# Small scale “big data” in the Finnish pharmaceutical product index compilation

*Kristiina Nieminen, Antti Suoperä, Satu Montonen*

*5th May 2017*

## Abstract

In recent years several countries have started to modernise their data collection for CPI by utilising scanner data. Statistics Finland’s long term target is to collect the prices, quantities and descriptive information of consumer product for all needed periods instead of just collecting a sample of the representative commodities. In short term, aim is to create new practices for scanner data collection, quality management and index compilation. First obtained scanner data is received from the Pharmaceutical Information Centre Ltd that provides reliable and up-to-date sales and price information of medicines in its’ information-hub.

First part of the paper explains how the scanner data of pharmaceutical products is processed and which are the main improvements done in the process. Based on our experience, this process-flow approach guarantees the transparence of the process and confirms easy access to the used methods and index formula. Examples of these new practices and the process flow will be described in short. In future, statisticians maintain and update the process flow, thus enabling more fluent adaption of the new data sources to the CPI production.

The second part describes the results of using the new data sources which shows substantive quality improvement in this specific commodity group. Different index formulas were tested in co-operation with the University of Helsinki/Department of Political and Economic Studies (professor Yrjö Vartia). These findings are described briefly.

The last part of the paper raises ideas for further development for using scanner data in statistical processes.

## 1. Introduction

Statistics Finland had few attempts to utilise transaction data in Consumer Price Index production ten years ago but so far to collect massive amounts of data and to process it in systematic manner has been a challenge. This challenge is now past due to the fact that first dataset containing thousands of products from one commodity group is implemented into production. New IT-systems enable the use of even bigger amounts of data in CPI production and with bigger datasets it is possible to test, to evaluate and to develop traditional compilation practices.

To define the size of big data is not unambiguous but typically this term refers to the data containing millions of unstructured observations. Instead in this case we refer with the term “big data” to the structured data set that contains all the commodities and essential information concerning the products. In Finland there are approximately 10 000 products available in pharmacies, but situation may differ depending on the size of the country or the commodity group concerned. Therefore we call this data “small scale big data”.

In this paper, the aim is to look at utilisation of scanner data from two perspectives: the first approach presents the practices used in processing of pharmaceutical data and the second approach then investigates the effect on changes made to the compilation strategy. The latter approach covers description of three separate calculation sets where different strategies were applied and shows the results from these.

The research of alternative compilation strategies were carried out in joint work with professor Yrjö Vartia[1] from Helsinki University and methodologist Antti Suoperä from Statistics Finland. In this paper these alternative strategies are shortly described and the results presented. More detailed information about the

[1] We want to express our gratitude to professor Yrjö Vartia for the co-operation, advices and for encouraging us to research the data further

methods and the results may be found from the document: *Index number theory and construction of CPI for complete micro data* by authors Vartia and Suoperä and from the document *Analysis of Change and Constructing Index Series for Complete Micro Data* by Suoperä, Nieminen, Montonen. These documents are available in Statistics Finland web-page: www.stat.fi/meta/menetelmakehitystyo/index_en.html.

In chapter 2, the background of this paper is highlighted. This outlines the starting point to the design of test calculations and analysis of the results that are main concern in this paper. After this, the orientation to the data is done in chapter 3 that gives an overview to the markets of pharmaceutical products, pricing policy and the identification of a product. The designed practices are introduced in chapter 4. The aim is here to demonstrate two practices (the definition of compilation strategy and use of metadata in data collection) that national statistical institutes (NSI) may find useful when they are taking new data sources taken into testing and finally to CPI production.

The selected strategies and the results from separate calculation sets are highlighted in chapters "5. Compilation of elementary aggregates" and "6. Results". Finally chapter "7. Conclusion" summarises the current state, the objectives for the becoming 3 year period and gives proposals for further research.

## 2. Background

### The first attempt to utilise transaction data in Statistics Finland in year 2000

Statistics Finland started its' first examinations to utilise transaction data (also known as scanner data) in Consumer Price Index (CPI) data collection and production at the end of year 2000. At the same time in 1999 a co-operation project was organised between Statistics Finland, AC Nielsen [2]Finland and Yrjö Vartia from Helsinki University.

The aim of research project was to analyse how AC Nielsen's transaction data may be utilised in CPI production. The received data consists of Scan Track data for every January for period 1995-1997 and for all months in 1998. The received research data was based on the weekly sales of the bar coded products including product prices and quantities.

Following item groups were selected to the further investigation

- Oils and fats (coicop 01.1.5)
- Mineral waters, soft drinks and juices (coicop 01.2.2)
- Cleaning and maintenance products (coicop 05.6.1.1)

These commodity groups were chosen as there was enough data available from longer time period and it was possible to divide selected products into sub-level categories in Classification of Individual Consumption by Purpose (COICOP).

On the basis of studies carried out in 2000-2001, it was concluded that the scanner data may be used as a basis in the CPI compilation. However, the research project discovered the need for further research before scanner data may be implemented into the production process. The project suggested following topics for further research: data collection, classification of items and index compilation as well as evaluation of the scanner data coverage to determine the commodity groups for which scanner data might be used. (Ylä-Jarkko, 2001)

After this, the research to utilise scanner data continued in other countries such as Norway, Netherland and New-Zealand.

[2] AC Nielsen studies consumers in more than 100 countries. Information from web-page http://www.nielsen.com/eu/en.html

## Eurostat's recommendations on obtaining and processing the scanner data

During the years 2015-2016 Eurostat composed, with HICP Expert Group members, recommendations to obtain scanner data and to process collected data. The recommendation is divided into two parts. The first part gives guidelines on what data to ask for, from whom, with what frequency and also gives guidance on checking the quality of the data (Eurostat, 2015) while the second part concentrates on how to process acquired data.

The first part of the recommendations is already accepted in Price Statistics Working Group so these recommendations may be used as a guideline in NSI when collecting scanner or similar type "big data". The second part of the recommendations is under development at the moment. Statistics Finland has taken part in Eurostat's joint-venture work and has taken account the recommendations when acquiring the scanner data.

## The new project, initiated in 2014, to modernise the CPI price collection

The new efforts to implement scanner data in Finnish Consumer Price Index (CPI) and Harmonised Indices of Consumer Prices (HICP) were once again set into action when Statistics Finland participated Eurostat's venture on "Modernisation of price collection and compilation". Fundamental reason to start utilizing the scanner data was to speed up the data collection, to lower the response burden, to decrease the data collection costs, to ensure the coverage, and most of all to improve quality of the results and to start using the new data sources in statistics production.

Statistics Finland initiated an internal project on the modernisation of data collection in the HICP in 2014. Project was titled as "Web Collection on Consumer Prices" and the working period for the project was from November 2014 to March 2016. The main goal of the project was to study, to develop and to test the capabilities on how to utilise scanner data and web scraping in statistics production.

More specifically, the tasks of the project were:

1. Stocktaking of the existing practices
2. Re-design of the data collection and data capture arrangements
3. Re-design of the index compilation process
4. Development (or purchase) the required software and implementation of the required changes to the current IT-environment.

The stocktaking task aimed to review existing practices in other NSI's regarding the use of scanner data and automated internet data collection like web scraping. During the project, Statistics Finland successfully negotiated and finally obtained two scanner-data for testing. First one, smaller in amount of commodities was received concerning pharmaceutical products (1) while the other concerned daily consumption goods (2). It was noted that same practices to obtain scanner data may be applied for any other scanner data deliveries to Statistics Finland. (Koskimäki & Saranpää, 2016)

The project group researched different index compilation strategies to utilize scanner data (no 1) in CPI production. First tests were carried out by following the guidelines given in ILO manual (ILO, 2004) and the Eurostat HICP regulation (Eurostat, 2016). After these initial results, testing was expanded to cover the other index formulas, especially the superlative price index formulas such as Fisher, Törnqvist and Vartia. The reason to test alternative strategies was that this was first time when a complete dataset containing prices, quantities and descriptive information from several year period was available.

The research work in project was based on the pharmaceutical products scanner data, but these tests and practices were designed to be transferrable for any other new data source.

