



# Large-scale loyalty card data in health research

Jaakko Nevalainen<sup>1</sup> , Maijaliisa Erkkola<sup>2</sup>, Hannu Saarijärvi<sup>3</sup>,  
Turkka Näppilä<sup>1</sup>  and Mikael Fogelholm<sup>2</sup>

Digital Health  
Volume 4: 1–10  
© The Author(s) 2018  
Article reuse guidelines:  
sagepub.com/journals-  
permissions  
DOI: 10.1177/2055207618816898  
journals.sagepub.com/home/dhj



## Abstract

**Objective:** To study the characteristics of large-scale loyalty card data obtained in Finland, and to evaluate their potential and challenges in health research.

**Methods:** We contacted the holders of a certain loyalty card living in a specific region in Finland via email, and requested their electronic informed consent to obtain their basic background characteristics and grocery expenditure data from 2016 for health research purposes. Non-participation and the characteristics and expenditure of the participants were mainly analysed using summary statistics and figures.

**Results:** The data on expenditure came from 14,595 (5.6% of those contacted) consenting loyalty card holders. A total of 68.5% of the participants were women, with an average age of 46 years. Women and residents of Helsinki were more likely to participate. Both young and old participants were underrepresented in the sample. We observed that annual expenditure represented roughly two-thirds of the nationally estimated annual averages. Customers and personnel differed in their characteristics and expenditure, but not so much in their most frequently bought items.

**Conclusions:** Loyalty card data from a major retailer enabled us to reach a large, heterogeneous sample with fewer resources than conventional surveys of the same magnitude. The potential of the data was great because of their size, coverage, objectivity, and long periods of dynamic data collection, which enables timely investigations. The challenges included bias due to non-participation, purchases in other stores, the level of detail in product grouping, and the knowledge gaps in what is being consumed and by whom. Loyalty card data are an underutilised resource in research, and could be used not only in retailers' activities, but also for societal benefit.

## Keywords

Health behaviour, loyalty card data, food expenditure, participation, purchases, sales data, SWOT

Submission date: 23 May 2018; Acceptance date: 6 November 2018

## Introduction

Evidence of the relationship between health behaviour and chronic diseases is well established.<sup>1</sup> The risk factors of diseases – unhealthy diet, smoking and alcohol consumption – are an inequality concern, as they are clustered in the population.<sup>2</sup> Therefore, it is essential to reach those at greatest risk of developing chronic conditions such as obesity, type 2 diabetes and cardiovascular diseases. Population health behaviour monitoring and evaluation play a critical role in understanding and addressing these challenges. It is imperative to have relevant, reliable and timely information on the risk factors derived from valid, yet cost-effective, sources.

Of all health behaviours, diet might be regarded as the most challenging to assess, due to its substantial variability in quantity and quality. Dietary data

<sup>1</sup>Health Sciences/Faculty of Social Sciences, University of Tampere, Tampere, Finland

<sup>2</sup>Department of Food and Nutrition, University of Helsinki, Helsinki, Finland

<sup>3</sup>Business Studies/Faculty of Management, University of Tampere, Tampere, Finland

### Corresponding author:

Professor Jaakko Nevalainen, Health Sciences/Faculty of Social Sciences, Arvo Ylpön katu 34, FI - 33014 University of Tampere, Finland.  
Email: jaakko.nevalainen@uta.fi



collection instruments, such as food frequency questionnaires and food diaries, are associated with large random variation and systematic biases because of imprecision in recall and reporting, as well as with tendencies to report healthier dietary habits than those actually practised.<sup>3,4</sup> It is also known that population surveys typically underestimate alcohol consumption.<sup>5</sup> Many dietary assessment methods (e.g. food records and diet interviews) are limited to short periods of measurement, which makes their validity for assessing long-term health behaviour poor. The strengths and the weaknesses of the methods are comprehensively presented elsewhere.<sup>6,7</sup>

Health-related surveys have generally indicated declining trends in response rates.<sup>8</sup> More specifically, the Finnish Drinking Habits Survey reported a response rate of 60% in 2016 (97% in 1968), and US data from seven National Alcohol Surveys showed a response rate of 52% in 2010 (77% in 1995).<sup>9,10</sup> The National Findiet Survey on food consumption also followed the same long-term trend of decreasing participation rates: 60–63% in 1982 and 2007, and 57% in 2012.<sup>11–13</sup>

Interest in overcoming these challenges by alternative means of data collection has grown.<sup>14</sup> The field of digital epidemiology studies how big data – social media data, wearable devices, Twitter data and GPS tracking, for example – can be used to effectively address public health problems.<sup>15–17</sup> However, innovative technologies do not necessarily overcome the inherent individual bias related to self-reported dietary and alcohol intake. Many measurements (e.g. using wearable devices) provide high-resolution data on a specific individual, but information on the population is limited, since this kind of data collection is resource intensive. On the other hand, instruments that provide data on many individuals tend to provide little information on the individual level and/or lack relevant structure (e.g. social media data). Ideally, alternative data sources should be sufficiently ‘big’ in terms of both the number of individuals and the amount of measurements, and importantly, provide unbiased details on these individuals.

