REVIEW

Reproducibility and Validity of A Posteriori Dietary Patterns: A Systematic Review

Valeria Edefonti,[1] Roberta De Vito,[2] Michela Dalmartello,[1] Linia Patel,[1] Andrea Salvatori,[1] and Monica Ferraroni[1]

[1] Branch of Medical Statistics, Biometry and Epidemiology "G. A. Maccacaro", Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Milan, Italy; and [2] Department of Computer Science, Princeton University, Princeton, NJ, USA

ABSTRACT

The effective use of dietary patterns (DPs) remains limited. There is a need to assess their consistency over multiple administrations of the same dietary source, different dietary sources, or across different studies. Similarly, their generalizability should be based on a previous assessment of DP construct validity. However, to date, no systematic reviews of reproducibility and validity of a posteriori DPs have been carried out. In addition, several methodological questions related to their identification are still open and prevent a fair comparison of epidemiological results on DPs and disease. A systematic review of the literature on the PubMed database was conducted. We identified 218 articles, 64 of which met the inclusion criteria. Of these, the 38 articles dealing with reproducibility and relative and construct validity of DPs were included. These articles (published in 1999–2017, 53% from 2010 onwards) were based on observational studies conducted worldwide. The 14 articles that assessed DP reproducibility across different statistical solutions examined different research questions. Included were: the number of food groups or subjects; input variable format (as well as adjustment for energy intake); algorithms and the number of DPs to retain in cluster analysis; rotation method; and score calculation in factor analysis. However, we identified at most 3 articles per research question on DP reproducibility across statistical solutions. From another 15 articles, reproducibility of DPs over shorter (≤1 y) time periods was generally good and higher than DP relative validity (as measured across different dietary sources). Confirmatory factor analysis was used in 15 of the included articles. It provided reassuring results in identifying valid dietary constructs characterizing the populations under consideration. Based on the available evidence, only suggestive conclusions can be derived on reproducibility across different statistical solutions. Nevertheless, most identified DPs showed good reproducibility, fair relative validity, and good construct validity. *Adv Nutr* 2020;11:293–326.

Keywords: a posteriori dietary patterns, cluster analysis, construct validity of dietary patterns, consistency of dietary patterns, factor analysis, generalizability of dietary patterns, reproducibility of dietary patterns, relative validity of dietary patterns, validity of dietary patterns

Introduction

Since the early 1980s, dietary patterns (DPs) have been used to synthesize multiple related dietary components in combined variables representing key dietary habits and/or the overall diet in free-living individuals. Interest in DPs is also motivated by well-known interactive effects of foods that are eaten together and by data dimensionality/multiple testing issues affecting the statistical analysis of many single food groups (FGs) or nutrients (1).

However, the lack of consistent methodology in deriving DPs has severely limited the ability to draw firm conclusions about the health risks or benefits associated with DPs (2). Indeed, only the most recent version of the Dietary Guidelines for Americans (3) has included evidence on DPs.

In 2012, the National Cancer Institute launched the Dietary Patterns Methods Project to support standardized and parallel analyses on selected a priori (or index-based) DPs and mortality outcomes in 3 large US cohorts (2). An index-based approach to DPs was chosen because results can be readily translated into dietary recommendations. Based on the application of multivariate statistical analysis to the available data, the a posteriori (or data-driven) DPs offer the advantage of representing actual dietary behavior in a population at a certain time point. If the population variability is well captured, the set of identified a posteriori DPs provides a realistic representation of eating choices (4). In addition, the a posteriori approach could capture rare,

VE was supported by Università degli Studi di Milano "Young Investigator Grant Program 2017." Author disclosures: VE, RDV, MD, LP, AS, and MF, no conflicts of interest. The funder had no role in any phase of this systematic review.

Supplemental Tables 1–4 are available from the "Supplementary data" link in the online posting of the article and from the same link in the online table of contents at https://academic.oup.com/advances/.

Address correspondence to VE (e-mail: valeria.edefonti@unimi.it).

Abbreviations used: ARI, adjusted Rand index; CA, cluster analysis; CFA, confirmatory factor analysis; DP, dietary pattern; DR, dietary record; EFA, exploratory factor analysis; FFQ, food-frequency questionnaire; FG, food group; m24HR, mean 24-h recall; PCA, principal component analysis; SMC, Swedish Mammography Cohort; 24HR, 24-h recall; 48HR, 48-h recall.

but well-characterized, dietary behaviors of subpopulations, including ethnic minorities (5).

Subjective decisions have been constantly reported as a limitation in studies deriving a posteriori DPs with principal component analysis (PCA), exploratory factor analysis (EFA), or cluster analysis (CA) (6). These decisions concern input variable format and potential transformation, number of input variables and food grouping schemes, and estimation method as well as criteria for model selection, including how to choose the number of DPs to retain (7). Although subjectivity in PCA/EFA and CA is often emphasized, very few articles have provided a formal comparison of different modeling strategies based on objective criteria. The reproducibility of DPs across different statistical solutions has rarely been a concern.