Next, a closer look is taken at the received pharmaceutical products data to form comprehend of the data content and the pricing mechanism in Finland.

## 3. Orientation to the obtained data

Traditionally it has not been easy to obtain complete datasets but luckily in Finland digitalisation has proceed efficiently. Companies have nowadays modern information systems in use for recording, monitoring and analysis of information captured from the transactions and the operations. Therefore, it is possible to ask for detailed scanner data of prices and quantities by products.

### Description of the ordered data

The data used in the compilation of the pharmaceutical index is obtained from the Pharmaceutical Information Centre[3] that maintains up-to-date drug information database (Pharmarket™) for the needs of consumers, healthcare professionals, companies and healthcare IT providers.

Pharmarket™ is a chargeable service produced by the Pharmaceutical Information Centre for monitoring the sales and prices of prescription medicines, over-the-counter medicines and free trade preparations delivered through pharmacies. The sales data of the service is collected from wholesalers and it is integrated with the Pharmaceutical Information Centre's commodity database.

The price of pharmaceutical product contains few elements that has effect on the total price. These are wholesale price, pharmacy margin, pharmacy fee and value added tax (Source: Pharma Industry Finland[4]). Wholesale price is determined by pharmaceutical companies but according to the law price must be the same in all regions and in every pharmacy in a month. It is not allowed to offer discounts in medicine prices in Finland.

The wholesale prices of medicinal products need to be confirmed by the Pharmaceuticals Pricing Board (Hila) that operates under the Ministry of Social Affairs and Health. Hila's decision-making is based on the Health Insurance Act. Hence pharmaceutical companies have to apply for confirmation before the effective prices may be changed. Consequence to this is that medicine price may change only twice a month.

The identification of the commodities is based on the barcode that equals with the commodity-level unique identification code (VNR[5]). This VNR-code separates commodities unconditionally, keeping abreast with disappearing and new commodities. Hence there is no possibility that a commodity disappears from the market and comes back later with different identification code, also known as relaunch of an item.

### Agreement to acquire monthly data

At the end of 2015, the agreement was signed with the supplier on continuous monthly data delivery. The content of the data, delivery times and the structure of the data were defined in the agreement in accordance with the ascertainment of needs of the consumer price index calculation. The recommendations given by Eurostat were followed in general.

The obtained data contain all prescription and over-the-counter medicines, as well as non-medical preparations (free trade products or common merchandise) from self-medication groups that were selected by Statistics Finland. The ordered data contain monthly data on sold commodities including their quantities and the prices. This data is complemented with additional register data where individual commodity is described separately. The structure of the acquired data is presented in Appendix 1.

---

[3] http://www.laaketietokeskus.fi/en/about-us

[4] http://www.pif.fi/en/statistics-and-reports/medicine-price

[5] The Nordic article number (Vnr) is a six-digit code given to all medicine packages. VNR -code is used to identify an individual medicine package at any point in the medicine distribution chain. Source: https://www.laaketietokeskus.fi/en/vnr

The obtained data contain 32 variables and approximately 10 000 individual commodities in a month. The data of quantities contains approximately 150 000 observations in a month because commodities are divided by hospital district and sales channel (hospital sales, pharmacy and retail sales). Hospital district may be used in compilation of regional indices.

### Description of the data collection

Previously, the data source for the index compilation of pharmaceutical products, has been the price list offered by Apteekkariliitto (Pharmacists' union) from where the prices for over-the-counter and prescription medicines have been selected manually. The price list of Pharmacists' union was based on the data from the Pharmaceutical Information Centre so the source of the data did not change but it became more extensive and up-to-date compared to the manual selection.

The first data delivery was made in January 2016 containing historical monthly data for the past seven years (2009 to 2015). After this, the monthly delivery is done by the 3rd day of month.

The Pharmaceutical Information Centre automatically compiles the required data onto the SFTP server as semicolon separated CSV files. The number of field separators in the CSV files is standard so all the data in the files is delivered even if the value of an variable is zero or missing (“ “). Automatic script extracts the files directly from the Pharmaceutical Information Centre’s SFTP server and transmits the data to Statistics Finland internal server.

## 4. Practices

In this chapter we will have a look at the practices that were used to design the index compilation strategy and to process the acquired data. Different approaches for the definition of strategy are widely described in literature but in this case we refer to the work of Yrjö Vartia in 1976. Ideas about various strategies for constructing index series are treated in his book *Relative changes and Index numbers* from where the first practice is adopted.

### Practice no 1. Define an index compilation strategy

First step after the scanner data is obtained is to define an index compilation strategy. Following analysis of pharmaceutical products, are based on the ten question principle defined by Yrjö Vartia in 1976 (Vartia, 1976, p. 93-95). The ten question principle is also presented in Statistics Finland’s Quality Guidelines for Official Statistics handbook, in the Section “Indices and indicators” (Statistics Finland, 2007).

The ten question principle is used to form a totality that fits the purpose of the index. The aim is to solve the questions before the design of the data processing. If there is need to change the selected strategy, it is convenient to come back to these definitions and re-evaluate them. This practice is very useful especially for the younger statistician who possess less experience in the area of index compilation.

The ten questions are grouped into following four sub-areas: 1) purpose of use of the statistics, 2) technical issues of the statistics, 3) index calculation methods, and 4) how quality changes, and new and disappearing commodities are treated in index calculation.

The questions relevant for each sub-area is presented in the list below followed by the replies that are used in current strategy of the CPI production and are in line with the EU HICP-regulation. These answers sets the starting point for the processing of pharmaceutical product data.

**10 Questions:**

The purpose for using the index should be decided by specifying:

1. the general nature of the commodities that are to be compared
2. the economic actors from whose perspective relative change will be measured
3. the lengths of the time periods over which relative change will be estimated

The technical problems of index calculation should be solved by specifying:

4. the classification that should be applied to the commodities that will be compared,
5. the method by which price data on these commodities will be collected,
6. the appropriate weight structure.

The index calculation methods should be decided by specifying:

7. the index formula,
8. the strategy for constructing the index series.

The special challenges

9. Quality changes in commodities
10. New and disappearing commodities

(Statistics Finland, 2007; Vartia, 1976)

**Answers:**

Following replies were defined and used in the compilation of pharmaceutical product indices in the initial calculations:

(1) The total prices for prescription and over-the-counter medicines are used in construction of statistics. The relevance of price data is clear – they are officially confirmed prices for consumers including value added tax.

The identification of the commodities is based on commodity-level unique identification code, VNR that separates products unconditionally. No relaunches and no discounts are present in this data.

The objective of the statistics is to measure the price change and price development of prescription and over-the-counter medicines sold in pharmacies in Finland. The calculation method is simplified by the fact that every pharmacy uses confirmed prices for prescription and over-the-counter medicines. In this situation, it becomes easier to measure the price change and development because no additional classifications are needed from the pharmacies for index calculation – data is sufficient for the consumer price index compilation.

(2) The primary perspective of the statistics is the consumer. The statistics are produced for the needs of the Consumer Price Index. However, these statistics also serve the needs of the data supplier, the Pharmaceutical Information Centre and Pharma Industry Finland.

(3) The indices for pharmaceutical products are constructed with a time span of one month.

(4) Commodities are classified according to the “European classification of individual consumption according to purpose” (eCOICOP) broken-down to 6-digit and 7-digit level. These most detail level

sub-classes are used mainly for national purposes.

There are six region in Finland that may be classified according to the Eurostat's NUTS2[6]-classification. The regional representativeness may be achieved because the data contain quantities by commodity and major region.

(5) The scanner data is obtained as a whole thus the data collection is not dependent on sampling methods. The obtained dataset is complete microdata in the sense that it contains prices and quantities by commodity and region. Therefore, it is possible to calculate relative changes in the prices and to estimate price development by commodity, major region and for the entire country.

(6) In the index calculation, the weights are constructed by commodity, based on the sales value. The index formula used is based on the requirements in EU Regulation, so the sales value from the previous year for uniquely identified commodities (VNR-code) is used in the index formula. In this case, when the index calculation is based on the aggregation of relative prices (log changes), the index weights are the relative value shares of the previous year by commodity in line with the presentation of Log-Laspeyres index formula.

