This review evaluates the methodological quality of current front-of-pack labeling research and discusses future research challenges. Peer-reviewed articles were identified using a computerized search of the databases PubMed and Web of Science (ISI) from 1990 to February 2011; reference lists from key published articles were used as well. The quality of the 31 included studies was assessed. The results showed that the methodological quality of published front-of-pack labeling research is generally low to mediocre; objective observational data-based consumer studies were of higher quality than consumer studies relying on self-reports. Experimental studies that included a control group were lacking. The review further revealed a lack of a validated methodology to measure the use of front-of-pack labels and the effects of these labels in real-life settings. In conclusion, few methodologically sound front-of-pack labeling studies are presently available. The highest methodological quality and the greatest public health relevance are achieved by measuring the health effects of front-of-pack labels using biomarkers in a longitudinal, randomized, controlled design in a real-life setting.

© 2012 International Life Sciences Institute

INTRODUCTION

Front-of-pack (FOP) nutrition labels are present on food products worldwide.1–4 These labels are developed as interpretational aids to supplement the back-of-pack nutrition facts panel. Some of these labels are assigned to products that have reduced levels of saturated fat, sugar, salt, or calories. The World Health Organization recommends limiting the intake of these nutrients to reduce the prevalence of diet-related chronic diseases.5 Although the present FOP labels have different designs, different product criteria, and different developers, they generally have the same two aims: to assist consumers in making healthier food choices, and to stimulate food manufacturers to produce healthier products.

In recent years, there has been an international debate about the preferred format and potential impact of front-of-pack nutrition labeling. Regulatory changes are currently being considered by the European Parliament.6,7 In July 2011, the Parliament voted in favor of new labeling legislation, which prescribes mandatory display of the nutrition facts panel but which has no FOP requirement. Regulatory bodies in Australia and New Zealand are likewise considering changes in labeling.8,9 The Institute of Medicine and the Food and Drug Administration in the United States are also currently conducting research in this area.1,10 In this highly political debate, policymakers, scientists, industry groups, and consumer organizations are looking for evidence-based information on the effects of FOP labels to support policy.7,11,12 As a
result, more and more studies testing the effectiveness
of FOP labels are being published. Researchers study
different aspects, such as (self-reported) consumer
understanding, liking, and use of FOP labels,13–23 obser-
vational label use,24–32 and effects of FOP labels
on reformulation,33–35 product sales,36–38 and health
outcomes.39–43 In this way, they aim to evaluate the impact
of the labels on the health of society.

Thus far, researchers have investigated the effects of
FOP labels in different ways. However, it appears no over-
view has been published to date regarding the methodo-
logical quality of these studies. Studying methodological
issues is important because good scientific studies can
provide reliable evidence about the effectiveness of FOP
labels. Some studies have already identified certain meth-
odological limitations; for example, it has been shown
that self-reported data do not accurately reflect actual
FOP label use.23,26,28 Consequently, more studies started to
make use of observational data, such as in-store observa-
tions,25,26,29,32 eye-tracking,27,31 or collection of supermar-
ket sales data.36,37 However, it is unclear whether these
data accurately reflect food purchases actually resulting
from FOP labels or whether other factors bias the results.
What is the most accurate way to study the effectiveness
of FOP labels, and which outcome measures are most
relevant for public health? In considering these questions,
this study aimed to provide a review of the methodologi-
cal quality of current FOP labeling research. The
strengths and limitations of the current studies are dis-
cussed, and future research challenges are proposed.

METHODS

Structure

For the purpose of this review, “effectiveness” was defined
as the measure of impact of FOP labels on consumer
behavior, product reformulation, and health outcomes.
Consumer behavior was subdivided into effects on con-
sumers’ self-reported understanding and use of FOP
labels, effects on consumers’ observational use, and effects
on sales. Figure 1 shows a schematic representation of this
subdivision, which was based on the designs and main
outcomes of the studies.

Search strategy

Peer-reviewed articles were located using a computerized
search of the databases PubMed and the Web of Science
(ISI) from 1990 to February 2011. The following key
words were used: “front of pack,” “nutrition logo,” “nutri-
tion label,” “nutrition symbol,” “on-package nutrition
information,” and “health logo.” The names of all current
existing FOP labels were also used as key words; for
example, Traffic Light, Guiding Stars, Canada’s Health
Check, AHA Heart Check, NuVal, Green Keyhole,
Choices logo, Guideline Daily Amount (GDA), Finnish
Heart Symbol, and Pick the Tick.1,3 In addition, the refer-
ence lists from key published articles were reviewed for
relevant articles.