Loyalty card data potentially possess these features. By loyalty cards we mean electronic customer cards used in grocery retailing, which automatically register grocery expenditure per purchased item every time the customer swipes their card at the shop. The incentive is a financial reward in exchange for data with the retailer. Customer loyalty programmes are widely used among retailers, with the aim of increasing customer loyalty. For example, in the UK, nearly 90% of retail customers belong to at least one customer loyalty programme.<sup>18</sup> By analysing the data, retailers can build customised marketing communications, identify the

most profitable customers, establish relevant segmentation and consumer profiles, and promote complementary or more expensive products to customers on the basis of their previous purchases.<sup>19</sup> Prior studies have indeed observed an association between card ownership and loyalty.<sup>20,21</sup>

An obvious problem with loyalty card data from a research perspective is the distribution of the purchases over several retailers. This problem may, however, not always be as restrictive as it first seems. The Nordic markets are highly centralised: the three largest market chains claim an average of 80–90% of the market share.<sup>22</sup> In Finland, the largest commercial operator (S Group) had a market share as high as 47.2% in 2017, which enabled investigation of larger population groups based on data from a single retailer.<sup>22</sup> This, combined with the fact that in Finland the loyalty card uptake among consumers is the highest of all the European countries,<sup>23</sup> provides a relatively reliable means with which to evaluate national-level food intake via loyalty card data. In addition, automated registration of expenditure and the long duration of data collection make the assessment of diet, smoking and mild alcohol consumption less subjective and free from recollection error. Tin et al.<sup>24</sup> reviewed 18 studies using supermarket sales data for various population food and nutrition monitoring purposes. Their findings support the feasibility of using supermarket sales data to monitor a population’s food purchasing patterns. Finally, it is possible that unplanned and unhealthy purchases – including alcohol and cigarettes – are made in smaller stores. In Finland, since the opening hour regulation (2016), smaller retailer concepts have become more competitive than kiosks. Purchases made in these smaller stores are also recorded to loyalty card databases, which reduces possible bias due to differences in the expenditure profiles of large and small stores.

The aim of this paper is to introduce a research project addressing the potential and the challenges of loyalty card data for health research. Based on large-scale loyalty card data obtained from Finland, we present our data collection process, as well as the first empirical results regarding participant and purchasing profiles. In the discussion, we also provide a SWOT (strengths, weaknesses, opportunities, threats) analysis of the value of loyalty card data.

## Methods

### Setting and recruitment

The loyalty card data were provided by the S Group (S-ryhmä), a major Finnish retailer co-operative operating in Finland, Russia and the Baltic countries (<https://www.s-kanava.fi/web/s/en/s-ryhma-lyhyesti>). We contacted the

card holders (owners of an S Group loyalty card) via email and asked for electronic informed consent to obtain selected background characteristics (age, gender and residential postal code) and grocery expenditure data concerning them from 1st of January to 31st of December, 2016, for research purposes, without personal identifiers.

An invitation to the study, with an electronic consent form, was emailed to 245,877 customers ('customers' sample) (i) with a known email address, (ii) who had granted permission to be approached with research queries, (iii) who were at least 18 years of age, and (iv) who held a loyalty card of the HOK-Elanto retail cooperative (S Group) operating in the Southern Finland region, reaching from the capital city of Helsinki about 50 km north. The total number of customers in the HOK-Elanto database was 915,797, and geographically they were mainly located in Helsinki and nine nearby municipalities. The 'personnel' sample was based on email contact with 13,763 S Group employees throughout Finland. At the time of the queries, the S Group had 40,482 employees. The emails were sent in April 2017 and the collection of electronic informed consents ended in May 2017.

### Ethical issues

Ethical approval was obtained from the University of Helsinki Review Board in humanities and social and behavioural sciences. We committed to following the 'Responsible conduct of research and procedures for handling allegations of misconduct in Finland' (The Finnish Advisory Board on Research Integrity). Both the research group and the S Group signed a contract on data transfer, ensuring the independence of the research and scientific publishing from business interests.

### Statistical analysis

The background characteristics and expenditure were mainly analysed descriptively using summary statistics and figures. To analyse the determinants of total expenditure from S Group stores using the individual-level data available, we applied a linear mixed model. The model was based on the customer sample and included age group (categorised as <25, 25–29, 30–34, ..., 80–84, 85– years), gender and their interaction as fixed effects, and postal code as a random effect to reflect spatial heterogeneity.

To assess the extent of the similarity of the customer sample to the general population in the same region, we obtained the age–gender distribution of the 10 municipalities in the HOK-Elanto region from Statistics Finland. We could not identify any relevant reference

population for the personnel sample. We then modelled the probability of being a HOK-Elanto loyalty card holder *and* participation among all residents of the region applying a logistic regression model, which enabled the investigation of the distribution of the dichotomous outcome variable (participant, yes/no) on several explanatory variables. We included gender, age, gender by age group interaction, and municipality in the model as fixed effects. In this analysis, to match the available statistics, the age groups were defined as 16–19, 20–24, 25–29, ..., 80–84, 85–. From the modelling result, we used inverse probability weighting to adjust the sample to the regional population and to reduce the non-participation bias in expenditure.