(7) The formula for the price index for prescription and over-the-counter medicines is Log-Laspeyres.

(8) One possible strategy for the correct construction of index series is presented in the formula (5.1). This formula follows the international recommendations thus it is selected as a starting point in the initial calculations.

The formula (5.1) shows the construction of index series that is based on measuring the relative price changes of consecutive months separately for each VNR commodity of the weight reference year. These changes are chained together as in typical index calculation. A separate index series is formed for each commodity where matching-pair comparison is feasible. From these index series, the relative change may be estimated precisely for any time span.

The index calculation is based on analysis of panel data, where panel time series are constructed for two-year periods. First year is weight reference year while the second year is current year. Each panel weight reference year consists of different VNR commodities. Commodities for the panel are selected from those VNR-codes that are available in the data for the first two months in weight reference year.

The commodities in panel are analysed using the principle of matching-pair comparison. The relative changes of the prices of commodities are aggregated together with value share weights. The value share weights used in the index calculation correspond with the consumption shares of prescription and over-the-counter medicines in weight reference year. Weight reference period is changed every year in accordance with the EU HICP Regulation.

(9) There is no quality change problem in the index calculation of prescription and over-the-counter medicines. In other words, the commodities identified with VNR codes are fully identical over time and thus qualitatively comparable in paired comparison.

(10) The challenge of new and disappearing commodities is solved in the statistics as follows: In the beginning of new weight reference year a group of VNR commodities is selected to the sample. The aim is to utilise the whole data thus those commodities are selected to the sample for which the relative price changes for the first two months of the weight reference year may be calculated.

If the commodity disappears from the market, the missing price is estimated according to the overall mean imputation. Therefore disappearing commodities do not change the price index estimate and

[6] NUTS, Nomenclature of territorial units for statistics. Information in web-page http://ec.europa.eu/eurostat/web/nuts/overview

index weights do not have to be changed in any comparison. The disappearing commodities that leave the market, are withdrawn from the panel in the beginning of following year.

New commodities are taken into consideration as follows: New commodities are not included in the index calculation until they are selected as index commodities of the weight reference year. In other words, new products are taken into account only when product has been on the market in the beginning of weight reference year. The gap between launch to the market and inclusion to the sample is at furthest ten months.

## Practice no 2: Standardise Data Collection

The processing logic and methods presented in the method description of the index calculation are programmed into SAS programmes. This tool enable statistician to edit programming logic when necessary and to insert data analysis checks in order to find measurement errors. The acquired data before and after the processing (input and output data) is easily accessed through user interface for visual examination. Most of all the processing logic and formulas used in compilation are in readable text format that statistician understands.

The first task in data collection is to transfer data from internal server to the readable format that is then stored into the shared server. At this first step data is also pre-analysed in order to recognize defects, errors and serious faults in data collection. If errors appear in the data, instant feedback is given to the supplier and data is retransmitted to Statistics Finland.

The pre-analysis are executed with uniform process (sas-program) that imports received csv-file into SAS-system and compares data with the data description that is stored in common metadata-database. The same practice is used for all administrative data and from now on also for scanner data.

**Practical Example**

Below we have an example that shows structure of the original data, content of the data description and the process flow that is used to import and validate the data content. After the validation is performed the pre-analysis report is automatically sent by email to the owner of the data.

1. In the figure below, is an example of the sequential file that is received from the supplier. The information about one specific product is presented in a row and columns are divided with semicolon. No descriptive information is available thus it may be a challenge to understand the content of the data and the variables in it. This data here contain products that are identified with VNR-code and detailed price information for the month in question.

```
VNR;Date;status;PriceNoTax;PriceTax;PriceWholeSale;SubstitutionGroup;SubstitutionCode;ReferencePrice;PriceUpperLimit;
421180;2017-02-01;5;8,61;9,47;5,94;;;;;207;1;AEK, PK;1
137340;2017-02-01;5;2,25;2,48;1,55;0849;0008490100;3,48;3,48;110;1;PK, YEK;1
521789;2017-02-01;5;8,61;9,47;5,94;1082;0010820100;8,32;8,32;110;1;PK, YEK;1
558709;2017-02-01;5;17,31;19,04;12,14;1069;0010690001;19,04;19,04;;1;PK;1
421495;2017-02-01;5;3,81;4,19;2,63;0322;0003220020;4,69;4,69;115, 116, 117, 128, 130;1;PK, YEK;1
520647;2017-02-01;5;23,44;25,78;16,68;1069;0010690003;25,79;25,79;;1;PK;1
421636;2017-02-01;4;120,65;132,72;92,09;0224;0002240100;;;;0;;1
173653;2017-02-01;5;567,95;624,75;483,00;;;;;;0;EK;1
```

2. Before the data delivery may be set to work, statistician have to describe all the variables and data content that relates to the new dataset. These descriptions are called Data Description and those are stored in the common metadata database. In the figure below, is presented small part of the metadata content describing the variables in the received price data.

Hinnat ja kustannukset /Kuluttajahintaindeksi

Tiedoston nimi — Dataset name: /TKSAS/SASDATA/Tilastot/khi/Import//DWFIN_Prices.csv

Dataset format — Tiedoston formaatti: sequential

Delimiter — Erotinmerkki: ;

Tiedostokommentti

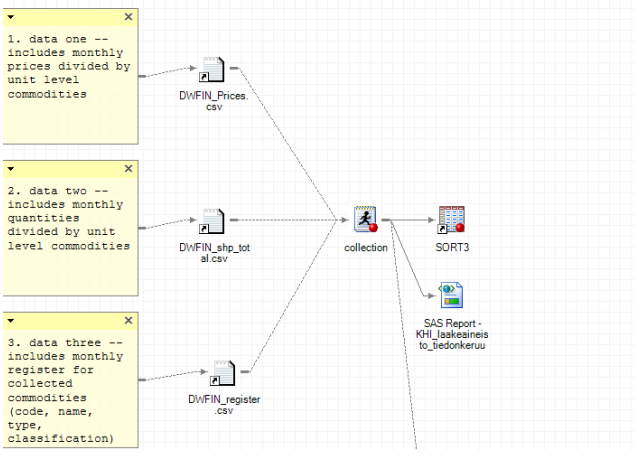
Muuttujia: 14 — Variable quantity

Havaintoja — Observation quantity

| Technical name / Tekninen nimi | Label / Muuttujan nimi | Group / Muuttujaryhmät | Data type / Tietotyyppi | Format / Esitysmu | Min / Minimiarv | Max / Maksimia | Values / Arvot-list | Length / Pituus | Alkupo | Puuttuva tieto (sallittu) | Puuttuv tietojen lkm |
|---|---|---|---|---|---|---|---|---|---|---|---|
| VNR | Product ID-number | Register;Prices; | character | | | | " | 6 | | no | |
| Date | Date | Register;Prices;Quantities;; | dateandtime | yymmdd10. | | | " | 10 | | no | |
| Status | Status | Prices; | numeric | | | | | 8 | | no | |
| PriceNoTax | Price without VAT | Prices; | numeric | | | | " | 8 | | no | |
| PriceTax | Price with VAT | Prices; | numeric | | | | " | 8 | | no | |
| PriceWholeSale | Wholesaleprice | Prices; | numeric | | | | " | 8 | | no | |

3. The process flow, that is used to convert csv-file and to pre-analyse the content, is very simple and easily understood for it is in SAS. Below the figure shows this flow.

   The flow starts with the yellow notes that shortly describe the obtained data. Then statistician follows the linking line from note onward getting to first file “DWFIN_Prices.csv”. At this point statistician may visually observe obtained data in text file format. Process continues to the first sas-program named “collection”. This is the point where processing rules are defined and executed. Statistician may examine these rules and assure that the data is treated according to the strategy plan. After the “collection” program is executed, two outputs may be identified in the flow: dataset and pre-analysis report.




If the execution fails, a clear indication of failure is marked to the flow. Statistician may assess the causes for the failure of execution by reading through the log-file.