Inclusion and exclusion criteria

Studies that evaluated the effectiveness of FOP labels
actually in use were included. Studies that investigated the
nutrition facts panel and other back-of-pack information,
health claims, calorie labeling, general on-package nutrition
statements (e.g., “low salt” or “healthy food”), the
general term “FOP labels” without mentioning the name
of any specific FOP label, or health logos not in use in real
life (e.g., a tick with the text “healthy choice” developed by
researchers and tested in a specific study design) were
excluded. Furthermore, FOP labeling studies that had no
clear effectiveness measure on consumer behavior or
Quality score

The quality of the studies included was assessed by two independent researchers (ELV and HEB) using the quality assessment tool developed by Sirriyeh et al.44 This tool is applicable to diverse research designs, including quantitative, qualitative, and mixed designs and enables comparison across a diverse range of studies. The tool consists of 16 criteria and is shown in Appendix S1 located online in this article’s Supporting Information. Each research paper was awarded a score on a scale from 0 to 3 for each of the criteria. There was initial coder disagreement for 55 of the 496 scores; discussion following the independent scoring of papers resolved all differences in agreement between the two researchers. The sum of the scores provided a quality score per paper, and this score was expressed as a percentage of the maximum score possible (range 0–100%).

RESULTS

The initial search generated a total of 622 citations, of which 122 titles appeared to meet the inclusion criteria and were reviewed. After reading the abstracts and/or full-text articles, 31 studies met the inclusion criteria and were included. Table 1 lists the characteristics of these 31 studies, subdivided as follows: 11 self-reported consumer studies, 9 objective consumer studies, 3 sales studies, 3 studies focusing on reformulation, and 5 studies focusing on health outcomes. The last column of Table 1 shows the quality scores per paper. Appendix S1, located online in this article’s Supporting Information, shows the calculations of the quality scores per paper. Table 2 lists the main items identified as contributing to a high quality score or a low quality score; these items are discussed under the subheadings below.

Self-reported consumer studies

The earliest studies evaluating effects of FOP labels, of which 11 studies were located,13–23 used consumer surveys. The outcome measures of these studies were based on self-reported data. Consumer studies using questionnaires generally aimed to provide insight into the understanding and use of FOP labels and to explore any differences in perception between consumer groups (high versus low educated, normal weight versus obese, men versus women, etc.). Most of the studies compared different FOP labeling formats and tried to identify which format can best guide consumers in making healthier food choices.13,15–18,21,22 The mean quality score of these studies was 48.8% (range 35.7–62.5%). A criterion on which the studies scored high in general was a “representative population sample of considerable size”: 9 of the 11 studies used large, well-balanced consumer panels consisting of 400 to up to 2,200 consumers.13–15,17–20,23 Criteria on which most studies received low scores were the “explanation for choice of data collection tools” and the “fit between research question and method of data collection”: many different tools were used to measure “use” and “understanding,” such as labeling tasks with photos, choice cards, computer tests, and comparison of mock packages.13–16,18 How well do these experimental self-reported data reflect actual understanding and use in real-life shopping environments? In general, when the limitations of the studies were discussed, there was a lack of critical reflection of the tools used. Moreover, the studies scored poorly with regard to validity of the measurement tools: although one study referred to some pre-tested scales,13 none of the questionnaires were validated.

Only two studies used focus group interviews.22,23 These qualitative group interviews were used to provide more in-depth insights as to how consumers understand and use FOP labels while viewing certain product packages.

Observational consumer studies

Nine studies that used observational methods to measure the understanding and use of FOP labels were located.24–32 Some of these studies appear to overestimate actual FOP label use in real-life settings,26,28,32 which highlights the importance of collecting more objectively assessed behavioral data. The mean quality score of these studies was 68.6% (range 50.0–76.2%). A criterion on which the studies in general scored high was a “clear description for the choice of data collection tools”: different observational data were collected, such as product observations in supermarkets,25,26,32 grocery receipts,29 records of reaction times during performance of a computer task,24 tracking of eye movements during performance of computer labeling tasks,27,31 thinking aloud data (shoppers were asked to “think aloud” during their shopping trip, and the conversations were recorded on tape),28 and food consumption data.30 Five studies were conducted in a real-life supermarket setting,25,26,28,29,32 leading to a higher quality score because real-life observations better reflect actual shopping behavior than studies conducted in laboratory settings.26,32