## Results

### Sample characteristics

The data on expenditure consisted of 14,595 consenting loyalty cardholders (5.6% of those contacted). We obtained the background data of 14,522 loyalty card holders altogether: 13,274 customers (5.4%) and 1248 members of personnel (9.1%). Supplemental figures 1 and 2 illustrate the geographical distribution of the participants. The majority of the customer participants lived in Southern Finland as expected, but the personnel sample was spread throughout the country. Employees were included in the sample to offer wider perspective into nationwide food purchases and as they form a loyal and interesting socio-economic subgroup, which is on the front-line facing changes in the labour market due to digitalisation.

Overall, more than two-thirds of the participants were women (Table 1). The gender distribution was particularly skewed in the personnel sample, of which 80.5% were women.

The average age of the participants was 45.7 years (Table 1). Not surprisingly, the personnel sample participants tended to be younger than those in the customer sample; the average age difference being approximately six years. The customer sample contained a substantial proportion of older people: 25% of the participants were at least 58 years old, and 10% were at least 67 years old. The same percentiles were 49 and 56 years in the personnel sample, respectively.

The population age–gender distribution in the 10 municipalities showed that women were more likely to participate in the study than men. Municipality statistics showed that 52.1% of the HOK-Elanto region residents, who were at least 15 years old, were women. Although the mean age of the residents was similar (46.2 years) to that in the sample, we observed that young and old participants were underrepresented (Supplementary Table 1). Residents of Helsinki were

**Table 1.** Observed background characteristics of customers and personnel. No background data were available for  $n = 73$  participants.

	<i>n</i> (%)	Number of distinct postal codes	Gender		Age (at 1.1.2016)	
			Female	Male	Mean (SD)	Range
Customers	13,274 (91.4%)	655	8937 (67.3%)	4336 (32.7%)	46.2 (14.7)	16–90
Personnel	1248 (8.6%)	669	1004 (80.5%)	244 (19.6%)	40.1 (11.6)	18–74
Overall	14,522 (100%)	962	9941 (68.5%)	4580 (31.5%)	45.7 (14.5)	16–90

the most likely to participate. All factors of the model were highly significant ( $p < 0.0001$ ; data not shown).

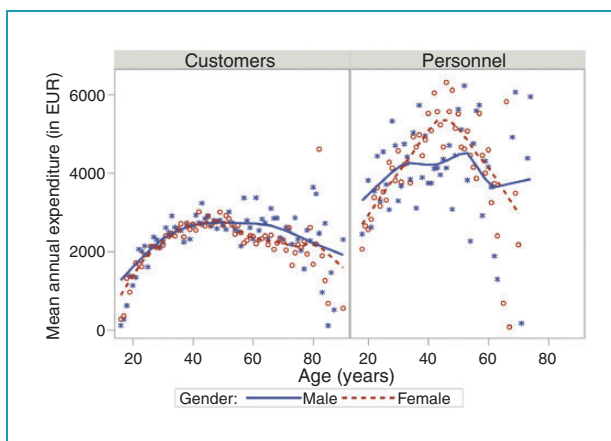
### Expenditure data and determinants

Expenditure data contained the total grocery expenditure of  $>13$  million purchase events, which were pre-classified into 184 product groups and accompanied by a timestamp: date of purchase and time of day. Each purchase was recorded in euros.

The overall means of total annual expenditure were 2443 EUR and 4419 EUR in the customer and personnel samples, respectively. When corrected to the age and gender distribution at each municipality, the weighted mean for the customers was lower: 2322 EUR. In the customer sample, we observed that both gender and age were significant determinants of total expenditure ( $p < 0.0001$ ). Mean expenditure peaked at middle age and declined towards both ends of the age range (Figure 1). Men's expenditure tended to be greater than women's, but the gender difference diminished in the 35–50-year age range (Figure 1;  $p = 0.016$  for the gender by age interaction term). We also noted some variation due to postal code areas ( $p < 0.0001$ ). However, this variation was not large. The high proportion of women and less widely spread age distribution in the personnel sample did not facilitate the same analysis of determinants as that of the customers. The right panel in Figure 1, however, suggests similar patterns to those in the customer sample. Most notably, the mean annual expenditure was substantially larger in this sample, which is probably due to the more attractive personnel reward system of an additional 3% discount.

When ranked by euros spent annually, we observed that the 10 most popular product groups were the same in the two samples, and their expenditure patterns were similar (Figure 2). Interestingly, beer and cigarette product groups were among these, and the mean cigarette expenditure of the personnel sample was approximately twice that of the customer sample.

Cyclic purchasing behaviour during the week was evident (Figure 3); Fridays and Saturdays were the



**Figure 1.** Mean annual expenditure by age and gender in the two samples. Observed age–gender specific means are denoted by markers, and solid lines indicate weighted loess curve fits.