Similarly, confirmatory factor analysis (CFA) still has limited use in the validation of EFA-based DPs and in the development of constructs representing correlation structures among FGs and among DPs. Even though this should be the first step for the generalization of DPs to other studies, their construct validity has been investigated in few articles.

More generally, the reproducibility of similar a posteriori DPs across time, studies, and/or countries has not been extensively assessed so far (5, 8). Although in the literature there is a distinction between consistency of DPs across multiple administrations of the same dietary assessment tool in a short period of time (reproducibility) (9) and consistency over longer time periods (stability over time) (10), unsolved methodological issues have been reported in both these analyses (11, 12). Similarly, the comparison of a posteriori DPs across different dietary assessment tools (relative validity) (9) poses unsolved methodological issues (13).

To our knowledge, no attempts have been carried out so far to collect and summarize the existing evidence on reproducibility and validity of a posteriori DPs. This article provides details of the literature search and selection process and also summarizes the evidence on reproducibility and relative and construct validity of DPs. A companion review will include information on stability of DPs over longer time periods and reproducibility of DPs across studies.

Methods

Literature search strategy

We carried out a systematic search through MEDLINE via PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) to identify all the articles on reproducibility and validity of a posteriori DPs, based on the following string: "(reproducibility or validity) and dietary pattern*". The search was restricted to human studies reported in the English language and published up to January 11, 2019 and followed the guidelines from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) group (14). Two authors (MD and VE) independently selected the articles and retrieved and assessed the potentially relevant ones. The reference lists of the identified articles as well as other systematic reviews focusing on similar topics were also scanned. Discrepancies in article selection were resolved by involving a third researcher (MF).

Inclusion and exclusion criteria

Articles were included or excluded according to the following criteria.

A posteriori dietary patterns.

We focused our systematic review on a posteriori DPs. However, in the absence of previous known reviews on this topic, we preferred not to add the term "a posteriori" to our search string. Therefore, we further excluded articles presenting reproducibility or validity of a priori DPs only or applying reduced rank regression. We included in the review articles comparing a priori and a posteriori DPs as far as they provided information on reproducibility and validity of a posteriori DPs. We also considered articles comparing PCA (or EFA) and CA, but we excluded them when concentrating only on the comparison between PCA/EFA- and CA-based DPs (e.g., reference 15).

Reproducibility and validity of a posteriori dietary patterns.

In recent years, disagreements in terminology across different scientific areas have characterized the concepts of reproducibility, replicability, and validity of scientific findings (16, 17). In **Supplemental Table 1**, we introduce the basic definitions adopted in the current review as well as the statistical tools used for their assessment. We integrate basic terminology within the scientific process of DP identification in nutritional epidemiology.

Figure 1 shows prototypical paths of DP identification processes related to reproducibility and validity of DPs. Dietary patterns are identifiable within any study design and starting from any dietary assessment tool source. If 1 dietary source is used at 1 time point, the assessment of DP reproducibility arises from the use of different statistical approaches for DP identification (Figure 1A). Within the validation study of a new food-frequency questionnaire (FFQ), the same FFQ was administered twice (within 1 y) and compared with a gold standard dietary assessment tool [a dietary record (DR) or (multiple administrations of) a 24-h recall (24HR)] carried out on the same time interval and sample; DP reproducibility is assessed comparing the 2 sets of FFQ-based DPs, whereas relative validity of DPs is assessed comparing FFQ-based and gold standard–based DPs (Figure 1B). When either cohort studies or multiple waves of the same survey are available, a dietary assessment tool is administered to the same subjects on multiple occasions over longer time periods and the comparison of sets of DPs at the available measurement occasions enables the evaluation of stability of DPs over time (Figure 1C). Finally, to assess cross-study reproducibility of DPs, comparison of different sets of DPs derived from comparable dietary sources (at similar time points) is possible across centers from the same study, or across different studies representing potentially