Most of these studies, however, struggled with the question of how accurately observational data reflect actual FOP label understanding and use, and most studies identified this in their study limitations. As stated by
<table>
<thead>
<tr>
<th>Type of study</th>
<th>Reference</th>
<th>Type of existing FOP label</th>
<th>Subjects (n)</th>
<th>Setting</th>
<th>Study design/control group</th>
<th>Main outcome measure(s)</th>
<th>Type of data collected</th>
<th>Authors' conclusions</th>
<th>Quality score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-reported consumer studies</td>
<td>Andrews et al., 2011</td>
<td>Smart Choice, traffic lights – GDA</td>
<td>US consumers (520)</td>
<td>Online survey, mock packages in questionnaire</td>
<td>Cross-sectional</td>
<td>Healthiness perception, nutrient use, purchase intentions</td>
<td>Self-reported questionnaire data</td>
<td>Smart Choices can lead to positive and potentially misleading nutrient and healthiness evaluations when compared with traffic lights – GDA or no FOP</td>
<td>54.8%</td>
</tr>
<tr>
<td></td>
<td>Balcombe et al., 2010</td>
<td>Traffic lights</td>
<td>UK consumers (477)</td>
<td>Experimental setting with choice cards in questionnaire</td>
<td>Cross-sectional</td>
<td>Understanding of traffic lights</td>
<td>Questionnaire data</td>
<td>Consumers prefer to reduce intake of any nutrient with a red light; different consumer groups respond differently to traffic lights</td>
<td>52.4%</td>
</tr>
<tr>
<td></td>
<td>Borgmeier &amp; Westenhoefer, 2009</td>
<td>Traffic lights, GDA</td>
<td>Consumers (420)</td>
<td>Experimental setting, tasks with photos of labeled foods</td>
<td>Randomized experimental design: four labeling groups and one control group</td>
<td>Identification of healthier food in pairwise comparisons and daily food selection</td>
<td>Photo selections by interviewer</td>
<td>Highest percentage of correct identifications of healthier foods with traffic lights; envisaged daily food consumption did not differ between conditions</td>
<td>45.2%</td>
</tr>
<tr>
<td></td>
<td>Feunekes et al., 2008</td>
<td>Traffic lights, Wheel of Health (+ fake labels)</td>
<td>Consumers from 4 European countries (1,630)</td>
<td>Online questionnaire</td>
<td>Cross-sectional</td>
<td>Consumer friendliness and usage intention</td>
<td>Self-reported questionnaire data</td>
<td>Simpler labeling formats more appropriate for facilitating quick purchasing decisions</td>
<td>40.5%</td>
</tr>
<tr>
<td></td>
<td>Gorton et al., 2009</td>
<td>Traffic lights, GDA</td>
<td>New Zealand shoppers (1,525)</td>
<td>25 supermarkets</td>
<td>Cross-sectional</td>
<td>Understanding and use of different FOP labels</td>
<td>Questionnaire data</td>
<td>High reported use of labels; traffic lights showed high understanding across ethnic and income groups</td>
<td>59.5%</td>
</tr>
<tr>
<td></td>
<td>Kelly et al., 2009</td>
<td>Traffic lights, GDA</td>
<td>Australian consumers (790)</td>
<td>Experimental setting, choice tasks with mock packages</td>
<td>Cross-sectional</td>
<td>Consumer ability to compare the healthiness of labeled mock products</td>
<td>Questionnaire data + interviews about the choice tasks</td>
<td>Traffic lights were most effective in assisting consumers to identify healthier foods</td>
<td>47.6%</td>
</tr>
<tr>
<td></td>
<td>Larsson &amp; Lissner, 1996</td>
<td>Keyhole</td>
<td>Swedish women (616)</td>
<td>Experimental setting</td>
<td>Cross-sectional</td>
<td>Knowledge of Keyhole and relation to dietary fiber and fat intake</td>
<td>Questionnaire data + 24 h recall</td>
<td>Most women understood the meaning of Keyhole, but there was no difference in total fat or fiber intake between the women with less and those with more knowledge</td>
<td>50.0%</td>
</tr>
<tr>
<td></td>
<td>Larsson et al., 1999</td>
<td>Keyhole</td>
<td>Swedish participants (1,591)</td>
<td>Experimental setting</td>
<td>Cross-sectional</td>
<td>Knowledge of Keyhole and relation to dietary behavior</td>
<td>Questionnaire data + food frequency questionnaire</td>
<td>Higher intake of Keyhole low-fat foods among people with knowledge of Keyhole, but this was not observed among less educated consumers</td>
<td>50.0%</td>
</tr>
<tr>
<td></td>
<td>Moser et al., 2010</td>
<td>GDA, traffic lights</td>
<td>Consumers from Germany (147) and Belgium (128)</td>
<td>Consumer surveys (face-to-face interviews + self-administered questionnaires)</td>
<td>Cross-sectional</td>
<td>Understanding and preference</td>
<td>Self-reported survey data</td>
<td>German consumers prefer traffic lights, while Belgian consumers prefer GDA. Socio-demographics also play an important role</td>
<td>35.7%</td>
</tr>
<tr>
<td></td>
<td>Signa et al., 2008</td>
<td>Pick the Tick, traffic lights</td>
<td>Maori, Pacific, and low-income New Zealanders (158)</td>
<td>Focus group interviews</td>
<td>Cross-sectional</td>
<td>Understanding and use of different labeling systems</td>
<td>Qualitative focus groups</td>
<td>Lack of use of nutrition labels due to lack of time, lack of understanding, shopping habits, and absence on low-cost foods</td>
<td>38.1%</td>
</tr>
<tr>
<td></td>
<td>Wyth et al., 2009</td>
<td>Choices</td>
<td>Dutch consumers (2,200)</td>
<td>Online questionnaires + focus group interviews</td>
<td>Cross-sectional design on two different times</td>
<td>Exposure, and reported logo use</td>
<td>Self-reported questionnaire data + qualitative focus groups</td>
<td>Exposure to logo increased over time; health-interested consumers use Choices logo</td>
<td>62.5%</td>
</tr>
<tr>
<td>Observational consumer studies</td>
<td>Balikova &amp; van Trijp, 2010</td>
<td>Choices, GDA</td>
<td>Dutch young adults (24)</td>
<td>Experimental setting, visual search task on computer</td>
<td>Cross-sectional</td>
<td>Determinants of consumer attention to labels</td>
<td>Reaction time, accurate answers on tasks</td>
<td>Display size, color, familiarity, and location are key determinants of consumer attention</td>
<td>64.3%</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>--------------------------</td>
<td>---------------</td>
<td>------------------------</td>
<td>----------------------------------------------------</td>
<td>----------------</td>
<td>--------------------------------------</td>
<td>-------------------------------------</td>
<td>------------------------------------------------------------------------------------------</td>
<td>-----</td>
</tr>
<tr>
<td>Grunert et al., 2010</td>
<td>GDA</td>
<td>Shoppers from 6 European countries (1,781)</td>
<td>14 major retailers in 6 countries</td>
<td>Cross-sectional</td>
<td>Use and understanding of nutrition information</td>
<td>In-store observations, in-store interviews and questionnaires</td>
<td>Understanding higher than use, possibly due to lack of motivation. Considerable national differences in both understanding and use</td>
<td>69.0%</td>
<td></td>
</tr>
<tr>
<td>Grunert et al., 2010</td>
<td>GDA, traffic lights, color-coded GDA</td>
<td>UK shoppers (2,019)</td>
<td>Three major UK retailers</td>
<td>Cross-sectional</td>
<td>Use and understanding of FOP nutritional information</td>
<td>In-store observations, in-store interviews and questionnaires</td>
<td>Understanding of FOP was high; actual use was lower; usage was related to interest in healthy eating and understanding of nutrition knowledge</td>
<td>76.2%</td>
<td></td>
</tr>
<tr>
<td>Jones &amp; Richardson, 2007</td>
<td>Traffic lights</td>
<td>UK participants from a university setting (92)</td>
<td>Experimental setting, labeling tasks on computer</td>
<td>Cross-sectional</td>
<td>Perceived healthiness of nutrition label and the areas of label examined</td>
<td>Eye movements + healthiness ratings</td>
<td>Traffic lights helped participants to pay attention to important nutrients and improved accuracy of healthiness ratings compared with standard label</td>