4. The dataset that is generated contains the imported data in the raw format. Only slight modifications to the imported variables are allowed. These modifications are done according to the metadata description. One example of most commonly used modification is date-variable that is edited into human readable format with specific format (see column Format in figure above). In the figure below a snap-shot of converted file is visible.

| | vnr | date | status | pricenotax | pricetax | pricewholesale | substitutiongroup | substitutioncode | referenceprice | priceupperlimit | reimbursementnumber | compensation | reimbursementcodes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 000005 | 2008-01-01 | 5 | 21.55 | 23.27 | 14.37 | 945 | . | 0 | 0 | | 0 | EK |
| 2 | 000019 | 2008-01-01 | 5 | 12.02 | 12.98 | 7.68 | 412 | 4120010 | 8.64 | 10.63 | | 0 | EK |
| 3 | 000051 | 2008-01-01 | 5 | 9.37 | 10.12 | 5.91 | 95 | 950100 | 4.65 | 6.64 | 201.205 | 1 | AEK. PK |
| 4 | 000052 | 2008-01-01 | 5 | 59.32 | 64.07 | 41.35 | 144 | 1440100 | 64.07 | 67.06 | | 0 | EK |
| 5 | 000054 | 2008-01-01 | 5 | 321.69 | 347.43 | 254.62 | . | . | 0 | 0 | 138.306 | 0 | YEK.LRPK |
| 6 | 000066 | 2008-01-01 | 5 | 20.29 | 21.91 | 13.47 | 144 | 1440030 | 21.91 | 23.9 | | 0 | EK |
| 7 | 000082 | 2008-01-01 | 5 | 14.51 | 15.67 | 9.34 | 203 | 2030100 | 14.98 | 16.97 | | 1 | PK |
| 8 | 000093 | 2008-01-01 | 5 | 9.29 | 10.03 | 5.86 | 203 | 2030060 | 10.03 | 12.02 | | 1 | PK |
| 9 | 000096 | 2008-01-01 | 5 | 6.77 | 7.31 | 4.18 | 95 | 950030 | 4.3 | 6.29 | 201.205 | 1 | AEK. PK |
| 10 | 000097 | 2008-01-01 | 5 | 8.17 | 8.82 | 5.11 | 97 | 970100 | 4.81 | 6.8 | 201.205 | 1 | AEK. PK |

5. The analysis report may be examined in SAS (report icon in process flow) or statistician may investigate it from the pre-analysis report that was automatically delivered by email after execution of the program. The pre-analysis report shows results for the checks that were carried out during the conversion of the csv-file. These checks are commonly defined thus same checks are used for all received data. To name some of these checks there are observation count, key-figures for numerical variables, frequencies for character variables, check for duplicate values and check for purposeless classification values.

**Source Data: /TKSAS/SASDATA/Tilastot/khi/Import//DWFIN_Prices.csv**
**Pre-analysis report based on the data description:**

| Observation count |
|---|
| 10 106 |

**Key figures for numerical variables**

| Obs | variable | variablename in Finnish | obs | missing | mean |
|---|---|---|---|---|---|
| 1 | date | Tietueen päivämäärä | 10 106 | 0 | 20 910.00 |
| 2 | pricenotax | Vähittäismyyntihinta, veroton | 9 998 | 108 | 237.03 |
| 3 | … | | 9 998 | 108 | 260.74 |
| 10 | substitutiongroup | Substituutioryhmä | 5 582 | 4 524 | 968.79 |

**Character variable frequencies**

| Obs | variable | variablename in Finnish | obs | missing |
|---|---|---|---|---|
| 1 | compensation | Tieto korvattavuudesta | 10 106 | 0 |
| 2 | reimbursementcodes | Kela-korvattavien läkkeiden korvausnumerot koodeina | 9 788 | 318 |
| 3 | reimbursementnumber | Kela-korvattavien läkkeiden korvausnumerot | 3 513 | 6 593 |
| 4 | vnr | Tuotteen yksilöintitunnus | 10 106 | 0 |

**Check of classification values**

| Compensation code | | | | |
|---|---|---|---|---|
| reimbursementcodes | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| AEK. LRPK | 38 | 0.39 | 38 | 0.39 |
| AEK. PK | 1372 | 14.helmi | 1410 | 14.41 |
| AEK. PK. YEK | 86 | 0.88 | 1496 | 15.28 |
| EK | 4805 | 49.09 | 6301 | 64.37 |

… report continues.

This report enables statistician to check whether the imported data contain relevant information or not. There is no need to visually check all the imported data on observation level instead just control the pre-analysis report. When working with massive amounts of data these reports compress essential information into readable format because otherwise it would be very time-consuming and difficult to gather similar information by visually browsing the output dataset in SAS.

## 5. Compilation of elementary aggregates

The calculation of the Harmonised Indices of Consumer Prices (HICP) is directed by the EU HICP framework regulation and its' 21 implementation regulations thus the index calculation is based on these guidelines. Same principles are used for the calculation of CPI that is compiled mainly for national purposes.

The aim is firstly to compile product level indices and secondly to compile these lowest level index series together with the relative value share weights.

Collection phase was presented in practice no 2 and here we will take a look at processing phase where all the essential work is done. This part is divided into two sections: At first is the short explanation on how the sample is drawn and panel data created and then the short summary of the data content. Finally the compilation of initial elementary aggregates is summarised.

### Introduction

The compilation phase is divided into four sub-steps:

1. Forming basic data,
2. Creating sample of VNR commodities for the weight reference year,
3. Constructing two year panel data for the weight reference year and current year,
4. Index calculation and chaining.

The first step contain small modifications that are carried out in order to generalize the programming logic. Variables are renamed to uniform names. Also the products are classified according to the eCOICOP-classification. To classify products is easy because obtained data contain enough information that may be utilised in the class value deduction.

The second step is to draw sample out of weight reference year VNR-commodities. The sample is withdrawn from the products sold in January and February. The method is very simple but because the attrition rate is low for pharmaceutical products, this simplification may be approved.

The third task is to compose two year panel that is used to control products replacements. New products are taken into account only when product has been on the market for almost a year. Outgoing products that leave the market, are withdrawn from the panel at the end of current year. The price change for the missing items are imputed according to the overall mean. Therefore, new and disappearing products do not have an effect on the compiled price indices.

It would be natural to link data editing with the index calculation so that the relative price changes and, in particular, the index series constructed for VNR index commodities may be used to assist in editing. Editing of statistical data has not been implemented because there is no need for data editing. However this may be done in the existing SAS application, as long as the use cases for data editing are defined by CPI experts in Finland.

Table 1 shows the number of individual commodities by COICOP 7-digit groups selected for year 2016 sample. Also the comparison to the traditional sample size that was valid before year 2017 is shown. It is well known fact that NSI's are forced to cut sample sizes to minimize collection costs and response burden so it is not surprise that these quantities differ this much.

Table 1. Number of individual commodities taken into the index compilation, by COICOP 7-digit commodity group year 2016

| code | Name of commodity | Commodities in scanner data, N | Commodities in traditional data collection, N |
|---|---|---|---|
| 06.1.1.0.1.2 | Non-refundable prescription medicines | 1574 | 28 |
| 06.1.1.0.2.1 | Over-the-counter medicines | 670 | 32 |
| 06.1.1.0.5.1 | Oral contraceptives | 119 | 8 |

As it may be noted from the table above refundable prescription medicines are extracted from the index compilation. The reason for this is that there is complicated compensation system used in Finland for refundable prescription medicines. It is not possible at this point compile index series for refundable prescription medicines with this data that would take account achieved compensations.

The fourth task is the index calculation that is described more closely in next sub-chapter “Index calculation”.

### Index calculation

In the initial index compilation for the purpose of CPI and HICP, the strategy used was introduced earlier in this document in chapter 4. Practices. Therefore here is only recap to the decision made in chapter 4.

The commodities identified with VNR codes are fully identical over time and thus qualitatively comparable in matching-pair comparison. The price for the base period is the average price of the VNR commodity in the weight reference year, and in the comparison period it is the total consumer price of the month in question.