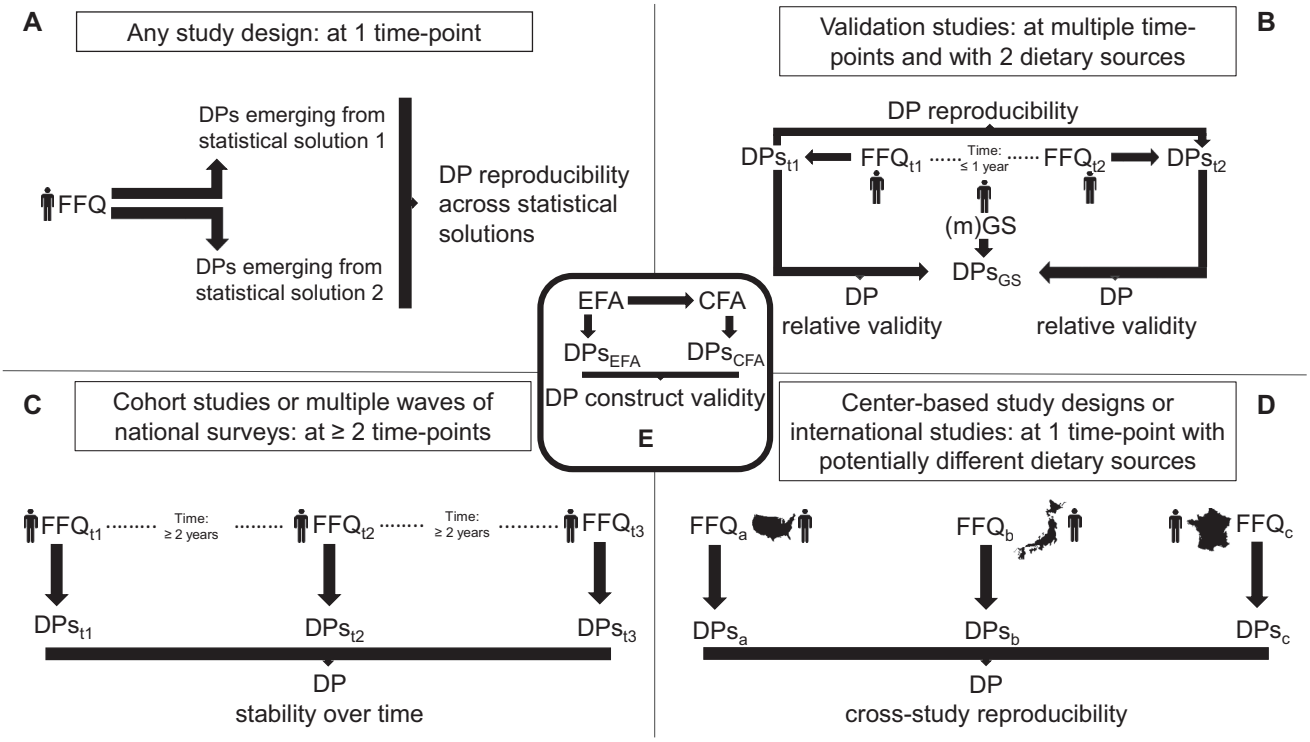

FIGURE 1 Schemes of dietary pattern identification processes related to the assessment of their reproducibility and validity. Specifically, reproducibility and/or validity of dietary patterns can be assessed in the following setups: (A) at 1 time point and with 1 dietary source; (B) at multiple time points (t1, t2, etc.) and with 2 dietary sources; (C) at multiple time points; and (D) across centers or studies. All these settings can include confirmation of the identified dietary patterns with confirmatory factor analysis (E). Abbreviations: CFA, confirmatory factor analysis; DP, dietary pattern; EFA, exploratory factor analysis; GS, gold standard dietary assessment tool; mGS, mean of intakes from multiple administrations of the same gold standard tool.

different populations or countries (Figure 1D). In any of these 4 settings, confirming EFA-based DPs is possible through CFA, which assesses construct validity of DPs; results from the 2 approaches can be formally compared with suitable statistical tools (Figure 1E). We reclassified the main findings from the articles included in the systematic review based on these definitions, regardless of the original definitions provided by the authors.

In summary, in the literature review, we distinguished the following definitions of reproducibility of DPs:

1) *Across different statistical solutions*: the extent to which similar DPs are consistently seen when a change occurred in: i) input variable format or scale; ii) number of input variables; iii) estimation method; or iv) criteria for model selection (including number of DPs to retain).
2) *Over time*: the extent to which similar DPs are consistently seen over short (i.e., ≤1 y) (traditionally called reproducibility in nutritional epidemiology) or longer (i.e., ≥2 y) time periods (stability over time).
3) *Across centers or studies (potentially representing different populations or countries)*: the extent to which similar DPs are common to diverse subsamples of interest, as opposed to study-specific DPs (cross-study reproducibility).

In the assessment of reproducibility across statistical solutions, we excluded articles that chose the number of clusters to retain with objective criteria (e.g., reference 18), within an analysis of the association between DPs and disease. In the assessment of cross-study reproducibility, we excluded articles based on a merged data matrix (generated by combining data from all the studies) approach (e.g., reference 19), where it was not possible to identify study-specific DPs and their potential reproducibility. Finally, we included articles using "internal validity" or "internal stability" indexes to choose the optimal number of clusters in the section on reproducibility of DPs across different statistical solutions. Although the terminology looks misleading, the research question was how to choose the number of clusters to retain and this was assessed with validity- or stability-based criteria for optimal solution identification.