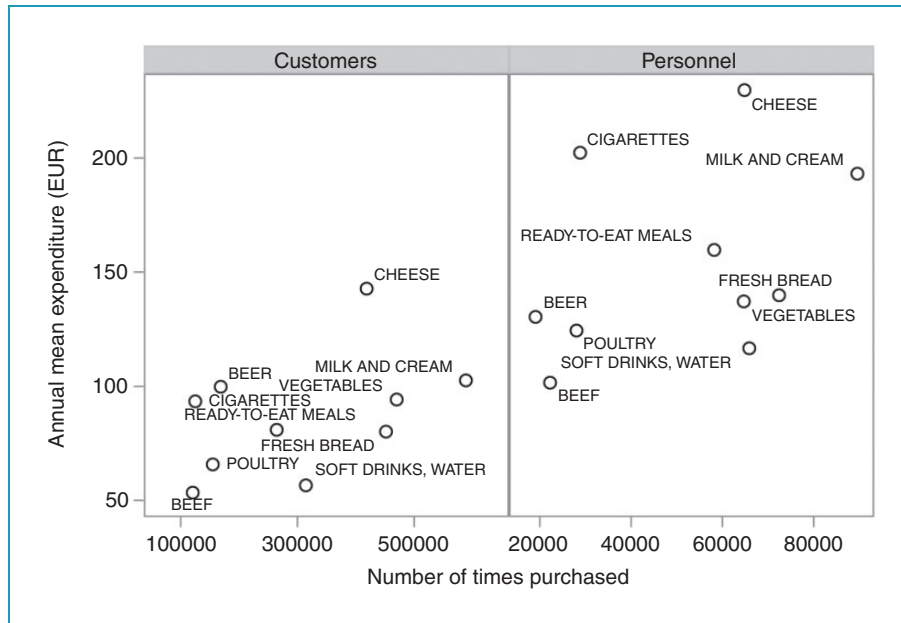
most active purchasing days, and Sundays and Tuesdays the least active. National holidays appeared to be preceded by a peak in expenditure. ‘Skiing holiday week’ was in week 8 in 2016, and the associated decline in expenditure is observable in Figure 3.

### Discussion

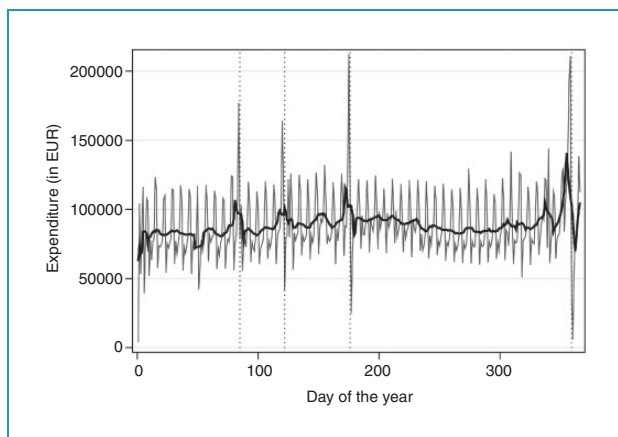
Our aim was to introduce a research project that would address the potential of loyalty card data for health research. The project was based on large-scale loyalty card data obtained in Finland: altogether 14,595 loyalty card holders consented to releasing their retrospective expenditure data.

### Participation in the study

In recent decades, participation rates in health-related surveys in Finland, as in most other countries, have declined.<sup>25</sup> In our study, the participation rate fell well below those of comprehensive surveys, and was highest among personnel, women, middle-aged people, and customers living in Helsinki. Several reasons could explain this. First, it is possible that emails



**Figure 2.** Most purchased product groups (top 10, ranked by expenditure) among customers and personnel.



**Figure 3.** Total expenditure of customers during calendar year. Daily sums (grey solid line) and centred seven-day moving average (black solid line). The vertical reference lines are positioned at Good Friday, 1st of May, Midsummer Eve and Christmas Eve.

do not reach everyone who is invited, because addresses may be outdated, or the invitation could be directed to the junk folder. Second, an invitation by email is easy to delete, forget or ignore, particularly if it does not raise the immediate interest of the recipient. Third, the recipients could indeed be less willing to share detailed and retrospective expenditure data on themselves than to provide answers to surveys. The determinants of the participation rates are partly in contrast with the results of the FINRISK surveys conducted in 1997–2012, in which women living in the capital region were more likely to be non-participants

in surveys than women living in rural areas.<sup>25</sup> In surveys, the non-participants are more likely to be young, male, unmarried, have a lower education level and income, and of foreign origin, factors that are also associated with less optimal health behaviour.<sup>8,26</sup> As we did not obtain the background characteristics of everyone invited, we could not directly compare the non-participants and the participants. However, we could obtain summary measures among HOK-Elanto region loyalty card owners as a whole, including those who did not fulfil the invitation criteria: their mean age was 46.0 years and 59.4% of them were women. This partly explains the difference in the gender distribution between the residents (of whom 52.1% were female) and the customers (67.3%), but not completely. The mean age fell very close to the means of the residents and of the customers. However, the availability of an average measure does not shed light on the observed underrepresentation of the young and old participants in the sample.

We had limited access to socio-demographic measures. However, we expected our personnel sample to be less educated than average Finnish employees, as 13% of people working in the retail sector have a low-level (less than high school), 62% a mid-level (high school, vocational school), and 25% a high-level (BSc degree or higher) education.<sup>27</sup> The respective proportions among all Finnish employees were 10%, 45% and 45%. Despite the apparent difference between the education frequency distribution of the two samples, we observed a higher participation rate in the personnel sample than in the customer sample.