The compilation method is based on matching-pair comparison of the prices of the price reference and comparison periods. The prices of commodities are not aggregated to COICOP-7-digit level as a first step, instead the relative change in prices is estimated for each commodity. This means that relative price change is calculated for each commodity that may be identified with unique ID-number.

The specific index series may be constructed for each commodity based on these relative changes. The index series show the price development of the commodity from the first month of the base year to the last month of the following year.

After the relative price change is calculated for each identified commodity, these commodity-specific relative changes are aggregated together using index weights that are calculated for each commodity for year *t-1* broken down by eCOICOP-sub-class levels. The formula for the price index for medicines is Log-Laspeyres, which is defined as

$$(5.1)\ \exp\left\{\sum_i w_{i0} \ln\left(p_{it} / p_{i0}\right)\right\},$$

where $w_{i0}$ is the index weight of the VNR commodity that is the commodity’s value share of consumption in the weight reference year, $p_{i0}$ its corresponding price for reference period and $p_{it}$ its corresponding price for the current month.

## 6. Results

This part of the document describes the results from the initial calculations and results from further researches carried out by Antti Suoperä and Yrjö Vartia. Hence, it need to be noted that three separate calculations sets were accomplished in 2016-2017.

First we take a look at the results from initial calculations where the strategy that was demonstrated in chapter 4. Practices was used. This strategy follows current practices in CPI-production. These results were thoroughly examined and based on the findings further research was decided to carry out.

Since obtained dataset was seen to be a complete data in the sense that it contains prices, quantities and register information from the pharmaceutical products sold in Finland, it was decided to test superlative index formulas. These results are shortly demonstrated in chapter 6.2 . For further investigation of the used methods and selected strategy may be read from the summary *Index number theory and construction of CPI for complete micro data* by Yrjö Vartia and Antti Suoperä.

After the tests introduced in chapter 6.2, Eurostat recommended us to check whether the chain-drift exist or not in pharmaceutical products data for there had been evidences of chain drift in scanner-data of daily products (Nygaard, 2010;Nygaard, 2011; de Haan & van der Grient, 2009). These tests were carried out and the results are demonstrated in chapter 6.3. It is proved that there may appear a chain drift depending on the selected compilation strategy and the commodity group concerned. Further information about the methods used and the conclusions may be read from the document *Analysis of Change and Constructing Index Series for Complete Micro Data* by Antti Suoperä, Satu Montonen, Kristiina Nieminen.

## 6.1 Final results

Let us first examine the descriptive characteristics of the Pharmaceutical Information Centre's data in a simplified manner. Table 2 presents the numbers of non-refundable prescription medicines, over-the-counter medicines and oral contraceptives in the panel data for different weight reference years in 2009 to 2016. The number of index commodities have grown over time in all groups. For demonstrative purposes only years 2009, 2010, 2015 and 2016 are presented.

Table 2: Number of individual commodities by COICOP-group in 2009 to 2016.

| year | Non-refundable prescription medicines | Over-the-counter medicines | Oral contraceptives |
|---|---|---|---|
| | N | N | N |
| 2009 | 717 | 423 | 49 |
| 2010 | 909 | 460 | 60 |
| … | | | |
| 2015 | 1477 | 651 | 100 |
| 2016 | 1574 | 670 | 119 |

Correspondingly, Table 3 presents the price development of corresponding medicines in 2009 to 2016. The first period 2009/1 equal one (= 1). Only for demonstrative purposes point figures from following years are limited to January while year 2016 is show for January and December.

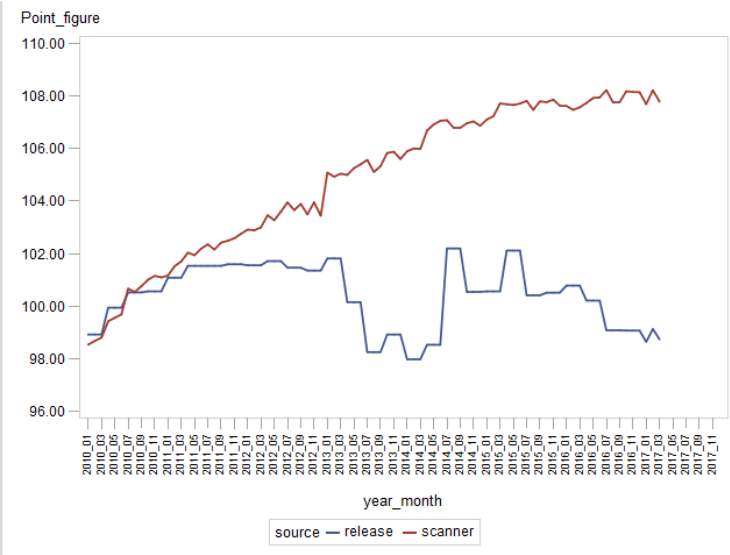
Table 3: Price development of non-refundable prescription medicines, over-the-counter medicines and oral contraceptives in the whole country in 2009 to 2016

| | | | | |
|---|---|---|---|---|
| | | Non-refundable prescription medicines | Over-the-counter medicines | Oral contraceptives |
| year | month | | | |
| 2009 | 1 | 1 | 1 | 1 |
| 2010 | 1 | 0,991 | 1,024 | 1,012 |
| 2011 | 1 | 0,982 | 1,051 | 1,068 |
| 2012 | 1 | 0,981 | 1,07 | 1,096 |
| 2013 | 1 | 0,986 | 1,092 | 1,096 |
| 2014 | 1 | 0,922 | 1,101 | 1,038 |
| 2015 | 1 | 0,925 | 1,113 | 1,031 |
| 2016 | 1 | 0,919 | 1,118 | 1,017 |
| 2016 | … | | | |
| **2016** | **12** | **0,917** | **1,124** | **1,002** |

The prices of prescription medicines have decreased by close on 8,5 per cent between 2009/1 and 2016/12. The downward trend in prices is explained by a typical feature for prescription medicines: New medicines are priced high until the competitive and technological advantage is lost over time and prices start to drop. For over-the-counter medicines prices have, in turn, grown by almost 12.5 per cent between 2009/1 and 2016/12.

When compared these new results with the published results quality improvement may be confirmed. In the figure 1 the difference between these two series is clear.

Figure 1. Comparison of price development in the commodity group: Over-the-counter medicines, series 2010=100




Blue line in figure 1 represents the price development of 32 products selected to the sample (traditional method), while red line shows the price development for all over-the-counter medicines in the market (new method). It is obvious that reason for this discrepancy lies in the sample size that has been far too small. The sample sizes by product were presented in table 1. All new index series show similar quality improvement in all three commodity groups: over-the-counter medicines, oral contraceptives and non-refundable prescription medicines.

## 6.2 Results from the superlative index formula tests

Different index formulas were tested in co-operation with the University of Helsinki/Department of Political and Economic Studies (professor Yrjö Vartia). Same data as in previous calculations was used in these tests. Data was taken as whole and no extraction was made before the calculations.

This chapter summarize the results from co-operation work thus all statements in this chapter are direct quotes from the research documentation. The detailed working document of the research and its' results is *Index number theory and construction of CPI for complete micro data* (Vartia & Suoperä, 2017).

### Background

This is a joint work of the authors Vartia and Suoperä that continues the co-work of University of Helsinki and Statistics Finland related more generally to aggregation of behavioral functions, see Vartia (2008) and Suoperä and Vartia (2011).

In this analysis, most popular index number formulas were analysed. Vartia & Suoperä showed that index number formulas may be classified to formulas having upward or downward bias for small changes compared to index formulas, that must be classified as 'unbiased' according to the index number theory. It is known that

quantity information is typically available only for the base period thus Laspeyres formula has been taken as a practical solution in CPI compilation. On the contrary “Big Data” opens up new potentialities to use other, more precise index formulas in CPI compilation. When the number of commodities increase 50-fold and also the quantities are observed, the index calculations must change accordingly.

**Methods used in analysis and compilation**

The index formulas that were used in this analysis are Fisher, Törnqvist, Stuvel, Diewert, Sato & Vartia, and Montgomery & Vartia. No questionable assumptions of “economic behaviour” were used in this analysis.