The current review included and summarized evidence on reproducibility of DPs over shorter time periods and reproducibility across statistical solutions.

We also distinguished between construct validity and relative (or comparative) validity of DPs (Supplemental Table 1). Construct validity indicates whether a test measures its targeted latent constructs through suitable operationalizations of the constructs; in nutritional epidemiology, it deals with the ability of the empirically derived DP scores to

resemble the latent DPs in their composition and correlation with the other DPs. The relative validity of DPs has borrowed its meaning from the relative validity of an FFQ; it indicates the ability of FFQ-based DPs to resemble those derived on the gold standard tool. We included articles assessing either construct or relative validity of DPs. We excluded articles that only assessed validity of DPs against sociodemographic characteristics, lifestyle habits, nutrient/food profiles from the same dietary source, nutritional biomarkers, markers of disease, or a disease of interest (e.g., reference 20).

Finally, we excluded those studies that, although focusing on the association between some identified DPs and a disease, provided assessments of internal reproducibility with the split-half approach and/or reliability measured as internal consistency with Cronbach α (e.g., references 20–22).

Data extraction

Quantitative and qualitative data were extracted from each of the studies selected for in-depth review by 3 independent researchers (LP, MD, and VE); any discrepancies were resolved after consultation with a fourth author (MF) to maintain consistency. Information extracted included the following: *1*) general characteristics of the studies (first author, year of publication of the article, country, and study name); *2*) study design (type of design, brief description of data collection, number and age of the participants, and years of follow-up); *3*) dietary assessment tools used; *4*) DP identification method; *5*) DP name and composition; *6*) statistical methods used for the assessment of reproducibility and/or validity of DPs; and *7*) main results on DP reproducibility and validity.

Quality assessment of the included studies

Each article that met the inclusion criteria was independently rated for quality by all researchers, except 1 (MF), using the Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies from the NIH National Heart, Lung, and Blood Institute (23). If the ratings differed, then the remaining author (MF) was consulted for quality adjudication. Involved researchers used the available study rating tools on the range of items included in each tool to judge each study to be of "good," "fair," or "poor" quality. The reference tools used depended on the study design and included the Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies and the Quality Assessment Tool for Case-Control Studies (23); for the quality assessment of validation studies, we adopted the Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies, in accordance with the presence of repeated dietary measures. Because our review was not focused on any specific outcome of interest, the rating system items that dealt with: *1*) the presence of an outcome, or *2*) the association between exposure and outcome were consistently given a "cannot determine/not reported/not applicable" score across all the studies. Thus, the maximum rating for cohort/cross-sectional studies was 7 (of the original 14 items) and that for case-control studies was 9 (of the original 12 items). In addition, we decided that the item asking about reliability, validity, and consistent definition of the exposure (number 9 in the cohort/cross-sectional design tool and 10 in the case-control design tool) was concerned with the dietary assessment tools used to measure dietary information. When the assessment of either reproducibility or validity was performed on an FFQ, we marked "yes" in correspondence to the tool item. When other dietary assessment tools were used instead of an FFQ, we marked "yes" when either multiple administrations of a 24HR or a DR were provided. When a validation study was assessed for quality, we marked this item with a "not applicable" in the absence of any previous publication on FFQ reproducibility and validity. We did not consider applicable to our quality assessment process the part of point 10 asking for reliability of the risk measure in the case-control study design tool.

In general terms, a "good" study has the least risk of bias due to flaws in study design or implementation, a "fair" study is susceptible to some bias deemed not sufficient to invalidate its results, whereas a "poor" rating indicates significant risk of bias. We followed the website guidelines (23) and did not base our final evaluation on a cutoff approach on the total score (calculated by summing the "1"s corresponding to "yes"), but we carefully evaluated the "no" items to assess the overall risk of bias of the examined study. Finally, we chose not to exclude studies on the basis of their quality, because of the lack of previous evidence on reproducibility and/or validity of DPs.

Results

An initial literature search of the PubMed database identified 218 articles, of which 181 remained when we limited the search to publications related solely to humans and written in the English language. Their full texts were retrieved for detailed evaluation. After the exclusion of 35 review articles, 124 original research articles were also excluded because they met the exclusion criteria indicated previously. In detail, the most frequent reasons for exclusion were as follows: DPs intended as a synonym of dietary habits; a posteriori DPs not identified in the article—that is, a priori DPs, DPs from reduced rank regression (either exploratory or confirmatory), treelet transform, or latent class models—or just compared with the a priori ones; PCA- or EFA-based DPs compared with CA-based DPs, with no separate analyses on either approach; reproducibility and validity of FFQs and not of DPs; split-half or Cronbach α only; DP validity assessed against subjects' characteristics or a disease of interest; or conference abstracts not published as a full-text article. Forty-two additional articles were identified from manual searches of reference lists of selected original and review articles. Thus, 64 articles were included in our systematic review. Of these, 38 articles were included in the current review and were concentrated on reproducibility and relative and construct validity of DPs; the 34 articles that focused on stability of DPs over time and on their reproducibility across studies were included in an additional review. Eight articles (10, 11, 24–29) were common to both reviews (**Figure 2**).

