WHITE PAPER

WHY WEARING GLASSES DOES NOT CAUSE GRAY HAIR!

FROM CORRELATION TO CAUSATION TO ARTIFICIAL INTELLIGENCE …

Xplain Data GmbH
Grünlandstr. 27
85604 Zorneding
Germany

info@xplain-data.com
xplain-data.com

WHAT TO EXPECT

SUMMARY

Correlation does not imply causation. Nevertheless, we are often making decisions based on correlation instead of “cause and effect”. Cause and effect, unfortunately, cannot be proven based on observational data. We may, however, obtain some evidence on causality: Direct causal factors cannot be “explained away”, and an intense search for alternative explanations reveals a small set of direct and potentially causal factors. Holistic data is therefore important for causal discovery.

Interventions which aim to drive a system towards a desired goal require knowledge about cause and effect. Causality will therefore be an important pillar for future systems of Artificial Intelligence. Precision medicine and individualized treatments are hardly conceivable without that. We show an example where we predict depressive episodes thereby revealing effects and side-effects of certain drug categories, and how those affect different patient groups.

GLASSES & GRAY HAIR

Sit down in a coffee bar, look at the people passing by and count how many of them are wearing glasses and how many have gray hair …

CORRELATION VS CAUSATION

You will find that there is a correlation between "wearing glasses" and "gray hair": Amongst those passers-by with glasses an increased fraction also has gray hair.

Are you wearing glasses?

Don't worry! You don't need to put your glasses aside to avoid your hair becoming gray. There is a correlation between "wearing glasses" and "gray hair", but glasses do not cause your hair to become gray.

The pattern "gray hair and glasses" is a very familiar one. Our brain is very good in pattern recognition, and – intuitively and correctly – we do not draw the conclusion "glasses are causing gray hair". There is a simple reason for that, as we will see later (you likely know it already). When analyzing data, however, people often draw false conclusions on causality based on correlations. Here are just some examples:

- Your marketing efforts at target customers might correlate positively with the amount sold at those customers. Does that mean that your marketing efforts caused increased sales? Marketing managers tend to assess the efficiency of a campaign based on such a correlation.
- Machine failures might correlate to certain operating parameters prior to failure – does it mean that those operating conditions cause failures?
- Expressed genes might be correlated to cancer – does that mean that those genes expressions cause cancer?

Apparently "correlation" and "causation" are two different things. This leaves us with a desperate question …

WHAT IS CAUSATION?

There is a precise mathematical definition for correlation, e.g., in terms of Pearson's correlation coefficient. Is there a similar precise definition for "causation" based on observational data? How can we measure "causation"?

Indeed, we can measure causation in a similar rigorous way – but only if we are allowed to do experiments. "Randomized controlled trials (RCTs)" are the gold standard how to assess causal effects. RCTs are widely used, e.g. in drug development where patients are randomly assigned to a control and intervention group.

If we can do randomized experiments, we are fine. Very often, however, this is not feasible. For example, to prove that glasses are not causing gray hair, we would need to grab people passing our coffee bar, randomly advise them to not wear or wear glasses henceforth, and later observe the effect in the two different groups. Practically, that is not feasible.

Experiments are rarely possible in real world settings, in particular not randomized experiments. In today's digital world data are piling up, and largely all of them are observational data (whereby observational means no controlled/randomized interventions/experiments). The crucial question therefore is: How can we define, or measure causation based on observational data?

BAD NEWS

There is bad news: We cannot measure causation based on purely observational data. Indeed, in case of purely observational data, we might even not be able to define the term "causation" in a rigorous mathematical way [1] .

How disappointing! We are in the ages of Big Data, and there is no way to use those data to understand "cause and effect"? Can't we get at least an estimate or bounds on causality? Can we exclude certain correlations to be non-causal, thus that we end up with a list of very hot candidates for potentially causal relationships?

If we can get only more or less good evidence on causality, then the next important question is: How do we need to collect data to obtain best possible insights on causality? Often tons of data are collected, only to later learn that a crucial little piece is missing.