The data used, is pharmaceutical product data (i.e. over-the-counter medicines). Dataset is *complete micro data* in the sense that it contains all prices and quantities (aggregated for the whole country) for all homogeneous commodities and periods, that are months in this analysis.

Perspective taken in this analysis was to concentrate on results and proposals that are applicable for similar CPI data having up to 100 times larger sets of commodities than in the current national CPI practices. Idea is to treat new and disappearing commodities in a systematic and simple way so that quality corrections are unnecessary when the commodities are homogeneous in commodity group concerned.

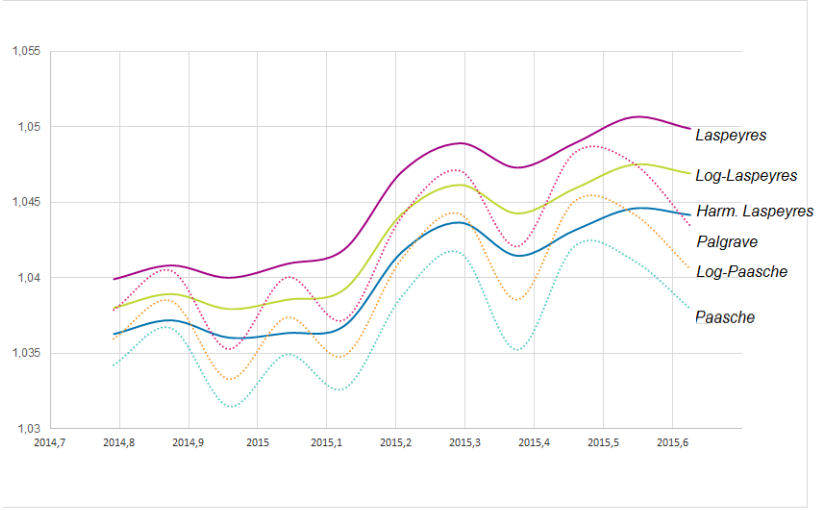
In the calculations, data was split in two by five different ways. In this report we will only investigate results from the one of these five ways. Split was done according to how much the values of the VNR-commodities have changed between base and observation periods. In the splitting, first part named 5S, consists of the VNR-commodities, whose values had larger relative change (up by multiple 5 or down by 1/5). Second part named 5N, is the complement of 5S where values of commodities stay relatively constant.

**Results**

Irving Fisher (1922) visualized the index numbers as forks containing a certain number of tines (Fisher Five-tined fork). According to Vartia & Suoperä this is an important visualization of differences between various indices and other choices of calculation. Hence next is presented three price indices based on arithmetic, geometric and harmonic means of price relatives (A, G, H) and old value shares followed with three price indices based on new value shares.

The three-tined fork having base period weights composes of indices Laspeyres, Log-Laspeyres, Harmonic-Laspeyres while with weights of observation period three-tined fork is composed of Palgrave, Log-Paasche, Paasche. Order of the three weighted averages is A, G, H in both example. These six index series are displayed in figure 2.

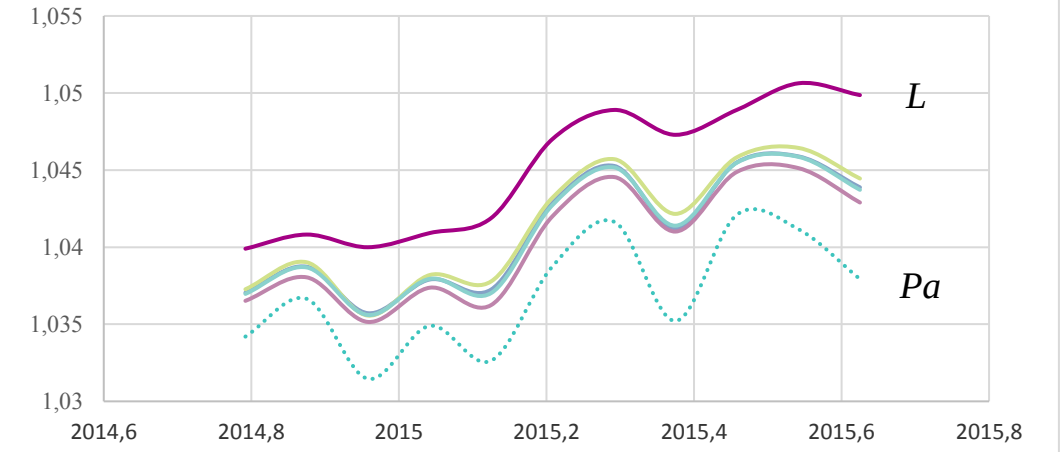
Figure 2. The Six-tined fork represented by Vartia and Suoperä



As it may be noticed from this example composition of these two forks is clear. Order of the arithmetic, geometric and harmonic mean is always the same, and so also in this case, but the composition of two forks varies depending on the data used. Here the Laspeyres-type fork with old weights is on top while Paasche-type fork with new-weights takes place below it. Laspeyres (*L*) on the top, Paasche (*Pa*) is the lowest index index in figure and all other basic indices remain between *L* and *Pa*.

This data of over-the-counter medicines showed that Laspeyres and Paasche had large biases up and down, respectively. Due to this superlative index formulas were tested with the same data. The superlative formulas Fisher (*F),*Stuvel *(Stu),* Montgomery-Vartia (*MV)* and Törnqvist *(t)* form a much tighter band within the six indices, figure 3.

Figure 3. Results from the tests of superlative index formula by Vartia and Suoperä




These results indicates that when complete datasets are available, superlative indices produce more precise series than traditionally used Laspeyres- and Paashe-type index formulas. Suitable superlative index formulas are for example Fisher, Törnqvist, Stuvel, Montgomery-Vartia, Diewert, Drobish, Walsh, Edgeworth. Beside the decision of index formula the definition of index compilation strategy includes the selection between base and chain methods.

Inclusion of complete micro data in CPI's will greatly change and simplify the current practices based on Laspeyres formula and complicated rules for elementary aggregates. When the number of commodities increases 50-fold and also the quantities are observed, the index number calculations must change accordingly.

We believe that using these new data sets and improved index number methodology, reliability and accuracy of CPI production will be raised to a new level.

## Results from the analysis of chain drift in pharmaceutical product data

### Aim in this research

As Norway and Netherland (Nygaard, 2010; de Haan & van der Grient,2009) have proved, the chained indexes may suffer from what is known as chain drift or chain link bias. The chain drift occurs if a chained index "does not return to unity when prices in the current period return to their levels in the base period" (ILO, 2004, p. 445). When there are large period-to-period fluctuations in prices, quantities and values, all kinds of chained price indices should be avoided, because the chain drift may occur. The base method never suffer from the chain drift whatever the index number formula is.

Therefore it is important to investigate pharmaceutical product data further and to control the existence of chain drift. Hence aim in this calculation round was 1) to analyse existence of the chain drift and 2) to construct new method that effectively eliminates the chain drift phenomenon. More detailed documentation of this research and its' results is available in *Analysis of Chain Drift for Complete Micro Data.*

### Methods used in analysis and compilation

The Törnqvist formula is used in our analysis by four different ways. First, *base* Törnqvist (commodity set $\{a_1, a_2, \dots, a_n\}$ excluding new and disappearing commodities) is applied. This method is selected for the chain drift never exists in base method. Second, the same set of commodities as in base method is selected and *chain in isolation* Törnqvist is applied for them. Third, *proper chain* Törnqvist is applied for the data including maximum number of matched pairs. Forth, we define *mixed* method, where we combine together the base method (first method) and typical price change calculation for new and disappearing commodities by Törnqvist formula. In the table 4 these different methods are summarised.