Our study shows that loyalty card data offer the opportunity to reach a large, potentially more heterogeneous sample with far less resources than other surveys of the same magnitude. However, the participation rates were low, and the data did not seem to fill all known participation gaps, for example, those related to age, and we cannot rule out the possibility of the socio-demographic gradient in non-participation. Therefore, in studies such as ours, with poor participation rates but cost effectively obtained large sample sizes, it would be even more important than in surveys to collect the variables related to non-participation in order to at least partially correct any biases that may result from this (e.g. by inverse probability of participation weighting). Unfortunately, this is not always possible, and such variables may differ from those that are predictive of non-participation in surveys, which means that they would have to be identified in advance. Parallel interest in the development of methods for correcting for data missing not at random may turn out to be useful for large-scale data with selective participation.<sup>28,29</sup>

In the future, increasing population diversity alongside declining trends in response rates will challenge large national dietary surveys, as a representative sample will become even harder to recruit, and the rapidly changing supply of food products will require comprehensive dietary assessment methods and continuously updated food composition databases. At the same time, digital development allows easy access to expenditure data with available detailed information, enabling the detection of differences in the diets of different population groups. However, this often requires additional information on customers, a type of data that is more difficult to obtain directly from retailers due to data confidentiality. A separate questionnaire on background variables or linkage with other relevant datasets could provide additional socio-demographic measures, which would increase the value of the data.

### *Expenditure profiles and their determinants*

We observed that the top products were the same in the two samples, even though the participants in the personnel sample purchased considerably more from the S Group. The total expenditure does not seem to be strongly associated with the most commonly purchased products, which could mean that the purchases from other retailers are similar to those in our data. Demonstrating a high proportion of expenditure within the retailer's stores combined with similar relative distribution of purchased products across different levels of retailer-specific expenditure add to the

credibility of the data. However, we cannot empirically rule out the possibility of differences between the levels.

Information on household size and members was not available, but it is likely that the number of children and adults in a household could largely explain the differences between age groups in mean annual expenditure.

The mean of cigarette expenditure of the personnel sample was approximately twice that of the customer sample. This could be partly due to the lower socio-economic status of the personnel and hence the higher prevalence of smokers, or because they purchase a larger share of their groceries and cigarettes in the store where they work. The variation due to postal code areas could be indicative of socio-economic differences or due to the geographical location of the stores.

### *Strengths*

One of the clear fundamental strengths of loyalty card data is the fact that they are not based on perception: without risk of impreciseness or over- or underestimation, loyalty card data capture the kind of items a particular person has bought from a specific retailer. They use retailers' existing data infrastructure, cash registers, and IT infrastructure to collect vast amounts of data that accumulate continuously over time. Loyalty card data can be regarded as a by-product: they result from retailers' existing processes and can be shared for health research purposes with marginal additional costs.

The sheer size of loyalty card data is a clear asset. It not only allows investigation of the whole sample, but also studies smaller subpopulations such as the elderly, on whom comprehensive data are often hard to obtain.

As customer loyalty programmes accumulate data on a household level, loyalty card data automatically provide longitudinal insight into how food expenditure evolves over time, given that the share and the composition of food expenditure for a particular retailer does not change. Traditional methods are often limited to cross-sectional data or focus on shorter time periods (e.g. one month). In contrast to loyalty card data, consumer panel data (e.g. Nielsen Homescan) cover all purchases for a certain time period but may suffer from selection bias of more affluent populations and misreporting, similar to traditional dietary assessment methods.

### *Weaknesses*

Customers often purchase food from different sources. Some may prefer buying vegetables from marketplaces, meats from primary retailers, and bread from a local bakery. Similarly, as customers may divide their

grocery shopping between several retailers, analysing data from only one retailer could give an incomplete picture of household-level expenditure. This is one limitation of loyalty card data in markets with many food retailers, such as the UK. However, in the Nordic countries in particular, where the level of market concentration in grocery retailing is high, loyalty card data obtained from even a single retailer could provide a good picture of household expenditure. Statistics Finland reports (preliminary data from 2016) that the average annual consumption of groceries and non-alcoholic drinks is 2916 EUR per household, and of alcohol and cigarettes 578 EUR per household.<sup>30</sup> Our estimates fall close to these national benchmarks. Although indirect, they are indicative of a very high proportion of expenditure among loyalty card holders in S Group stores. Interestingly, the observed means of personnel exceeded the national estimates. This may indicate that personnel heavily concentrate all their shopping in S Group stores due to personnel discounts.

Although loyalty card data capture what is *bought*, they cannot directly reveal exactly what is eventually *consumed* and *by whom*. Consumers living in the same household may eat (or drink) out, share dinners with friends and relatives, or buy food for pets or others not living in the same household. Loyalty card data therefore only partially reflects consumption by household members. Good compatibility between respondent-collected household-level food purchase data (covering all purchases) and individual-level dietary data has been demonstrated,<sup>31–33</sup> and thus, household-level food purchase data can reasonably reliably be used to model individual-level dietary patterns if all purchases are recorded. In addition, consumers with specific socio-economic backgrounds may be prone to specific types of retailers. For example, the customers of a premium food retailer or a hard discounter may offer heavily biased consumption patterns in relation to both food consumption and healthiness. Therefore, to address this weakness, loyalty card data should be compared with population data whenever possible. We are currently planning a validation study that will address these limitations more closely.