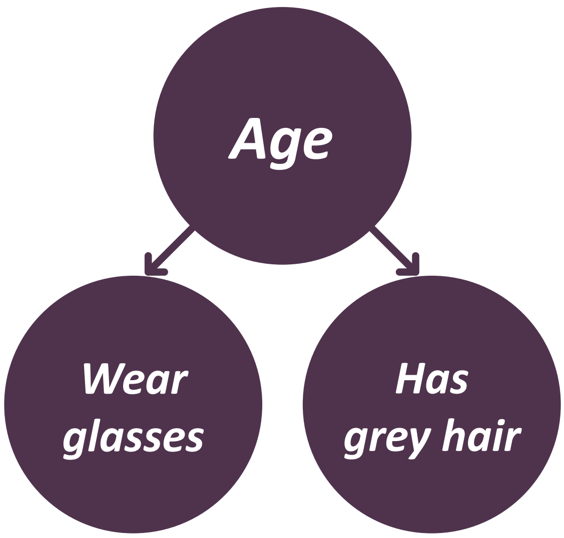
CAUSALITY – "NO OTHER AVAILABLE EXPLANATION"

According to Kenny[2] three conditions must be satisfied to infer that variable (X) has a causal influence on variable (Y):

1. X must precede Y temporally
2. X must be reliably correlated with Y (beyond chance)
3. Alternative explanations for the relationship between X and Y must be ruled out.

How does this help us in our glasses and gray hair example? Condition 3. means that, if we can find an alternative explanation for the correlation between glasses and gray hair, then this rules out glasses as a cause for gray hair. In other words, we need to find a better explanation or the "true cause" for gray hair, which makes glasses obsolete as the explanation for gray hair.

As you might have guessed already, the true cause for gray hair is "age". When we get older, eventually our hair is becoming gray and many of us need to wear glasses. Age is a common cause for both, for the need of glasses and hair becoming gray (a so-called confounder).



[1] Some researchers might object, and, indeed, with several very theoretical premises it might be possible to identify a "structural causal model". Practically, however, we must state that we cannot prove causal effects based on purely observational data – randomized controlled trials remain the gold standard.

[2] D. A. Kenny: Correlation and Causality, New York: John Wiley & Sons, 1979

In theory we could prove that by asking all passers-by about their age and then evaluate the correlation between glasses and gray hair within each age category. If there is no correlation in any of the age groups, then age has completely "explained away" glasses as a direct causal factor for gray hair.

Vice versa – if we would unexpectedly find that within passers-by of same age there is still a correlation between glasses and gray hair, then age cannot be the only reason for gray hair. Beyond age, there need to be other factors existing, and the factor glasses would still be amongst the potential candidates causing gray hair.

In that case, we would need to keep on searching for other factors (other confounders). Let's ask all the passers-by for lots of additional data, their gender, their nutrition and smoking habits in recent years, their sports activities ... if nothing of all the available information explains away glasses as a factor for gray hair – indeed then we would need to consider the possibility that glasses are causing gray hair.

If a factor cannot be explained away by any other available information (temporarily preceding the target) then this factor is becoming a likely candidate as a causal factor, the more other information is available the more likely it will be. In that sense we might define "observational causality" as "no other available explanation within given scope of information" [3] .

Lucky!? This sounds like "lots of information helps a lot". Indeed, while the bad news is that cause and effect cannot be proven based on observational data, the good news is that "Big Data" (in the sense of diverse data) might at least help to rule out many hypotheses and quickly narrow down to the interesting ones.

[3] This is still no mathematical definition and likely also differs to potential other "definitions" of causality. Indeed, the herein used definition of causality or that of Kenny (1979) refers to "direct" causes only (i.e. not including variables which have an indirect effect "via others").

A PRACTICAL APPROACH TO CAUSAL DISCOVERY

A HOLISTIC VIEW IS MANDATORY

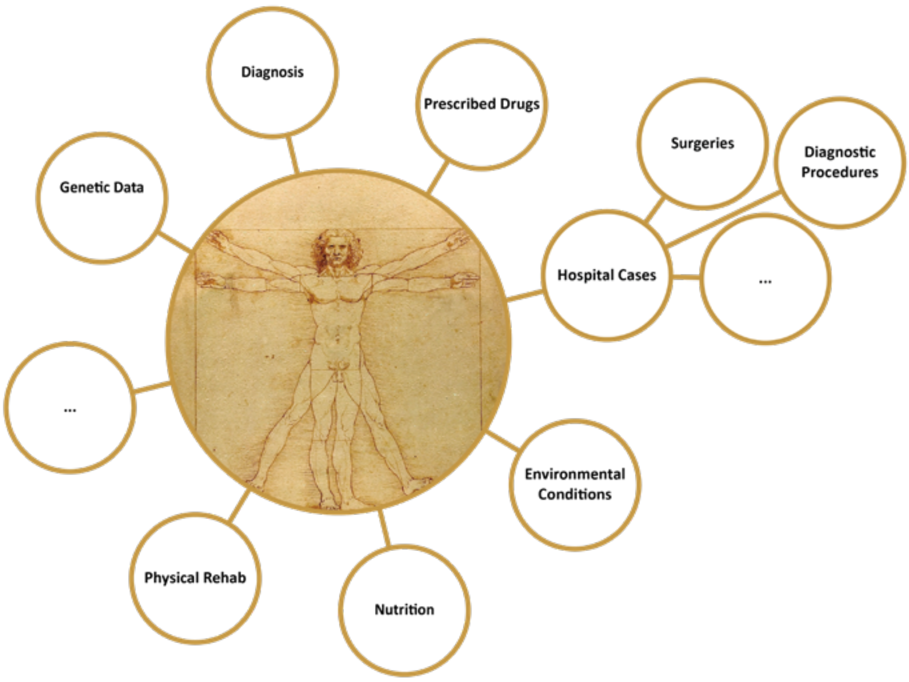
We have learned that a factor is a good candidate for a potentially causal factor if it cannot be "explained away" by any other variable. The more diverse data we have, the more intense we can search for "alternative explanations" (search for confounders). An approach to causality therefore requires holistic data.