Table 4. Four methods that were used in chain-drift analysis

| Method | Formula | Sample strategy |
|---|---|---|
| Base Törnqvist (1) | $t_{Base}^{t/0} = exp\left\{\sum \frac{1}{2}(w_i^0 + w_i^t)\log(p_i^t / \bar{p}_i^0)\right\}$ | commodity set $\{a_1, a_2, \dots, a_n\}$ excluding new and disappearing commodities |
| Chain Törnqvist (2) | $t_{Chain}^{t/(t-1)} = exp\left\{\sum \frac{1}{2}(w_i^{t-1} + w_i^t)\log(p_i^t / p_i^{t-1})\right\}$ | commodity set $\{a_1, a_2, \dots, a_n\}$ excluding new and disappearing commodities |
| Chain Törnqvist (3) | $t_{Proper\ chain}^{t/(t-1)} = exp\left\{\sum \frac{1}{2}(w_i^{t-1} + w_i^t)\log(p_i^t / p_i^{t-1})\right\}$ | Maximum number of matched pairs in base and observation periods |
| Mixed Törnqvist (4) | In next row, below | All commodities except new and disappearing (base Törnqvist) + new and disappearing (price ratio) |
| $t_{Mixed}^{2/1} = exp\{\frac{1}{2}(w_{Base}^1 + w_{Base}^2) log t_{Base}^{2/1} + \frac{1}{2}(w_{N\&D}^1 + w_{N\&D}^2) log t_{Chain,N\&D}^{2/1}\}$ | | |

In (1) the base period 0 is previous year and *t* is observation month in current year. We compare arithmetic mean prices $\bar{p}_i^0$ (i.e. quantity weighted) and current month prices $p_i^t$ and the index weights are arithmetic mean of base and observation periods value shares $w^0$ and $w^1$. We change the weights every January. Index series with a different base year are chained by transforming each single base year index series to begin from December in base year (i.e. $t_{Base,Dec} = 1$). In base method, like in this example, chain drift never exists.

In (2) these *measured* price changes can be easily chained to get index series based on chained method.

In (3) the price changes are calculated by (2) and they are chained together to get the index series for prescription and over-the-counter medicines. In this method we include new and disappearing commodities in the most dynamic way. The $t_{Proper\ Chain}$ includes set $\{a_1, a_2, \dots, a_n\}$ and new and disappearing commodities.

In mixed method (4), we combine the *derived* [7]price change $t_{Base}^{2/1}$ and the *measured* price change $t_{Chain,N\&D}^{2/1}$ estimated by (2), that is

(4) $$t_{Mixed}^{2/1} = exp\left\{\frac{1}{2}(w_{Base}^1 + w_{Base}^2)logt_{Base}^{2/1} + \frac{1}{2}(w_{N\&D}^1 + w_{N\&D}^2)logt_{Chain,N\&D}^{2/1}\right\}.$$

$w_{Base}^t$ and $w_{N\&D}^t$ are value shares for commodity set $\{a_1, a_2, \dots, a_n\}$ and for new and disappearing commodities respectively. These value shares sum up to unity i.e. $\frac{1}{2}(w_{Base}^1 + w_{Base}^2) + \frac{1}{2}\left(w_{N\&D}^1 + w_{N\&D}^2\right) = 1$. The index series for mixed Törnqvist is easily derived by chaining the price changes properly.

We use these four methods or strategies to reveal the existence of chain drift. These methods are compared in the following order:

- First we compare base to chain Törnqvist in isolation for the same set of commodities. Differences between base and chain index series for the same set of commodities reveal the chain drift.
- After that we compare the *chain Törnqvist in isolation* (i.e. for commodity set $\{a_1, a_2, \dots, a_n\}$) to *the proper chain* Törnqvist allowing maximum number of matched price pairs in proper chain analysis. This will give information about price changes between commodity set $\{a_1, a_2, \dots, a_n\}$ and maximum number of matched price pairs.
  In case of these two chained methods have different price developments, these differences will be accounted to new and disappearing commodities.

If the base and both chain methods differ from each other, we need a *mixed* method or strategy to solve weak points of other three methods. Especially when the value shares of new and disappearing commodities are stable in time, the mixed method is superior compared to the other methods.

**Results**

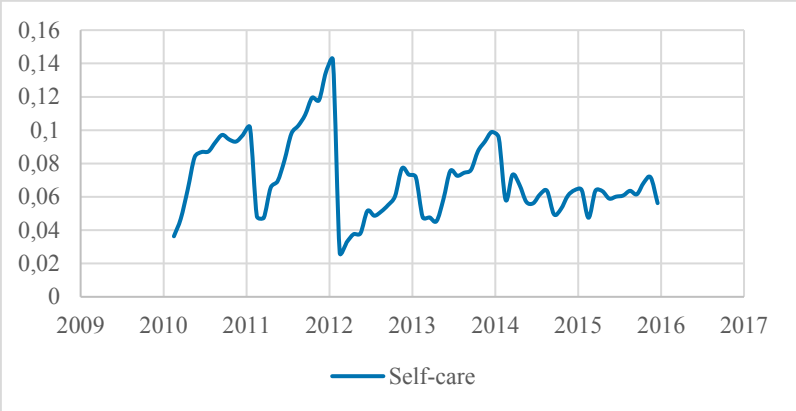
Let us first examine the changes in value shares of new and disappearing commodities. This analysis will tell us when the launch of new products realises and how new commodities impact on the total value share.

As we may notice in figure 4, for any base year the value shares (Törnqvist) of new and disappearing commodities are in the beginning of year about 2 – 4 percent increasing during a year about 2 – 9 percent. The year 2011 is unusual, but in general the value shares of new and disappearing commodities (in commodity group pharmaceutical products) changes quit moderately compared to the example in Nygaard (2010) and de Haan & van der Grient (2009) cases.

[7] For mixed method, we need to *derive* the price change for the base Törnqvist for all $i, t = 1,2, \dots$ For simplicity we analyze only base (i.e 0) and two other periods, say $t$ =1, 2 and the *derived* price change $logt_{Base}^{2/1} = logt_{Base}^{2/0} - logt_{Base}^{1/0}$ in the logarithmic form is

$$logt_{Base}^{2/0} - logt_{Base}^{1/0} = \sum \tfrac{1}{2}(w_i^0 + w_i^2)\log\left(p_i^2/\bar{p}_i^0\right) - \sum \tfrac{1}{2}(w_i^0 + w_i^1)\log\left(p_i^1/\bar{p}_i^0\right)$$

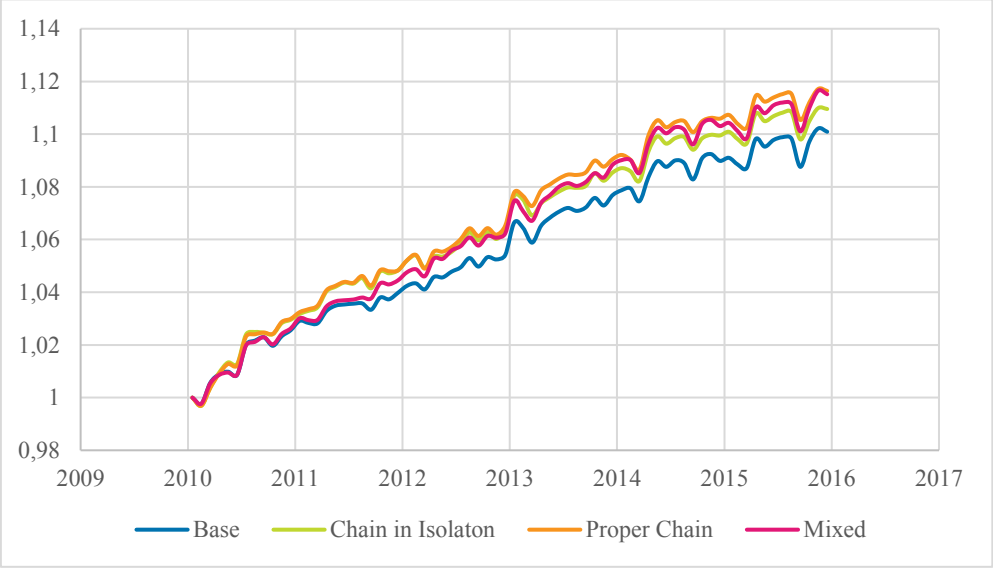
Figure 4. Over-the-counter-medicine value shares for new and disappearing commodities




During the years 2014 and 2015 the value shares keep in 6 percent level almost whole-year. Increase of value shares are mainly explained by increase of number of new commodities and not by increase of values (i.e. expenditures) of a single commodity. Instead in de Haan & van der Grient (2009) case, problems arise from very dramatic changes of quantities (i.e. more than 10 times increases and declines), expenditures and value shares of a single commodity.