The value of loyalty card data is highly dependent on data accuracy. For example, loyalty card data can be collected on a total sum level (i.e. expenditure of 43.50 EUR), product category level (24.20 EUR worth of vegetables), or product level (4.20 EUR worth of carrots). However, food expenditure data may not be grouped into clearly demarcated functional categories, which would be essential for linking with health data. In nutritional science, commonly used food groupings are mainly based on earlier findings regarding the associations between dietary components and health. The use of different categories weakens

comparability between studies. Too vaguely or incorrectly (from the research point of view) categorised data – such as plant-based protein products being classified into the same category as whole meat products – even in large volumes, have little value. In this respect, access to loyalty card data does not necessarily mean high health research potential.

### Opportunities

Loyalty card data may offer a unique, underutilised data source for health research that significantly complements existing data sources. Second, utilising loyalty card data for the benefit of health research could uncover the societal potential of customer loyalty card data for collective benefit. In addition to their business value, the data can be used as a resource for creating societal value. This could also offer companies another way in which to fulfil their corporate social responsibility: Through their own actions, they can leverage existing data assets for societal benefit. Third, combining loyalty card data with other data sources could further amplify their potential for any research. For example, linking loyalty card data with health outcome data, either on the individual or regional level, could offer new perspectives to the association between specific diseases and diet/food (un)healthiness. This could be done using data from national health surveys or medicine statistics conducted in the same region, which could be linked by, for example, postal code area. This strategy would enable ecological analysis of the association between regional expenditure and health outcomes. Another possibility would be to approach the loyalty card owners directly through surveys or invitations to health examinations. However, this could prove to be difficult in practice. Fourth, loyalty card data may provide a lens through which researchers can evaluate the impact of various health promotion campaigns and policy interventions (e.g. taxes) on actual food consumption. For example, in close collaboration with food retailers, researchers could design a set of interventions to guide consumers towards healthier diets, and through loyalty card data evaluate their impact on different customer groups. This would help us understand what types of interventions, mechanisms or incentives influence consumers' food consumption.

In addition to these opportunities, Tin et al.<sup>24</sup> identified other areas of nutrition monitoring that could use supermarket sales data: the assessment of food purchase patterns within and between population groups, longitudinal comparison of population food purchase patterns with regard to policy or economic changes, the derivation of nutrient availability by linking food purchase data to food composition data, the validation of

self-reported dietary data by comparing actual food purchases with reported purchases, and the assessment of key determinants of the healthier food choices of a population. We add to this the possibility of tailoring personal feedback by using customers' usual food expenditure and hence influencing their food purchases. Ideally, an online feedback system could enable a customer to check the dietary quality or carbon footprint of their own expenditure. Such interaction between the customer, researchers and the retailer would substantially add to the public health value of the data. In New Zealand, individualised electronic supermarket sales data were used to tailor culturally targeted nutrition resources for ethnically diverse shoppers in a large supermarket trial.<sup>34</sup>

The potential of purchase data would be greater if they were reliably linked with other relevant datasets, and with correction factors to provide valid estimates of consumed food from bought food.<sup>24</sup> Keeping up with the modern food landscape requires systematic, meaningful linkages across data on food purchase/sales, food intake, food composition, and nutrition fact panels.<sup>35</sup> Ng and Popkin<sup>35</sup> concluded that the ability of researchers and nutrition professionals to properly integrate different data sources and to fully capitalise on the arising opportunities remains somewhat undeveloped. However, the number of studies based on purchase data supplemented with other data sources is increasing. A recent Danish study<sup>36</sup> combined comprehensive household food purchase data with nutrition information and individual register data when assessing the effect over time of unemployment on food purchase behaviour and diet composition. The study provided valuable methodological examples to be utilised with similar datasets. The Diet-Related GHG Index has also been modelled on the basis of the same Danish household food purchase data and a food frequency questionnaire.<sup>33</sup> However, digital epidemiology is still in its early stage of development. Methodological robustness is an ethical as well as a scientific requirement, involving, for example, the validation of algorithms, a better understanding of confounding, filtering systems for noisy data, management of biases, and the selection of appropriate data streams.<sup>37</sup>

### Threats

In the past decade, data privacy has become an important issue for both companies and consumers; recent discussion on the issue may have made consumers more aware of possible misuses of data. For example, in terms of social media usage, several reports and news items have highlighted how consumers should be more concerned about what data are being collected on

them.<sup>22,38,39</sup> In general, most consumers are concerned about the data collected on them on the internet. In particular, recent data privacy issues with Facebook, which linked consumers' social media usage to targeted political advertisements may have spillover effects on other forms of customer data usages, including loyalty programmes. However, with written consent approved by an ethical board, purchase data could be utilised for the common good in a transparent manner, while respecting individual rights and liberties, which is the crux of the debate on the ethics of big data.<sup>37</sup>

From a company's point of view, customer data have become a key component of almost any business. Consequently, data privacy and protection have become a focal competence area. The General Data Protection Regulation provides new guidelines for organisations in the European Union for improving the management of risks and practices related to customer data. The regulation may also limit retailers' willingness to share data to third parties, including researchers, and companies may become afraid of the negative publicity that could result from allowing researchers to investigate their customer data. However, sharing valuable data offers a company the opportunity to carry societal responsibility.