Holistic data, in turn, means a complex data model. Think of data which represents patient lives. This data will necessarily require a model of multiple tables and related entities – by no means just one flat table. Most statistical approaches, however, assume a set of N variables, i.e., a flat table with N columns, and each row representing an observation (one passer-by in our coffee bar example with variables such as "wearing glasses" or "age"). In real world data models, however, there is nothing like this existing a priori. Complex data first need to be mapped into a flat table, which today still is a manual, assumption-biased process outside the scope of statistical learning algorithms.

"Object Analytics" bridges this gap. It refers to the representation of data as holistic objects instead of rows and tables and facilitates analysis of objects (e.g. "Patients") as a whole. Specificaly for causal discovery this means that we can search an entire object with all its dependent data for "alternative explanations". Object Analytics thus offers novel opportunities to understand potential cause and effect relationships.

AND DOMAIN EXPERTS NEED TIE-IN

Based on comprehensive data, we can rule out many correlated factors to be non-causal and thereby obtain evidence that others are likely candidates for a very direct and potentially causal dependency. But even if we have a plethora of observational data, causality can still not be proven. A practical approach which assesses causal effects therefore needs to incorporate expert feedback.



Example of a complex, real-world object: "The patient" with various data streams attached to it. "Object analytics" means being able to analyze the different data areas in relation to each other.

Those experts need to be able to intuitively understand results. A primary goal of our approach therefore is to summarize the dependency of a target event on prior events in terms of a model as simple as possible with parameters intuitive to comprehend. A network of so-called “probabilistic OR- and AND-gates” serves this purpose.

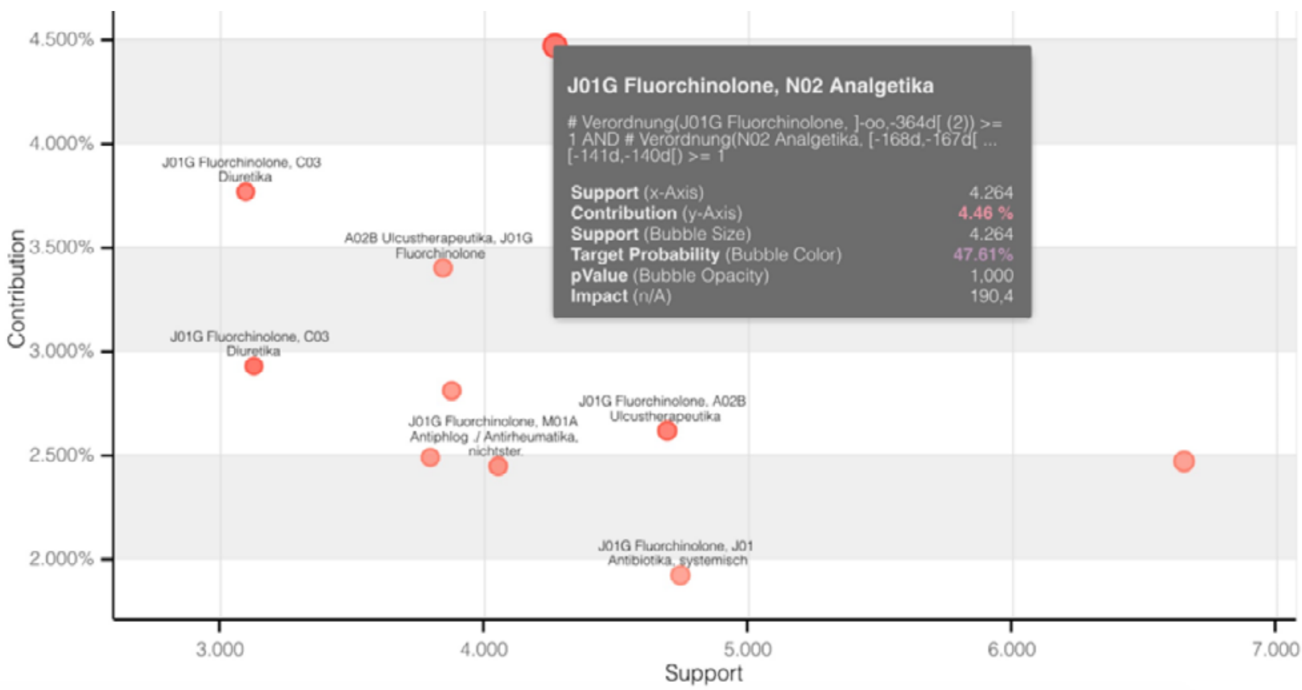
The overall approach is designed as a close interaction with experts. Initially, the typical situation is that the view to the essentials is obscured by myriads of meaningless correlations. The fog lifts quickly by building a first model which boils down the many correlations to a small set of direct (potentially causal) factors. The expert may then assess factors by rejecting some and marking others which he believes are cause and effect in nature and which also represent potential interventions. (Age might cause gray hair, but we cannot intervene on age. However, if smoking causes gray hair, we can stop smoking to avoid gray hair.) With expert feedback included detailed models are re-built which then can be deployed, e.g., as elements in an intelligent business process.

