andrea Wohlsen, Christelle Haziza, Michael Peck, Rolf Weitkunat, and Zheng Sponsiello-Wang.

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 CRITERI A

- 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

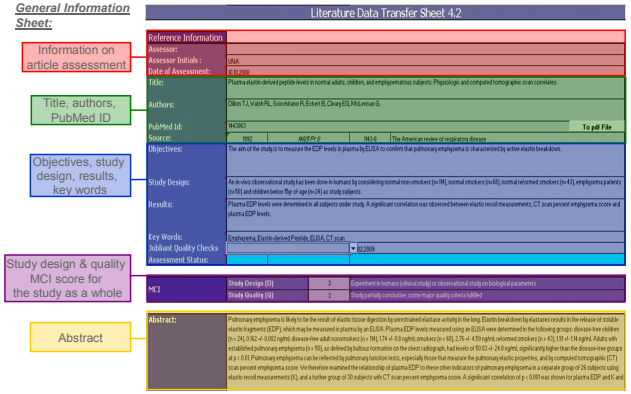
- Articles dealing ONLY with immortalized cell lines.
- Articles dealing ONLY with cell-free assays.
- Case reports, reviews, editorials, comments, and letters.
- Epidemiological data and meta-analysis.
- Articles dealing ONLY with COPD patients with exacerbations AND/OR alpha-1 anti-trypsin deficiency.

Conclusion

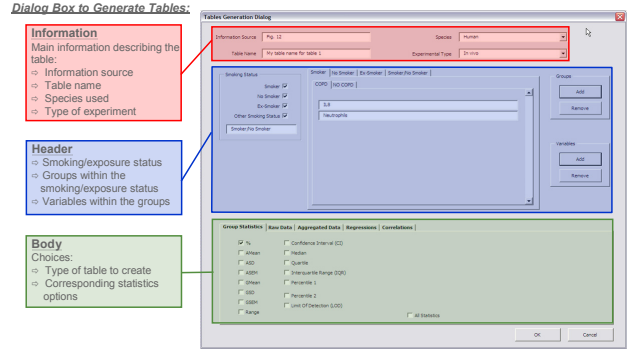
We have set up a process to build a COPD data warehouse. The process **captures quantitative data and associated context information**, such as type of experiment (in vivo, in vitro), species used (human, mice, rat), characteristics of the study group (smoking status, demographic data, disease phenotype, 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

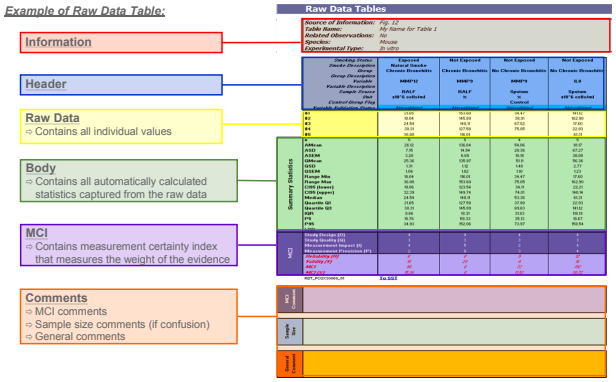
General Information Sheet:



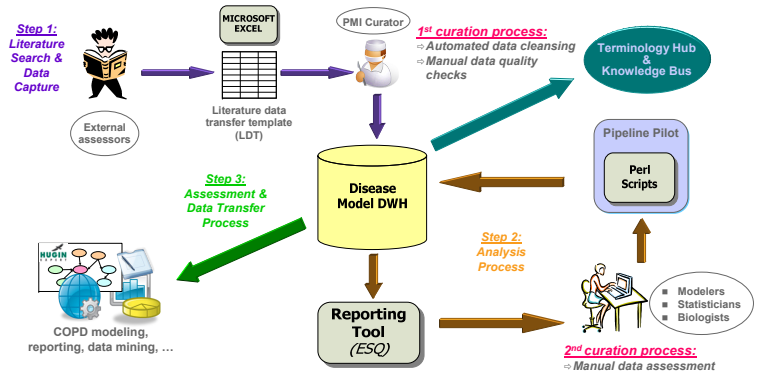
Dialog Box to Generate Tables:



Example of Raw Data Table:



Data Curation & General Process



Types of Quantitative Data

Individual measurements for individual subjects / experimental units and for one or several variables

Aggregated measurements

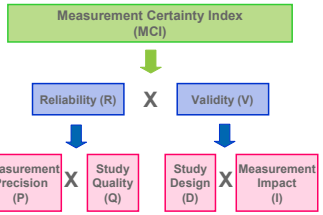
- **Group statistics data** describing the groups or used as inclusion/exclusion criteria
- **Aggregated data** for a group of individual subjects / experimental units and for one or several variables

Regression equations involving two or more variables

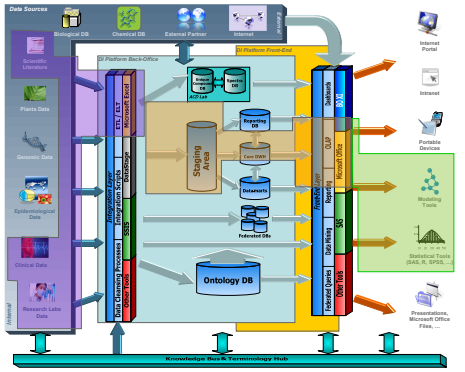
$(y = \alpha x + \beta;\ y = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_m x_m + \beta;\ y = \alpha \cdot e^{\beta x})$

Correlation coefficients between two variables

Measurement Certainty



COPD DWH in PMI Data Integration Platform



8th Annual International Conference on Computational Systems Bioinformatics (CSB)
Stanford University, CA
10-13 August 2009

PMI RESEARCH & DEVELOPMENT

Philip Morris International Research & Development, Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland
T: +41 58 242 21 11, **F**: +41 58 242 28 11, **W**: www.philipmorrisinternational.com