<td>50.0%</td>
<td></td>
</tr>
<tr>
<td>Rayner et al., 2007</td>
<td>Logos from Tesco, Sainsbury, and Pick the Tick</td>
<td>Shoppers from UK and Australia (44)</td>
<td>Supermarket</td>
<td>Cross-sectional</td>
<td>Thinking aloud two times: shopping normally and shopping healthily</td>
<td>Reported use and actual use of FOP</td>
<td>Only Tesco shoppers rarely used the FOP label. Shoppers claim to use FOP labels, although thinking aloud revealed hardly any actual use</td>
<td>75.0%</td>
<td></td>
</tr>
<tr>
<td>Reid et al., 2004</td>
<td>Canada's Health Check symbol</td>
<td>Canadian shoppers (200)</td>
<td>Supermarket</td>
<td>Cross-sectional</td>
<td>Number of Health Check products purchased by shoppers</td>
<td>Grocery receipts + questionnaire data</td>
<td>Shoppers purchasing Health Check products had lower fat intake; logo awareness related to use and to interest in healthy foods</td>
<td>73.8%</td>
<td></td>
</tr>
<tr>
<td>Steenhuis et al., 2010</td>
<td>Choices</td>
<td>Female consumers from a university setting (36)</td>
<td>Experimental lab setting, taste experiment</td>
<td>Cross-over design with two conditions: logo and no logo</td>
<td>Weighed consumption of chocolate-cake</td>
<td>Consumption data + questionnaire data</td>
<td>Use of logo did not increase consumption and had no effect on taste rating</td>
<td>69.0%</td>
<td></td>
</tr>
<tr>
<td>Visschers et al., 2011</td>
<td>General FOPs on breakfast cereals in Swiss</td>
<td>Swiss students (32)</td>
<td>Experimental setting, food choice task on computer</td>
<td>Between-subject design, respondents randomized over two conditions: health or taste motivation</td>
<td>Eye movements</td>
<td>Eye movements + questionnaire</td>
<td>66% perceived nutrition label and/or FOP. Health motivation and package design directed consumer’s attention towards on-package nutrition information</td>
<td>71.4%</td>
<td></td>
</tr>
<tr>
<td>Wyth et al., 2010</td>
<td>Choices</td>
<td>Dutch consumers (404)</td>
<td>Nine supermarkets</td>
<td>Cross-sectional</td>
<td>Proportion of purchased logo products</td>
<td>Product observations + questionnaire data</td>
<td>Health-interested consumers purchase most logo products; hedonists purchase least logo products</td>
<td>69.0%</td>
<td></td>
</tr>
<tr>
<td>Sales studies</td>
<td>Sacks et al., 2009</td>
<td>Traffic lights</td>
<td>Supermarket store chains of 1 retailer (over 1,000) in the UK</td>
<td>One retailer</td>
<td>Natural experiment with measurements during 8 weeks</td>
<td>Sales of 6 ready meals and 12 sandwiches</td>
<td>Introduction of traffic lights had no effect on relative healthiness of consumer purchases</td>
<td>66.7%</td>
<td></td>
</tr>
<tr>
<td>Sutherland et al., 2010</td>
<td>Guiding Stars</td>
<td>Supermarket chain stores in the US (1,68)</td>
<td>One store chain</td>
<td>Natural experiment with three measurements in 2 years</td>
<td>Proportion of food items with star rating</td>
<td>Sales data</td>
<td>Proportion of purchased star items increased at 1- and 2-year follow-ups</td>
<td>59.5%</td>
<td></td>
</tr>
<tr>
<td>Wyth et al., 2011</td>
<td>Choices</td>
<td>Dutch workers (213) and employees (368)</td>
<td>Worksite cafeterias</td>
<td>Randomized controlled trial with one labeling and one control condition for 9 weeks</td>
<td>Proportion of Choices sandwiches and soups</td>
<td>Sales data + questionnaire data</td>
<td>No effect on sales when comparing logo with nonlogo cafeterias; health-interested employees reported using the logo</td>
<td>83.3%</td>
<td></td>
</tr>
<tr>
<td>Type of study</td>
<td>Reference</td>
<td>Type of existing FOP label</td>
<td>Subjects (n)</td>
<td>Setting</td>
<td>Study design/control group</td>
<td>Main outcome measure(s)</td>
<td>Type of data collected</td>
<td>Authors' conclusions</td>
<td>Quality score</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>--------------------------------</td>
<td>---------------------------</td>
<td>-------------------------------------------</td>
<td>----------------------------------------------</td>
<td>----------------------------</td>
<td>------------------------------------------</td>
<td>-------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Reformulation studies</td>
<td>Wyth et al., 2010&lt;sup&gt;33&lt;/sup&gt;</td>
<td>Pick the Tick</td>
<td>Dutch Choices products: 47 food manufacturers (821)</td>
<td>Cross-sectional</td>
<td>Reason of logo assignment and nutrient composition data</td>
<td>Data provided by food manufacturers</td>
<td>Choices had stimulated healthier product development, especially regarding sodium and dietary fiber</td>
<td>69.0%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Williams et al., 2003&lt;sup&gt;34&lt;/sup&gt;</td>
<td>Pick the Tick</td>
<td>Pick the Tick products from Australia (12)</td>
<td>Cross-sectional</td>
<td>Sodium content before and after reformulation</td>
<td>Chemical analyses + data provided by food manufacturer</td>
<td>Tick influenced food manufacturer to remove 235 tons of salt from breakfast cereals</td>
<td>26.2%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Young et al., 2002&lt;sup&gt;35&lt;/sup&gt;</td>
<td>Pick the Tick</td>
<td>Pick the Tick products from New Zealand (23)</td>
<td>Cross-sectional</td>
<td>Sodium content before and after reformulation</td>
<td>Chemical analyses + data provided by food manufacturers</td>
<td>Tick influenced food companies to exclude 33.3 tons of salt through reformulation of breads, breakfast cereals, and margarines</td>
<td>50.0%</td>
<td></td>
</tr>
<tr>
<td>Modeling and health outcomes studies</td>
<td>Ireland et al., 2010&lt;sup&gt;36&lt;/sup&gt;</td>
<td>Pick the Tick</td>
<td>Adult Australian consumers (49)</td>
<td>Real-life setting, free living</td>
<td>One group randomized to Tick products and others to low-salt foods for 8 weeks</td>
<td>Sodium excretion</td>
<td>24 h urine samples</td>
<td>Simple dietary education was effective in reducing sodium intake in free-living individuals</td>
<td>88.1%</td>
</tr>
<tr>
<td></td>
<td>Roodenburg et al., 2011&lt;sup&gt;37&lt;/sup&gt;</td>
<td>Choices</td>
<td>Intake data from 7 countries</td>
<td>Virtual setting, modeling</td>
<td>Comparisons of typical daily menus with Choices daily menus</td>
<td>Nutrient intakes</td>
<td>Food composition + food consumption data combined</td>
<td>Replacing typical daily menus by Choices menus can potentially lead to improved nutrient intakes towards recommendations</td>
<td>69.0%</td>
</tr>
<tr>
<td></td>
<td>Roodenburg et al., 2009&lt;sup&gt;38&lt;/sup&gt;</td>
<td>Choices</td>
<td>Dutch young adults (750)</td>
<td>Virtual setting, modeling</td>
<td>Evaluation of different Choices-compliant scenarios</td>
<td>Nutrient intakes</td>
<td>Food composition + food consumption data combined</td>
<td>Replacing foods by Choices-compliant foods can potentially lead to improved nutrient intakes</td>
<td>59.5%</td>
</tr>
<tr>
<td></td>
<td>Sacks et al., 2011&lt;sup&gt;39&lt;/sup&gt;</td>
<td>Traffic lights</td>
<td>Australian adult population</td>
<td>Virtual setting, modeling</td>
<td>Evaluation of traffic lights and junk-food tax scenario</td>
<td>Weight reduction, DALYs averted, and costs</td>
<td>Data from national nutrition survey + costing data</td>
<td>Traffic light labeling and junk-food taxes are highly cost effective as obesity prevention measures</td>
<td>69.0%</td>
</tr>
<tr>
<td></td>
<td>Temme et al., 2011&lt;sup&gt;40&lt;/sup&gt;</td>
<td>Choices</td>
<td>Dutch young adults (750)</td>
<td>Virtual setting, modeling</td>
<td>Evaluation of different Choices scenarios by using market shares</td>
<td>Nutrient intakes</td>
<td>Food composition + food consumption data combined</td>
<td>With Choices foods available in 2007, SAFA and sugar intake can be slightly reduced</td>
<td>69.0%</td>
</tr>
</tbody>
</table>

