Unlocking Valuable Clinical Trial Study Data with AI Data Extraction

Bayer Consumer Health and Capgemini Engineering with Acodis Case Study

With over 170 consumer health brands in its innovative global portfolio, Bayer Consumer Health empowers people to manage their health needs in the areas of dermatology, nutritional supplements, pain management, cardiovascular risk prevention, digestive health, cough, cold, and allergy care.

Capgemini engineering is a global leader in partnering with companies to transform and manage their business by harnessing the power of technology.It is a responsible and diverse organization of nearly 350,000 team members in more than 50 countries.

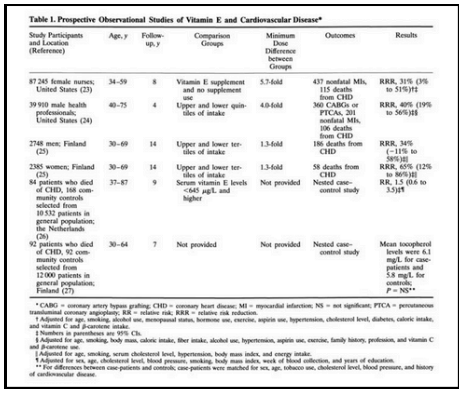
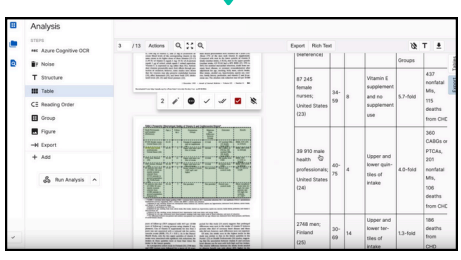


Table 1. Prospective Observational Studies of Vitamin E and Cardiovascular Disease*

Study Participants and Location (Reference)	Age, y	Follow-up, y	Comparison Groups	Minimum Dose Difference between Groups	Outcomes	Results
87 245 female nurses; United States (23)	34–59	8	Vitamin E supplement and no supplement use	5.7-fold	437 nonfatal MIs, 115 deaths from CHD	RRR, 31% (3% to 51%)†‡
39 910 male health professionals; United States (24)	40–75	4	Upper and lower quintiles of intake	4.0-fold	360 CABGs or PTCAs, 201 nonfatal MIs, 106 deaths from CHD	RRR, 40% (19% to 56%)‡§
2748 men; Finland (25)	30–69	14	Upper and lower tertiles of intake	1.3-fold	186 deaths from CHD	RRR, 34% (−11% to 58%)‡§
2385 women; Finland (25)	30–69	14	Upper and lower tertiles of intake	1.3-fold	58 deaths from CHD	RRR, 65% (12% to 86%)‡§
84 patients who died of CHD, 168 community controls selected from 10 532 patients in general population; the Netherlands (26)	37–87	9	Serum vitamin E levels <645 μg/L and higher	Not provided	Nested case–control study	RR, 1.5 (0.6 to 3.3)‖¶
92 patients who died of CHD, 92 community controls selected from 12 000 patients in general population; Finland (27)	30–64	7	Not provided	Not provided	Nested case–control study	Mean tocopherol levels were 6.1 mg/L for case-patients and 5.8 mg/L for controls; $P = NS$**

* CABG = coronary artery bypass grafting; CHD = coronary heart disease; MI = myocardial infarction; NS = not significant; PTCA = percutaneous transluminal coronary angioplasty; RR = relative risk; RRR = relative risk reduction.
† Adjusted for age, smoking, alcohol use, menopausal status, hormone use, exercise, aspirin use, hypertension, cholesterol level, diabetes, caloric intake, and vitamin C and β-carotene intake.
‡ Numbers in parentheses are 95% CIs.
§ Adjusted for age, smoking, body mass, caloric intake, fiber intake, alcohol use, hypertension, aspirin use, exercise, family history, profession, and vitamin C and β-carotene use.
‖ Adjusted for age, smoking, serum cholesterol level, hypertension, body mass index, and energy intake.
¶ Adjusted for sex, age, cholesterol level, blood pressure, smoking, body mass index, week of blood collection, and years of education.
** For differences between case-patients and controls; case-patients were matched for sex, age, tobacco use, cholesterol level, blood pressure, and history of cardiovascular disease.