### Conclusions

More attention needs to be paid to the potential of loyalty card data for customers, research or society at large, rather than for retailer purposes only.<sup>40</sup> In this paper, we have introduced a unique dataset of 14595 S Group (retail markets) loyalty card holders. These data consist of all food purchases, spanning over a period of one year, in total >13 million purchase events. They accumulate automatically, with high resolution, and without participant-based reporting bias. Although the potential for these kinds of digital big data is considerable, challenges also arise: 1) the data collected represent household purchases for most customers, 2) only a part of all food purchases is from one single retail chain; and 3) possible future restrictions and concerns regarding individual data privacy. In concentrated markets such as those in the Nordic countries, these resources are nevertheless highly intriguing for researchers. In the future, efforts to overcome their limitations, and balancing size with sufficient level of detail could lead to new viewpoints regarding population exposure, behaviour and lifestyle.

**Acknowledgements:** We thank the S Group for their collaboration.

**Contributorship:** JN, ME and MF conceived the study. JN wrote the first draft of the manuscript and conducted the data



analysis. ME and HS researched the literature. TN was responsible for data management and participated in the data analysis. All authors reviewed and edited the manuscript and approved the final version.

**Conflict of interests:** The authors declare no conflicts of interest.



**Ethical approval:** The ethics committee of the University of Helsinki Review Board in the humanities and social and behavioural sciences approved this study (REC number: #43/2016).

**Funding:** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**Guarantor:** JN

**Peer review:** This manuscript was reviewed by Cliona Mhurchu, The University of Auckland, New Zealand, and one other individual who has chosen to remain anonymous.

#### ORCID iD

Jaakko Nevalainen  <http://orcid.org/0000-0001-6295-0245>  
Turkka Näppilä  <http://orcid.org/0000-0002-0562-7254>

**Supplemental Material:** Supplemental Material for this article is available online.

#### References

- World Health Organization. Global action plan for the prevention and control of noncommunicable diseases 2013–2020. Geneva: WHO, [http://www.who.int/nmh/events/ncd\\_action\\_plan/en/](http://www.who.int/nmh/events/ncd_action_plan/en/) (2013, accessed 1 February 2018).
- Schuit AJ, van Loon AJM, et al. Clustering of lifestyle risk factors in a general adult population. *Prev Med* 2002; 35:219–224.
- Willett W. Correction for the effects of measurement error. In: *Nutritional Epidemiology*. Oxford, UK: Oxford University Press, 2013.
- Bennett DA, Landry D, et al. Systematic review of statistical approaches to quantify, or correct for, measurement error in a continuous exposure in nutritional epidemiology. *BMC Med Res Methodol* 2017; 17:146.
- Livingston M and Callinan S. Underreporting in alcohol surveys: whose drinking is underestimated? *J Stud Alcohol Drugs* 2015; 76:158–64.
- Lovegrove JA, Hodson L, Sharma S, et al. *Nutrition research methodologies*. Hoboken, NJ: Wiley Blackwell, 2015.
- Willett W. *Nutritional epidemiology*. 3rd ed. Oxford, UK: Oxford University Press, 2012.
- Tolonen H, Helakorpi S, Talala K, et al. 25-year trends and socio-demographic differences in response rates: Finnish adult health behaviour survey. *Eur J Epidemiol* 2006; 21: 409–415
- Härkönen J, Savonen J, Virtala E, et al. Suomalaisten alkoholinkäyttötavat 1968–2016: Juomatapatutkimusten tuloksia [In Finnish, abstract in English]. Helsinki, Finland: Yliopistopaino, [https://www.julkari.fi/bitstream/handle/10024/134585/URN\\_ISBN\\_978-952-302-873-9.pdf?sequence=1](https://www.julkari.fi/bitstream/handle/10024/134585/URN_ISBN_978-952-302-873-9.pdf?sequence=1) (2018, accessed 23 February 2018).
- Rossow I, Mäkelä P and Kerr W (2014). The collectivity of changes in alcohol consumption revisited. *Addiction* 2014; 109:1447–1455.
- Pietinen P, Vartiainen E, Seppänen R, et al. Changes in diet in Finland from 1972 to 1992: impact on coronary heart disease risk. *Prev Med* 1996; 25: 243–250.
- Paturi M, Tapanainen H, Reinivuo H, et al. (eds). The National FINDIET 2007 Survey [In Finnish, tables, figures and summary in English]. Publications of the National Public Health Institute B23/2008. Helsinki: Yliopistopaino, <http://www.julkari.fi/bitstream/handle/10024/78088/2008b23.pdf> (2008, accessed 01 March 2018).
- Helldán A, Raulio S, Kosola M, et al. The National FINDIET 2012 Survey [In Finnish, tables, figures and summary in English]. Publications of the National Institute of Health and Welfare 16/2013. Tampere, Finland: Yliopistopaino, [https://www.julkari.fi/bitstream/handle/10024/110839/THL\\_RAP2013\\_016\\_%26sliitteet.pdf](https://www.julkari.fi/bitstream/handle/10024/110839/THL_RAP2013_016_%26sliitteet.pdf) (2013, accessed 23 February 2018).
- Illner AK, Freisling H, Boeing H, et al. Review and evaluation of innovative technologies for measuring diet in nutritional epidemiology. *Int J Epidemiol* 2012; 41: 1187–1203.
- Signal LN, Smith MB, Barr M, et al. Kids' Cam: an objective methodology to study the world in which children live. *Am J Prev Med* 2017; 53: e89–e95.
- Salathé M, Bengtsson L, Bodnar TJ, et al. Digital epidemiology. *PLoS Comput Biol* 2012; 8: e1002616.
- Lee EC, Asher JM, Goldlust S, et al. Mind the scales: Harnessing spatial big data for infectious disease surveillance and inference. *J Infect Dis* 2016; 214: S409–13.
- Tonini T. Consumer's perception of digital rewards in loyalty programs: insights from a multi-country research, 2nd Edition, *Target Research*. <http://www.imaeurope.com/wp-content/uploads/2017/05/TargetResearch.pdf> (2017, accessed May 15, 2018).
- Peppers D, Rogers M and Dorf B. Is your company ready for one-to-one marketing? *Harvard Business Review* 1999; 77: 151–160.
- Meyer-Waarden L. The influence of loyalty programme membership on customer purchase behaviour. *Eur J Marketing* 2008; 42: 87–114.
- Turner JJ and Wilson K. Grocery loyalty: Tesco Clubcard and its impact on loyalty, *Brit Food J* 2006; 108: 958–964.
- Statista. Statista: the portal for statistics, <http://www.statista.com> (2018, accessed 15 February 2018).
- Nielsen. Share of consumers that have loyalty cards in selected European countries in 2016, <https://www.sta>