<sup>a</sup> The quality of the 31 included studies was assessed by two independent researchers, using the quality assessment tool developed by Sirriyeh et al.<sup>44</sup> The tool consists of 16 criteria and is shown in Appendix S1 (located online in this article's Supporting Information). Each research article evaluated was awarded a score on a scale from 0 to 3 for each of the criteria. Discussion following the independent scoring of articles resolved any differences in agreement by the two researchers. The sum of the scores provided a quality score per paper within a range of 0–100% (with 100% being the maximum score possible).

Abbreviations: DALYs, disability-adjusted life years; GDA, Guideline Daily Amount; SAFA, saturated fatty acids.
Grunert et al.,\textsuperscript{26} even though shoppers may have looked at nutritional information in-store, this does not necessarily mean that this information had an impact on their choice. Or, looking at the observational data of Vyth et al.\textsuperscript{32} and Reid et al.,\textsuperscript{29} neither of the studies was able to conclude whether health-conscious participants purchase FOP labeling products due to the logo or due to another reason. In addition, the eye-tracking studies report that the results do not indicate whether respondents understood the information they perceived correctly, hampering the interpretation of the data.\textsuperscript{27,31} The studies stress the importance of conducting more longitudinal studies in which causality can be assessed. Further, although most studies used some validated items from prior research—e.g., the food choice questionnaire,\textsuperscript{32} the dietary restraint scale,\textsuperscript{30} questions to assess the use of food package information,\textsuperscript{29} or an instrument measuring nutritional knowledge\textsuperscript{25,26}—none of the measurement tools or the questionnaires used were reported as being validated for the specific research purpose.