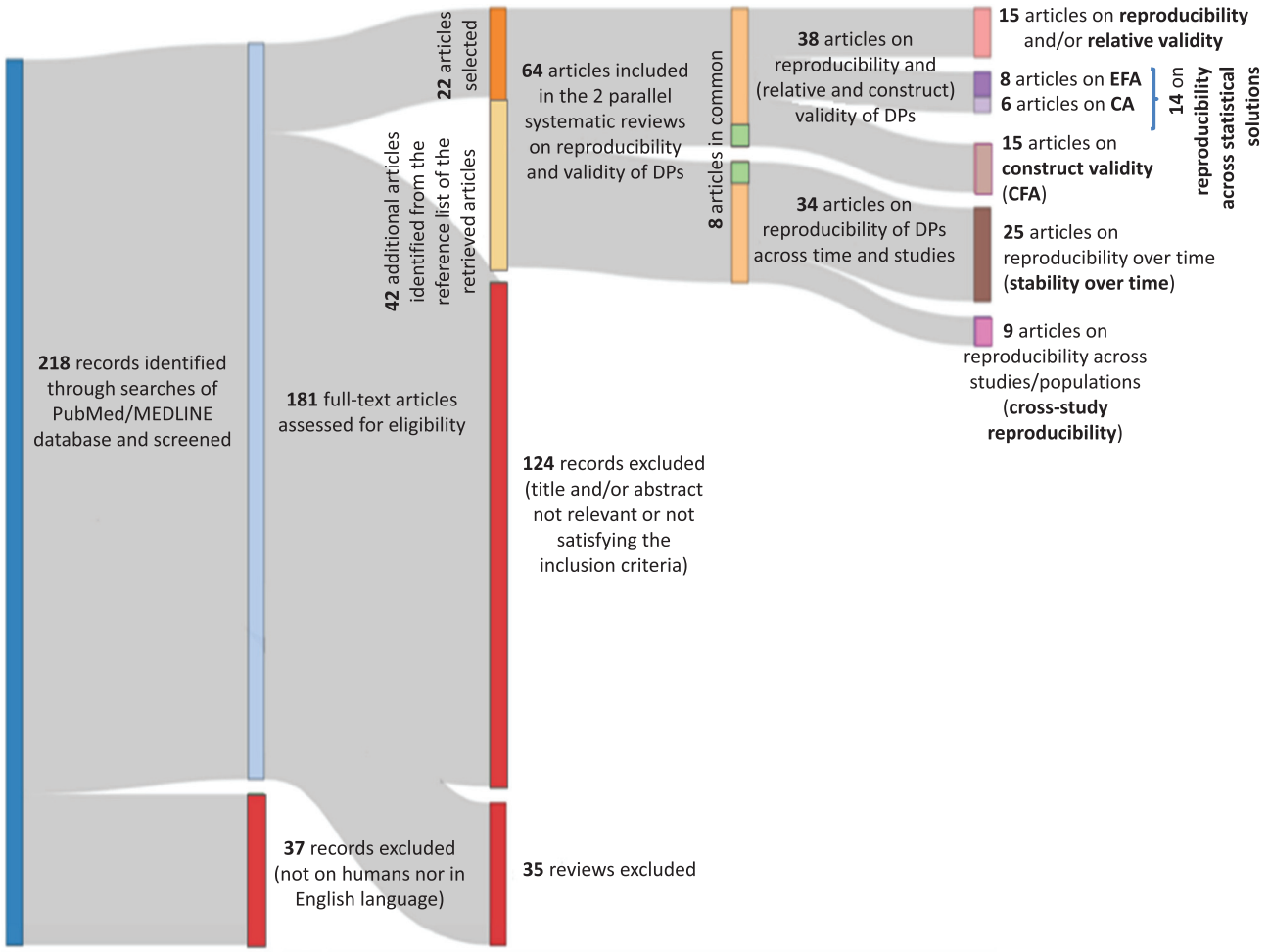

FIGURE 2 Sankey diagram showing the selection process used in the systematic review on reproducibility and validity of dietary patterns. In the current review, we have provided details of the 38 articles that assessed reproducibility and relative and construct validity of a posteriori dietary patterns. Abbreviations: CA, cluster analysis; CFA, confirmatory factor analysis; DP, dietary pattern; EFA, exploratory factor analysis.