Example of table recognition and extraction from publicly available comparable clinical data scanned pdf document.

Task

Extract the raw data from clinical study reports from historic studies and map multiple studies to a common data model to perform meta-analysis across several studies.

Scope

Data extraction of over 200 tables, which ranged in formatting and length in over 3500 pages of documents and data validation for a finished common data model in 3 weeks.

Key Benefits

Scalability, flexibility, and speed. No coding is required, data extraction is done by a subject-matter expert.

The Challenge

- Conducting clinical trial studies is a costly process for pharmaceutical companies, requiring extensive effort, time, and resources. Costly patient recruitment, screening and retention, research and development, and regulatory requirements contribute to the high expense of conducting clinical trials.
- Over-the-counter (OTC) consumer health products are often based on data generated years or even decades ago. In many cases, the raw data from these older studies in electronic format has been lost, and the clinical data are only accessible through scanned copies of the listings in the appendices of integrated study reports.
- Bayer Consumer Health and Capgemini Engineering aimed to **extract the raw data from clinical study reports from historic studies**. While the clinical trial studies were conducted many years prior, the studies held valuable key insights, such as patient demographics, drug exposures, and outcomes, that could be used for analytical purposes. In addition, mapping multiple studies to a common data model allowed to pool together studies to perform meta-analysis across several studies.
- The team faced a challenge, as much of the clinical trial data was stored in unstructured data formats. The data was not readily available or accessible for analysis and required extraction from old, **scanned copies of PDF documents**. The team required a solution that allowed for high-quality extraction from multiple different tabular formats. In addition, some tables had differing (or lack of) border and row structuring, and/ or a high level of skewness due to the scanning. Furthermore, some scanned PDF files were of poor quality, with noisy data.

" Acodis allowed for the data to be extracted without the use of manual hard coding each page (>3500)." ★★★★★

Dr. Shannon Montgomery, Data Scientist, Capgemini Engineering

"By extracting raw data from scanned PDF copies from historic study reports, we were able to re-use the clinical data and conduct further exploration and analysis by combining datasets from various studies. Without the new technology this would not have been possible within reasonable costs and timelines"

Gregor Bieri, Head Clinical Data Sciences, Analytics & QM at Bayer Consumer Health

Solution

- While identifying multiple available options that can be used to extract data, the challenge remained that data could not be extracted from the unstructured formats of the PDF files within this case study. In addition, the team faced limitations with the scalability and flexibility of a solution that allowed for consistent data extraction of over 200 tables, which ranged in formatting and length.

- The team partnered with Acodis to use the data extraction solution, which allowed for the AI supported **identification** of data from each table. This solution provided **flexibility** across the ranging data formats and ensured that the data extracted from the PDF files was of the best available quality. As the data extraction tool allowed for identification of rows and columns, without the need for hard coding, it allowed for the extraction to be **driven by a subject-matter expect**, with no reliance on technical resources. This provided an additional benefit as it allowed for flexibility in the use of the tool and reduced architectural limitations.

- Furthermore, the Acodis table recognition solution improved the **scalability** and **speed** of the high volume of data to be extracted, by recognizing tables that had a similar format, and reduced the level of fine-tuning required. The solution became familiar with the format, recognizing the columns, rows, and cells and reducing the manual effort required.

Outcome

With the Acodis solution, the team extracted data from over

> 200	> 3500	3
complex tables	pages of documents	weeks

This resulted in the ability to clean, map, and standardize raw clinical data. As a result, the team could analyze a

high volume of data

that was previously inaccessible and provided additional business insights, helping to reduce the need for conducting additional clinical trials.