**Sales studies**

Only two studies have been published that collected supermarket sales data before and after the introduction of an FOP label to study whether the FOP label influences sales. These studies had a quality score of 66.7\%\textsuperscript{36} and 59.5\%.\textsuperscript{37} Both scored high on “rationale for choice of data collection tools,” since both studies collected a large amount of objective longitudinal sales data from real-life supermarket chain stores. The study by Sacks et al.\textsuperscript{36} scored lower on “representative sample of products” than Sutherland et al.\textsuperscript{37} because Sacks et al.\textsuperscript{36} only collected sales data from 6 ready meals and 12 sandwiches, whereas Sutherland et al.\textsuperscript{37} looked at all products with the Guiding Stars symbol.\textsuperscript{36,37} Further, both studies scored low on “fit between research question and method of data collection”: neither study had a control group. The change in sales may have been attributable to new product and package introductions, possibly in combination with the FOP label or other on-package nutritional information such as “low fat” or “light” statements, effects of price discounts, product group promotions, and/or product life cycles.\textsuperscript{45} Sutherland et al.\textsuperscript{37} do not discuss this limitation, but Sacks et al.\textsuperscript{36} note that attributing the observed increase in sales to the introduction of the FOP labels is not completely accurate, as the products examined were also reformulated at the time the labels were introduced, and the product packaging and manufacturer were changed.\textsuperscript{36}

One other study measured changes in sales after the introduction of an FOP label, but in worksite cafeterias, not in supermarkets.\textsuperscript{38} This randomized controlled study in 25 worksites measured objective sales data, consisted of a sample of reasonable size, and did include a control group in the longitudinal design, which makes it possible to link any change in sales to the label (quality score 83.3\%).