General characteristics and study design information from the 38 studies on reproducibility and relative and construct validity of DPs (9–11, 13, 24–57) are presented in **Table 1**. The articles were published between 1999 and 2017, with 53% of them published from 2010 onwards; the studies were carried out in several areas of the world, including Europe and North America, but Asia and Oceania were also well represented, with 6 and 2 articles, respectively. A few articles were based on the same studies, including those from the Swedish Mammography Cohort (SMC) (10, 26–28, 33), from the MONItoring of Trends and Determinants in CArdiovascular Disease (MONICA) study (29, 47), and those from the European Prospective Investigation into Cancer and Nutrition (EPIC) study (49, 51, 55). All the articles were based on observational studies, including 1 case-control (45), 18 cohort (10, 13, 24, 25, 27, 28, 35, 36, 42, 43, 46, 48, 49, 51, 53–56), and 9 cross-sectional (38, 39, 41, 44, 47, 48, 50, 52, 57) studies, 1 multiple administration of the same survey (29), and 9 validation studies of FFQs (9, 11, 30–34, 37, 40). One study included adult men only (9), 11 studies included adult women only (10, 27, 28, 30, 33, 36, 37, 39, 49, 52, 55), with some of them based on pregnant women (36, 37, 39); 1 article was based on children (56) and another on adolescents (13). When available, the (total) follow-up time ranged from 1 mo (30) to 14 y (51). Dietary assessment instruments were administered between 1982–1983 (29) and 2014–2015 (34), with assessments equally carried out in the 1980s, 1990s, and 2000s, and a few in 2000–2010. With a few exceptions (35, 38, 42, 46, 50), the FFQ was the main dietary assessment tool used; in most studies, the FFQs were self-administered (8 FFQs were interviewer-administered only) and had a reference period of 1 y, with the obvious exception of the FFQs assessing diet during pregnancy (37, 39) and of the SMC FFQ (6 mo) (10, 26–28, 33). The number of food items inquired in the FFQs ranged from 26 (29, 47) to 284 (43), with 56% of the FFQs showing ≥100 items. When 2 FFQ administrations were available, the median time interval between them was 12

TABLE 1 Basic characteristics of observational studies on reproducibility and relative and construct validity of a posteriori dietary patterns[1]

Reference, location, title of study, quality	Study design	Subjects: *n*, age (y), and follow-up	Questionnaire
Ambrosini et al., 2011, Australia (13) Western Australian pregnancy cohort (Raine) study Fair quality	14-y follow-up of the Raine cohort study, including adolescents from 2900 pregnant Fs originally recruited at 16–20 wk of gestation between 1989 and 1991	1613 adolescents who completed the FFQ, 822 adolescents who completed the DR, 783 adolescents who completed both FFQ and DR Mean: 14 ± 0.2 Follow-up: not applicable	FFQ: 1 y; SA; validity assessed but no comments on the results; 212 FIs; FFQ completed by primary caregiver and adolescent 3-d DR completed by adolescents, and verified by a dietician; interest on representative DR 38 FGs common to all dietary sources
Asghari et al., 2012, Iran (11) TLGS Fair quality	TLGS: cohort study on urban residents in Tehran in 1999–2001; validation study of the TLGS FFQ based on a random sample of participants who were proportionately distributed across five 10-y age intervals and 2 sexes plus extra wave of the cohort study with FFQ administration	132 (89 completed FFQ3) 20–70 (mean: 35.6 ± 16.8) Follow-up: 8 y, until 2011	FFQ (based on a Willett format): 1 y; SA; reproducibility and validity to be assessed in this study, but validity granted for the analysis of stability over time; 168 FIs Twelve 24HRs: collected monthly on 2 formal weekend days and 10 weekdays FFQ1: completed 1 mo before collection of the first 24HR FFQ2: completed 1 mo after the last 24HR, 14 mo between FFQ1 and FFQ2 FFQ3: completed at the end of the follow-up; 19 FGs common to all dietary sources
Bailey et al., 2006, USA (Pennsylvania) (42) Geisinger Rural Aging Study Fair quality	Geisinger Rural Aging Study: longitudinal cohort study of rural older adults in Pennsylvania enrolled within a Medicare-managed health maintenance organization; random sample of participants to an intensive cross-sectional research study, not depressed or with functional limitations	179 66–87 (mean: 73 ± 5) Follow-up: none	Five 24HRs collected on random and nonconsecutive days over 10 mo using a multipass technique; m24HRs used for the analysis; 24 FGs for all time points
Balder et al., 2003, Netherlands, Sweden, Finland, and Italy (26) DIETSCAN (NLCS, SMC, ATBC, ORDET) Good quality	Parallel analysis of 4 studies (no pooled analysis); NLCS (random subcohort of): population-based cohort of Ms and Fs from Dutch municipalities; SMC: population-based cohort of Fs based on a mammography screening in 2 counties in central Sweden from 1987 to 1990; ATBC: randomized placebo-controlled intervention study conducted among M smokers who lived in southwestern Finland; ORDET: cohort study of Italian healthy volunteer Fs from the province of Varese, northern Italy	NLCS: 3123 (1598 Fs and 1525 Ms); SMC: 61,469 Fs; ATBC: 27,111 Ms; ORDET: 9208 Fs NLCS: 55–69 at baseline in 1986 (mean: 61.4 ± 4.2 for Ms and 61.4 ± 4.3 for Fs); SMC: 40–74 when invited to mammography screening in 1987 to 1990 (mean: 53.7 ± 9.7); ATBC: 50–69 at baseline between 1985 and 1988 (mean: 57.7 ± 5.1); ORDET: 35–69 between 1987 and 1992 (mean: 48 ± 8.5) Follow-up: 7 for NLCS (baseline: 1986); 13 for SMC (baseline: 1987–1990); NA for ATBC (baseline: 1985–1988, intervention ended in 1993 after 5–8 y, follow-up later on); 9 for ORDET (baseline: 1987–1992)	Four different but validated FFQs: NLCS-FFQ: 1 y; SA; NA reproducibility but valid; 150 FIs (51 FGs, but final number equal to 49); SMC-FFQ: 6 mo; SA; NA reproducibility but valid; 67 FIs (51 FGs, but final number equal to 42); ATBC-FFQ: 1 y; SA; reproducible and valid; 276 FIs (51 FGs, but smaller final number of FGs); ORDET-FFQ: 1 y; SA; reproducible and valid; 107 FIs (51 FGs, but final number equal to 32)
Beck et al., 2012, New Zealand (30) NA Poor quality	Validation study of a new FFQ; convenient sample of Fs living in Auckland in 2009 free of chronic disease, recruited with a magazine advertisement or invitation to potential volunteers	115 Fs 18–44 (median: 33) Follow-up: 1 mo	FFQ: 1 mo; SA; reproducibility and validity to be assessed in this study FFQ1: completed at baseline FFQ2: completed 1 mo later 4-d weighted DR: completed between FFQ1 and FFQ2 144 FIs for FFQ and DR (30 FGs—most frequently consumed on FFQ1)