Figure 5 represent index series for base (i.e. for commodity set $\{a_1, a_2, \dots, a_n\}$), chain in isolation (i.e. for commodity set $\{a_1, a_2, \dots, a_n\}$), proper chain (i.e. maximum number of matched pairs) and mixed methods.

Figure 5 Comparison between alternative methods used with Törnqvist index formula for over-the-counter medicines, 2010-2016




The difference between base and chain in isolation can be seen in mid of the years 2010 and 2011. This comparison shows the impact of chain-drift. So it's true – slight chain drift happens but here it is not as remarkably as was seen in researches by Norway and Netherland. Chain-drift corresponds about one percent increase of prices during six years compared to the base method.

Comparing chained index series (i.e. chain in isolation and a proper chain), we notice that new and disapering commodities increases price changes about one percent during six years. Figure shows also that there are 72 periods (month/year) where price changes are repeatedly the same between these four methods.

The Figure 5 tells that the base Törnqvist is downward biased. That happens because new and dissapearing commodities are excluded from compilation. We suggest using a *superlative mixed method*, because the base Törnqvist under estimates the price changes and both chain methods include a slight chain drift.

## Conclusions

This chapter is brief summary of the results and the conclusion made.

It has been proven that new scanner data is more comprehensive than ever offering NSI's to evaluate different compilation strategies. Scanner data includes up-to-date information about the prices (incl. tax), quantities sold in previous month and additional register based information about the pharmaceutical products. This kind of data opens up new approaches that may be taken when compiling the elementary aggregates.

Following commodity groups were implemented into the production of CPI and HICP from the beginning of year 2017:

| code | Name of commodity |
|---|---|
| 06.1.1.0.1.2 | Non-refundable prescription medicines |
| 06.1.1.0.2.1 | Over-the-counter medicines |
| 06.1.1.0.5.1 | Oral contraceptives |

No revisions backwards were made due to the contractual commitment in CPI in Finland.

CPI-weight, for commodity group 06.1. Pharmaceutical products, is 18,52 ‰ from the total expenditure in 2017. Value share is very small and there may arise questions why to bother putting so much time and effort to research this commodity group.

We see that especially at the starting point, when there is very little competence and experience for processing big data it seems reasonable to start with "small scale" big data and to enhance statisticians competence to work with new data sources. In the future less and less data is examined and validated at observation level instead thousands of observations are processed as a group. Current practices ought to be re-evaluated and improved by taking each process phase and step under consideration one after another.

It is also shown that superlative index formulas produce more accurate index series than Laspeyres-type index

Therefore we suggest that a *superlative mixed method* may be a solution to the named challenges. *The superlative mixed method* combines base superlative index strategy and complements it with chain superlative strategy where new and disappearing commodities are taken account instantly. This way the effect of chain drift may be minimised yet utilising the benefits of superlative index formulas.

Statistics Finland will continue the research on using scanner-data in CPI-production and to further investigate the proposal to use a superlative mixed method in index compilation. Next data sources that will be taken into testing are 1) the daily products data obtained from the major retail chain, 2) the alcoholic beverages obtained from monopoly owner and 3) the hardware store data obtained by web-scraping the internet-pages of some big chain stores. Also the co-operation between Yrjö Vartia and Antti Suoperä continues.

Finally we want to encourage all other NSIs to test proposed mixed method and to give feedback on the findings.

## References

Ylä-Jarkko, Mari (2001) *Final report of scanner data research*, Helsinki:Statistics Finland, (internal report)

Koskimäki, Timo & Saranpää, Tuukka (2016), *Final report on implementation of the action Grant agreement 04151.2014.002-2014.422*, Helsinki:Statistics Finland, (internal report)

ILO, International Labour Office (2004), *Consumer Price Index Manual: Theory and Practice,* expanded version of "Consumer Price Indices: An ILO manual" (1989)

Eurostat (2015), *Draft recommendations on obtaining scanner data,* June 2015

Eurostat (2016), Legislation related to Harmonised Indices of Consumer Prices, Available from: http://ec.europa.eu/eurostat/web/hicp/legislation

Vartia, Yrjö (1976), *Relative changes and index numbers,* The research Institute of the Finnish Economy, Serie A 4, Helsinki, Finland

Statistics Finland (2007), *Quality Guidelines for Official Statistics*, Handbook 43B, 2nd edn., Helsinki:Multiprint, Available from http://tilastokeskus.fi/meta/qg_2ed_en.pdf#_ga=1.56757484.716102791.1478086303 [Accessed 30th March 2017]

Nygaard, Ragnhild (2010): *Chain Drift in a Monthly Chained Superlative Price Index*, Statistics Norway.

Nygaard, Ragnhild (2011): *Dealing with bias in Norwegian superlative price index of food and non-alcoholic beverages*, Statistics Norway.

De Haan, Jan & van der Grient, Heymerik (2009), *Eliminating Chain Drift in Price Indexes Based on Scanner data*, Statistics Netherlands

Vartia, Yrjö and Suoperä, Antti (2017), *Index number theory and construction of CPI for complete micro data*

Suoperä, Antti (2017), *Analysis of Change and Constructing Index Series for Complete Micro Data*

Appendices

Appendix 1. The structure of scanner-data concerning pharmaceutical products

| Source | Name | Type | |
|---|---|---|---|
| Register | *Pharmarket_ID* | Character | Identification code of packages (the same as the VNR number for pharmaceuticals, own coding for non-pharmaceuticals) |
| Register | *VNR* | Numeric | Identification code of product |
| Register | ProductName | Character | Identification code of packages (the same as the VNR number for pharmaceuticals, own coding for non-pharmaceuticals) |
| Register | ProductInformationIncludingName | Character | Long name of product including product data |
| Register | MarketingCompany | Character | Enterprise that markets the product |
| Register | UnitsPerPackage | Numeric | Number of units in package |
| Register | Strength | Character | Strength |
| Register | DosageForm | Character | Form of medication |
| Register | PackageSize | Character | Package size |
| Register | ATC5_Description | Character | Anatomical Therapeutic Chemical Classification System of medication defined by the WHO. |
| Register | Date | Date | Date of the data |
| Prices | *VNR* | Numeric | Identification code of the product |
| Prices | Date | Date | Date of the data |
| Prices | status | Numeric | Status of the product's sales permit |
| Prices | PriceNoTax | Numeric | Retail price of package excluding VAT |
| Prices | PriceTax | Numeric | Retail price of package including VAT |
| Prices | PriceWholeSale | Numeric | Wholesale price excluding VAT |
| Prices | SubstitutionGroup | Numeric | Substitution group of interchangeable medication |
| Prices | SubstitutionCode | Numeric | Substitution code of interchangeable medication |
| Prices | GenericSubstitution | Character | Generic substitution |
| Prices | ReferencePrice | Numeric | Reference price confirmed by the Pharmaceuticals Pricing Board used as basis for drug reimbursement |
| Prices | PriceUpperLimit | Numeric | Maximum price |
| Prices | ReimbursementNumber | Character | Reimbursement numbers of reimbursable drugs |

| | | | |
|---|---|---|---|
| Prices | Compensation | Character | Data on reimbursement, 1 = reimbursable 0 = not reimbursable |
| Prices | ReimbursementCodes | Character | Reimbursement numbers of reimbursable drugs as codes |
| Sales | year | Character | Statistical reference year |
| Sales | month | Character | Statistical month |
| Sales | shp_code | Character | Identification code of hospital district |
| Sales | Pharmarket_ID | Character | Identification code of packages (the same as the VNR number for pharmaceuticals, own coding for non-pharmaceuticals) |
| Sales | unit | Numeric | Quantity |
| Sales | SalesChannelCode | Character | Identification code of sales channel, 1 = pharmacy and retail trade 2 = hospital |
| | Region | Character | Major region |

*s*