**Reformulation studies**

Only three studies have been published that evaluated the effects of FOP labels on product development.\textsuperscript{33–35} The mean quality score of these studies was 48.4\% (range 26.2–69.0\%). Young and Swinburn\textsuperscript{35} and Vyth et al.\textsuperscript{33} scored high, for example, on a “clear description of the
procedure for data collection” (data provided by food manufacturers) and a “clear explanation for choice of data collection tools” (nutrient composition data before and after assignment of the FOP labels). Clear explanations for these items were lacking in the study by Williams et al.34 Low scores were assessed for the sample size of Young and Swinburn35 and Williams et al.34: sample sizes were quite small (23 and 12 products, respectively), while the sample of Vyth et al.33 was larger ($n = 821$), though still not exhaustive. Further, only Vyth et al.33 had a clear section on limitations. An important limitation mentioned was that most data were self-reported by the food manufacturers.

**Health outcomes studies**

Epidemiological modeling is a way to investigate potential effects of FOP labels on nutrient intakes and health outcomes. Four studies that evaluated the effects of FOP labels by modeling were found.40–43 The mean quality score of these studies was 66.6% (range 59.5–69.0%). These studies scored high, for example, on “fit between research question and method” and “representative sample size”: they estimated the effects of FOP labels on nutrient intakes and health outcomes on the basis of national databases with population data instead of small consumer groups. Nevertheless, these studies scored low on “validity of the measurement tools,” mainly because modeling studies are based on so many highly selective assumptions. Assumptions in these four studies were related, for example, to compensation behavior, food replacement procedures, scenario development, costs estimations, and the associations between nutrients and health from limited literature sources. Only one study looked at the actual effects of FOP labels on biomarkers in the real-life setting.39 Ireland et al.39 reached the highest quality score of all studies in this overview (88.1%). This high score is attributable to its longitudinal design with free-living individuals in the real-life setting and to its validated measurement tools: the investigators used the biomarker “24 h urinary sodium excretion,” which is considered the most reliable method of assessing sodium intake compared with more subjective measures, such as dietary recall methods.46

**DISCUSSION**

This is the first study that provides an overview of the methodological quality of current FOP labeling research. Quality assessments were used to identify certain challenges for future research. Table 3 lists these research challenges, which are discussed in this section according to each subheading. The least relevant challenge is outlined first, and the overview ends with the most relevant research challenge from a public health perspective. Public health relevance is illustrated by Figure 2. The upper part of this figure is based on the theoretical framework for studying consumer responses to nutrition labeling, developed by Grunert and Wills.47 Figure 2 is further clarified in the subheadings below.

**Self-reported consumer studies**

Self-reported consumer studies provide interesting initial insights into the understanding and intention of FOP label use. Providing a questionnaire to a consumer panel is considered a relatively easy, quick, and inexpensive way to collect research data. However, the relatively low quality scores and the relatively low public health relevance of these studies (see Figure 2) make them scientifically less interesting, and thus the results should be used with caution with regard to policy recommendations. Whether these experimental self-reported data reflect actual use in real-life shopping environments is highly questionable.26,28,32,36

**Observational consumer studies**

Observational consumer studies have higher quality scores and higher public health relevance than self-
reported consumer studies (see Figure 2). Although observational studies lack the ability to assess causality, these studies better reflect actual behavior in real-life settings, in which consumers are influenced by many food choice motives, such as price, taste, time, and convenience.48 Nevertheless, in both self-reported and observational consumer studies, the lack of a validated methodology and the lack of a validated questionnaire to measure FOP label use were identified. Future studies should attempt to validate labeling questionnaires by using, for example, consumer scan panels that scan the amount of labeled products that were purchased. Biomarkers of intake represent another way of validation and are described at the end of the Discussion.

Sales studies

This type of research is considered to be of higher public health relevance than small-scale consumer studies because of its large objective databases. The inclusion of a control group is essential to attribute the sales effect to the introduction of the FOP label alone. Regarding the study of Sutherland et al.,37 it would have been interesting, for example, if the FOP label had been introduced in half of the chain stores initially and in the other half 6 months later. Then, one could compare the sales of the labeled stores with the sales of the nonlabeled stores (while keeping all other factors equal). However, sales databases are quite crude and cannot be used to reflect individual food intake,36,38 explaining their lower public health relevance compared with data from studies that measure food intake and health outcomes (Figure 2). This stresses the importance of collecting individual dietary intake data as well.