(*Continued*)

TABLE 1 (*Continued*)

Reference, location, title of study, quality	Study design	Subjects: *n*, age (y), and follow-up	Questionnaire
Bedard et al., 2015, France (49) E3N (EPIC-France) Fair quality	1993 wave of the prospective cohort study E3N, after exclusion of current or former smokers, and of Fs with prevalent asthma at baseline	30,589 Fs 40–65 at baseline (mean: 53) Follow-up: 1993–2005	FFQ: NA reference period; SA; reproducible and valid; 208 FIs (27 FGs)
Bountziouka et al., 2011, Greece (40) NA Poor quality	Validation study based on a convenience sample, representative of the general population of Athens residents (stratified sample by age group and gender according to 2001 Census)	500 Mean: 46 ± 16 Follow-up: none	FFQ: 1 mo; IA; reproducible and valid; 76 FIs 3-d DR: based on 2 weekdays and 1 weekend day, over the same time span of the FFQ; DR FIs matched with FFQ FIs 24 FGs common to all dietary sources
Bountziouka and Panagiotakos, 2012, Greece (41) NA Fair quality	Nutrition survey	500 Mean: 37 ± 15 Follow-up: none	FFQ: 1 mo; IA; reproducible and valid; 76 FIs (24 FGs); FFQ completed twice, within a 15-d interval
Castro et al., 2015, Brazil (50) Health Survey of the City of São Paulo Poor quality	Health Survey of the City of São Paulo: cross-sectional population-based survey (using a complex multistage sampling design to have a representative sample of the city residents)	1102 (424 Ms; 678 Fs) ≥20, 46% ≥60 y Follow-up: none	Two nonconsecutive 24HRs, former collected face to face (USDA 5-Step Multiple Pass Method) and latter with telephone interview; 1169 FIs (38 FGs, but final analysis on 34 FGs)
Crozier et al., 2008, UK NA Fair quality (39)	Cross-sectional study including Fs in early pregnancy (median gestation: 15.3 wk) booked for delivery under 2 consultants in Southampton	585 Fs in early pregnancy with complete information on FFQ and DR ≥16 (mean: 26.4 ± 4.9) Follow-up: not applicable	FFQ: 3 mo (first trimester of pregnancy); IA; NA reproducible and valid; 100 FIs (49 FGs) 4-d DR: filled in immediately after completion of the FFQ, at the end of the first trimester of pregnancy; DR FIs mapped into the 100 FFQ FIs and then grouped in the 49 FGs used for the FFQ data
Dekker et al., 2013, Netherlands (25) Doetinchem Cohort Study Good quality	Three successive surveys (surveys 2, 3, and 4, at 3, 11, and 16 y after the first one) within the Doetinchem population-based cohort study including at baseline an age- and sex-stratified random sample of residents from Doetinchem town; follow-up available for 2/3 of the original random sample by design	4007 subjects with information available for the 3 rounds. In detail: 1993–1997: 6113 (survey 2); 1998–2002: 4916 (survey 3); 2003–2007: 4520 (survey 4) 47–66 Follow-up: 6, 11, 16 y after the first survey, so 10-y follow-up from survey 2 to survey 4	FFQ: 1 y; NA SA; reproducible and valid; 178 FIs (32 FGs)
Fransen et al., 2014, Netherlands (51) EPIC-NL Fair quality	Cohort study consisting of Prospect-EPIC and the MORGEN-EPIC cohorts	39,678 (Prospect-EPIC Fs, MORGEN-EPIC Ms and Fs), of which 19,837 in the derivation sample and 19,841 in the replication sample Prospect-EPIC: 50–69; MORGEN-EPIC: 20–64 Follow-up: 1993–2007	FFQ: 1 y; SA; reproducible and valid; 178 FIs (31 FGs)
Greve et al., 2016, Germany (56) IDEFICS Fair quality	Baseline survey of the German subsample of the IDEFICS study (a European longitudinal multicentre study in children and infants from 8 European countries)	1791 children 2–9 Follow-up: none	FFQ: NA reference period; SA (caregiver); NA reproducibility and validity; 45 FIs (15 FGs)

