



Deliverable report

NOVA (WP2)

D-2.3.2. (D-JRP6-2.4)

**Simulation analyses to
assess the use of
consumer purchase data
as an analytical tool for
outbreak investigations**

JRP6 - NOVA - FBZ1 - 1st Call

Responsible Partner: SSI

Contributing partners: FHI



GENERAL INFORMATION

European Joint Programme full title	Promoting One Health in Europe through joint actions on foodborne zoonoses, antimicrobial resistance and emerging microbiological hazards NOVA
European Joint Programme acronym	One Health EJP
Funding	This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.
Grant Agreement	Grant agreement n° 773830
Start Date	01/01/2018
Duration	60 months (OH-EJP) 36+6 months (NOVA)

DOCUMENT MANAGEMENT

Title of this OHEJP (NOVA) deliverable	Simulation analyses to assess the use of consumer purchase data as an analytical tool for outbreak investigations.
WP and task	JRP6-WP2-T3 NOVA-WP2: Analysis of food purchase data. Task 3: Big data analysis of risk factors for sporadic disease
Leader	SSI (DK)
Other contributors	FHI (NO)
Due month of the deliverable	36
Actual submission month	42 (project extended with 6 month due to covid-19)
Type <i>R: Document, report</i> <i>DEC: Websites, patent filings, videos, etc.; OTHER</i>	R, DEC, other Save date: 10-Sep-21
Dissemination level <i>PU: Public (default)</i> <i>CO: confidential, only for members of the consortium (including the Commission Services).</i>	PU See updated Grant Agreement



<p>Dissemination</p> <p><i>Author's suggestion to inform the following possible interested parties.</i></p>	<p>OHEJP WP 1 <input type="checkbox"/> OHEJP WP 2 <input type="checkbox"/> OHEJP WP 3 <input type="checkbox"/></p> <p>OHEJP WP 4 <input type="checkbox"/> OHEJP WP 5 <input type="checkbox"/> OHEJP WP 6 <input type="checkbox"/></p> <p>OHEJP WP 7 <input type="checkbox"/> Project Management Team <input type="checkbox"/></p> <p>Communication Team <input checked="" type="checkbox"/> Scientific Steering Board <input type="checkbox"/></p> <p>National Stakeholders/Program Owners Committee <input type="checkbox"/></p> <p>EFSA <input checked="" type="checkbox"/> ECDC <input checked="" type="checkbox"/> EEA <input type="checkbox"/> EMA <input type="checkbox"/> FAO <input type="checkbox"/> WHO <input type="checkbox"/> OIE <input type="checkbox"/></p> <p>Other international stakeholder(s):</p> <p>.....</p> <p>Social Media:</p> <p><u>Other recipient(s):</u></p>
--	---



SIMULATION ANALYSES TO ASSESS THE USE OF CONSUMER PURCHASE DATA AS AN ANALYTICAL TOOL FOR OUTBREAK INVESTIGATIONS

0. Please note

Due to the COVID-19 epidemic, work on this project was paused for a significant amount of time and did not finish within the NOVA project period. However, the work continues and will be concluded in the coming months. The work and its results will be published, a manuscript is close to completion. This deliverable therefore only briefly describes the status of the project.

1. Introduction

Background

Foodborne outbreak investigations with a widely distributed food source generally rely on interviews with cases. This is a difficult and work consuming process, which often is not successful. Challenges often include lack of time and resources to be able to reach a sufficient number of cases, do the interviews within a reasonable time frame (to reduce incomplete recall) and secure samples of suspected outbreak sources. When successful, a pattern within the many types of food, people report to have consumed, may be recognised, allowing outbreak investigators to establish a hypothesis. If a hypothesis about a food source is reached, based in these initial interviews, an analytical study is the next step. This too is an expensive and time consuming process. For this reason many outbreaks are not investigated at all, or as well as desired. To be able to conduct outbreak investigations by tapping into electronic records of supermarket purchases of cases and background population is therefore hugely desired.

The level of electronic payment of grocery purchases in Scandinavian (and in many countries elsewhere in Europe) is by now very high with almost all transactions being digital. Thus, purchases of individual persons may be identified based on loyalty card programs, or directly by their payment cards (in agreement with the individuals).



The aim of the work presented here has been to explore the possibilities that lies in using large datasets of consumer purchases from supermarket chains, to conduct analyses for locating foods that are likely vehicles of outbreaks based on sales associations with known cases. We've sought to obtain real-life datasets and conduct simulation analyses based hereon.

2. Methods

In this project, the project group has aimed to obtain large real-life consumer dataset from Scandinavian supermarket chains. Initially work was done, using a very large anonymised dataset which was provided from a major Danish supermarket chain to the SSI. Following the introduction of the EU GDPR regulations, however, this work had to stop halfway into the project. Regulations were not breached, but in the climate that existed at the time (with some uncertainty as to how GDPR rules should be understood in detail) our supermarket partner though it best to not run any risks and asked us to stop working with the data. This was unfortunate in project terms, but we're nevertheless very thankful for this cooperation and the initial insights it gave us. Following a year of negotiations by the FHI, a second large anonymized consumer purchase dataset was offered to our use by a group of Norwegian supermarket chains. It contains individual weekly purchase histories available for analysis (920,834 unique customers from six Norwegian municipalities in 2019, totalling 222 million purchases of 30,929 different products).

The studies have aimed to use this consumer purchase dataset to simulate outbreaks, identify outbreak vehicles using automated regression methods, and examine how performance of selected analytic strategies varied across outbreak parameters.

3. Results & Discussion

The results will be presented in full elsewhere. In brief, by use of simulations, we generated outbreaks with different attack rates, proportion of background cases (where consumer purchase data (CPD) were not available), and food expiry dates. Analyses was conducted to evaluate the effect of item purchase frequency. The proportion of background cases (*i.e.* those with no available consumer data), was the most important parameter for the number of cases needed to point towards a possible source of the outbreak. Sensitivity analyses were also conducted to address different attack rates (0.01-0.1), expiry length (0-300 days) and



incubation time (0-10 days). Not surprisingly, the higher the item detail level (primary product group e.g. 'diary', secondary product group e.g. 'milk', and GTIN level: product ID, and brand specific product), the less cases were needed before it was possible to point towards an outbreak source. Thus, a logistic regression method can identify outbreak vehicles in a real-world CPD, and is, until further methods are explored, a good consensus candidate. Also, analysis of outbreaks using product groups with a granularity that is comparable to most trawling questionnaires, reduces performance compared to outbreaks analysed on item level. The study also indicated that CPD analysis requires participation from all major grocery chains for optimal performance.

The results of the analysis was presented as an oral presentation at the OH_EJP annual meeting in May 2021 in Copenhagen: "Dougherty PE et al: *Simulation and identification of foodborne outbreaks in real consumer purchase data*".

As manuscript exist in draft and is being finalised for submission, expected to happen November 2021: "Dougherty PE, Møller FT, Rø GØI, Ethelberg S and Jore S: Use of supermarket purchase data for foodborne outbreak source identification, a simulation study using large sized real life data".