Reformulation studies

Although the reformulation studies had relatively low quality scores due to methodological weaknesses, they have relatively high public health relevance: reformulation can increase the availability of healthier products and consequently may have a large impact on all consumer groups without necessitating a change in consumer behavior.33 The challenge is to collect objective food composition data. Ideally, these data should be considered chemically analyzed data rather than data self-reported by manufacturers. It is recommended that such reformulation data be collected from the start of the introduction of an FOP label, because afterwards they can be difficult to retrieve. Moreover, it is especially likely that those manufacturers that participated in the prior reformulation studies are the ones that significantly improved their products. Therefore, it is recommended that data regarding the numbers of unhealthy products that were...
introduced in the same timeframe be collected in order to allow the overall picture of the food supply to be evaluated.

**Health outcomes studies**

What are the effects of FOP labeling on health outcomes? Modeling studies provide some potential insights. If updated food consumption and food composition data are available, these studies can provide interesting insights into the potential effects of FOP labels on nutrient intakes and health.\(^{40,42,43}\) Nevertheless, the most relevant question from a public health perspective remains as follows, as illustrated by Figure 2: What are the actual effects of FOP labels on a population’s health? The most interesting research challenge is the measurement of the health effects of FOP labels in real-life settings by using biomarkers of intake and biomarkers that are good predictors of disease risk.

**Study limitations**

First, this study is limited because it is not a systematic review. In addition, one could discuss the decision not to include, for example, studies about FOP calorie labeling or studies just stating “FOP labels.” Nevertheless, it is unlikely that the studies missed in this review would have had a major impact on the conclusions.

Second, the quality assessments of the studies can be discussed. Although many validated quality assessment tools are currently available, they are largely limited to the assessment of studies with a specific research design: 50–60 tools are currently available to assess the quality of randomized controlled trials, along with a range of other tools for other single research designs.\(^{49,50}\) Because FOP labeling studies tend to encompass diverse research methods and designs, the tool of Sirriyeh et al.\(^ {44}\) was chosen, which is able to evaluate overall quality of different designs. This tool is limited in that it relies on the researcher’s knowledge and expertise to enable fair and consistent assessments to be drawn. Nevertheless, this limitation was overcome by assessing the quality scores assigned by two independent researchers. Large-scale validation of the tool is still needed.\(^ {44}\) This tool is also limited because it does not take into account the goals of different kinds of FOP labeling research: a well-performed modeling study based on theoretical assumptions receives a higher score than a reformulation study that has a weak methodology yet uses actual reformulation data. Finally, it does not take into account that it is much easier and less costly to perform a simple desk-based modeling study than a multiple-country field study with in-store observations and a large sample size. This stresses the need for a multimethod approach that can validate the criteria of FOP labels by conducting modeling studies and can also assess the effects of FOP labels on consumer understanding, reformulation, and health outcomes in real-life settings.

**The research challenge for the coming years**

The two studies with the highest quality scores\(^ {38,39}\) have interesting methodological characteristics in common. Both have a longitudinal randomized design and used observational methods to measure the effects of FOP labels in real-life settings. The inclusion of a control group enables the effects to be attributed to the FOP label alone. Although there is no single, universally accepted hierarchy of evidence, there is broad agreement that randomized, controlled, longitudinal research designs in real-life settings provide one of the highest forms of scientific evidence.\(^ {51}\) When the public health relevance of FOP labeling studies is also taken into account, the research challenge for the coming years will be the measurement of the health effects of FOP labels in real-life settings by using biomarkers. It would be interesting to develop a randomized, controlled, longitudinal design in which one group of consumers consumes products that comply with FOP label criteria, and the other group consumes their regular diets. The main outcome measures could be changes in urinary sodium excretion as a biomarker for sodium intake, and levels of blood lipids as a marker for saturated fat intake. It would also be interesting to investigate the effects of FOP labels on cardiovascular risk factors such as cholesterol levels and blood pressure. Comparing the effects of different types of labels, such as traffic lights and health logos, would provide useful information for the international debate about FOP labeling.

**CONCLUSION**

Evaluations of FOP labels vary greatly in methodological rigor, and few methodologically sound studies are presently available. The highest levels of methodological quality and public health relevance are achieved through measuring the health effects of FOP labels by using biomarkers in a longitudinal, randomized, controlled design in a real-life setting. The research recommendations presented here can be used to challenge future researchers to expand the current knowledge of FOP labeling and its effects on consumers.

**Acknowledgments**

*Authorship.* ELV conducted the literature search, assessed the quality scores, and wrote the manuscript.
HEB assessed the quality scores and reviewed and critiqued the manuscript. IHMS, AJCR, JB, and JCS supervised data collection and reviewed and critiqued the manuscript. All authors approved the final version of the manuscript.

**Funding.** ELV is working at the VU University and is funded by the Choices Foundation. The Choices Foundation played no role in the study design, analysis, or interpretation of the data.

**Declaration of interest.** AJCR is employed by Unilever R&D, the Netherlands. The other authors have no relevant interests to declare.

**REFERENCES**


SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Quality scores per paper.