AN EXAMPLE

Predicting health risks is important – understanding causes for health risks is even more important. Knowledge about cause and effect is the basis for targeted interventions. The promises of “precision medicine” will materialize only once we understand cause and effect of drugs and treatments in individual patient groups (including side effects).

In the example in the below figure, we have predicted the risk for a depressive episode based on a patient’s prior health history (drugs and treatments prior to the first administration of an anti-depressive drug). Data of 3 million patients have been used with ~200 million prescriptions for thousands of different drug categories.

The algorithm searches the object model (all data streams attached to the object “Patient”) for factors related but prior to first anti-depressive treatment. While searching the object tree, 100.000s of hypotheses are formed. Each hypothesis for a potential causal factor is tested against all the others to evaluate whether any of the other factors may server as “alternative



Identified factors directly related to later anti-depressive treatment depicted in a bubble chart: x-axis: number of patients with this factor, y-axis: contribution to risk. Only factors related to fluoroquinolones are shown.

[4] See for example: www.akdae.de/Arzneimittelsicherheit/Bekanntgaben/Archiv/2004/20040528.html

explanations" (as confounders). Only factors which survive the extensive search (i.e., correlation to target cannot be explained via other factors) are presented to the expert. Depending on depth of search ~50 to 100 factors remain as "cannot be explained away" and potentially causal contributions to the risk of depression.

Interestingly different sorts of antibiotics appear as risk drivers, in particular so-called fluoroquinolones. They have very different risk contribution in different combinations with other drugs resp. for different groups of patients. Mood swings and depression are indeed a known side effect of fluoroquinolones[4]. Those side-effects might, however, be considerably underestimated, and they seem to manifest very differently in different patient groups. An individualized decision on treatment with fluoroquinolones therefore seems compulsory.

APPLICATIONS OF CAUSALITY: INTELLIGENT MEASUREMENTS, INTELLIGENT INTERVENTIONS

The world is full of talk about "Artificial Intelligence". No one talks about "Causality". Wouldn't intelligent interventions require the performed actions to be cause-and-effect-related to the target?

Evidently, being able to find likely causal relationships offers numerous opportunities. The above "precision medicine" example is just one of them. But we don't need to go that far and fancy an AI system that autonomously finds best actions based on cause and effect. There are much more nearby applications of causality, for example an adjusted measurement of diverse types of performed interventions.

Assume, as an example, there are many customers, and you are targeting those with diverse types of actions. Any measurement of the effect of different types of actions should consider causality. You don't want to credit results to an action which have not been caused by that action. Adjusted measurement is an obvious application of concepts for causality. Based on that you can then start allocating resources where they have the most effect.

We are only just at the beginning to understand causality based on real-world data. The products of Xplain Data introduce viable concepts for causality into the domain of Artificial Intelligence.

We offer algorithms that may process complex objects "as they are" and live in real world instead of an artificially prepared analytics environment. Causality will soon become an important pillar for Artificial Intelligence.

Xplain Data
Discover Causality

Xplain Data GmbH
Grünlandstr. 27
85604 Zorneding
Germany

info@xplain-data.com
xplain-data.com