- tista.com/statistics/792653/loyalty-card-uptake-europe-an-countries/ (2016, accessed 31 March 2018).
24. Tin ST, Mhurchu CN and Bullen C. Supermarket sales data: feasibility and applicability in population food and nutrition monitoring. *Nutr Rev* 2007; 65: 20–30.
  25. Tolonen H, Koponen P, Borodulin K, et al. Differences in participation rates between urban and rural areas are diminishing in Finland. *Scand J Public Health*. Epub ahead of print 06 January 2018. DOI: 10.1177/1403494817748737.
  26. Strandhagen E, Berg C, Lissner L, et al. Selection bias in a population survey with registry linkage: potential effect on socioeconomic gradient in cardiovascular risk. *Eur J Epidemiol* 2010; 25: 163–172.
  27. PAM palveluajon taskutilasto, <https://www.pam.fi/media/1.-materiaalipankki-tiedostot-nakyvat-julkisessa-materiaalipankissa/tilastot-ja-tutkimukset/palveluajon-taskutilasto-2017.pdf> [In Finnish] (2017, accessed 06 March 2018).
  28. Kopra J, Karvanen J and Härkänen T. Bayesian models for data missing not at random in health examination surveys. *Stat Model* 2018; 18: 113–128.
  29. Tchetgen Tchetgen EJ, Wirth KE. A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics* 2017; 73: 1123–1131.
  30. Suomen virallinen tilasto (SVT): Kotitalouksien kulutus [In Finnish]. ISSN=1798-3533. Helsinki: Tilastokeskus, [http://www.stat.fi/til/ktutk/2016/ktutk\\_2016\\_2017-12-28\\_tie\\_001\\_fi.html](http://www.stat.fi/til/ktutk/2016/ktutk_2016_2017-12-28_tie_001_fi.html) (2016, accessed 26 January 2018).
  31. Nelson M, Dyson PA and Paul AA. Family food purchases and home food consumption: comparison of nutrient contents. *Br J Nutr* 1985; 54: 373–387.
  32. Becker W. Comparability of household and individual food consumption data: evidence from Sweden. *Public Health Nutr* 2001; 4: 1177–1182.
  33. Lund TB, Watson D, Smed S, et al. The Diet-related GHG Index: construction and validation of a brief questionnaire-based index. *Climatic Change* 2017; 140: 503–517.
  34. Mhurchu CN, Blakely T, Jiang Y, et al. (2010). Effects of price discounts and tailored nutrition education on supermarket purchases: a randomized controlled trial. *Am J Clin Nutr* 2010 91: 736–747.
  35. Ng SW and Popkin BM. Monitoring foods and nutrients sold and consumed in the United States: dynamics and challenges. *J Acad Nutr Diet* 2012; 112: 41–45.
  36. Smed S, Tetens I, Bøker Lund T, et al. The consequences of unemployment on diet composition and purchase behaviour: a longitudinal study from Denmark. *Public Health Nutr* 2018; 21: 580–592.
  37. Vayena E, Salathé M, et al. Ethical challenges of big data in public health. *PLoS Comput Biol*, 2015; 11: e1003904.
  38. Price R. Billionaire ex-Facebook president Sean Parker unloads on Mark Zuckerberg and admits he helped build a monster. *Business Insider*, 09 November, <http://nordic.businessinsider.com/ex-facebook-president-sean-parker-social-network-human-vulnerability-2017-11?r=UK&IR=T> (2017, accessed 15 February 15 2018).
  39. Shinal J. Ex-Facebook privacy manager says company cares more about data collection than protecting users. *CNBC*, 20 November, <https://www.cnbc.com/2017/11/20/former-facebook-privacy-manager-sandy-parakilas-criticizes-zuckerberg.html> (2017, accessed 15 February 2018).
  40. Saarijärvi H, Kuusela H, Kannan PK, et al. Unlocking the transformative potential of customer data in retailing. *Int Rev Retail, Distrib Consum Res* 2016; 26: 225–241.
-