(*Continued*)

TABLE 1 (*Continued*)

Reference, location, title of study, quality	Study design	Subjects: *n*, age (y), and follow-up	Questionnaire
Hong et al., 2016, China (34) NA Good quality	Validation study of FFQ; subsample of 250 participants from the community-based, cross-sectional, nutrition and health survey in Nanjing, presenting a multistage random sampling design based on 6 communities of residents	203 31–80 (mean: 50.4 ± 12) Follow-up: 1 y	FFQ: 1 y; IA; reproducibility and validity to be assessed in this study; 87 FIs; FFQ completed twice (FFQ1 and FFQ2), at the beginning (June 2014) and end (May 2015) of the studyFour 3-consecutive day (including 2 weekdays and 1 weekend day in a usual week) 24HRs collected at intervals of 3 mo during the 1-y period by trained interviewers 28 FGs common to all dietary sources
Hu et al., 1999, USA (Massachusetts) (9) HPFS Good quality	HPFS: prospective cohort study of US M health professionals started in 1986; validation study of the FFQ used in the 1986 wave of the HPFS cohort study; random sample of cohort members (men) from the Boston area	127 Ms 40–75 y at baseline in 1986 Follow-up: 6–7 mo for validity analysis, 1 y for reproducibility analysis	FFQ: 1 y; SA; reproducibility and validity to be assessed in this study; 131 FIs FFQ1: completed during the following years FFQ2: completed 1 y after FFQ1 Two 7-d DRs 6–7 mo apart DR1: completed ∼3 mo after FFQ1 DR2: completed 2–3 mo before FFQ2; 1217 DR food codes used for creating FGs 40 FGs common to all dietary sources
Judd et al., 2014, USA (24) REGARDS Fair quality	REGARDS: population-based random sample of black and white individuals designed to oversample black participants and people residing in the stroke belt (8 US states)	21,636 >45 Follow-up: none	FFQ: 1 y; SA; NA reproducibility, but valid; 107 FIs (58 FGs, but final analysis on 56 FGs due to low communalities and zero consumption)
Khani et al., 2004, Sweden (33) SMC Fair quality	SMC: population-based cohort based on a mammography screening in 2 counties in central Sweden from 1987 to 1990 with 57,881 Fs who had completed the baseline SMC FFQ Validation study of the SMC FFQ; 2 random samples, 1 for FFQ reproducibility assessment and the other for FFQ validity assessment, reference FFQ completed at baseline for both samples	197 Fs included in the FFQ reproducibility sample; 111 Fs included in the FFQ validity sample 40–74 at baseline Follow-up: 1 y	FFQ: 6 mo; SA; reproducibility and validity to be assessed in this study; 60 FIs FFQ1: completed at baseline within the reproducibility sample FFQ2: completed 1 y after FFQ1 within the reproducibility sample FFQ: completed at baseline within the validity sample Four 7-d open-ended weighted DRs administered 3 mo apart, covering a year; 543 DR food codes matched to the FFQ items 26 FGs common to all dietary sources
Lau et al., 2008, Denmark (48) Inter99 Study Fair quality	Age- and sex-stratified random sample of participants of a health survey derived from baseline data of the population-based intervention study Inter99 (1999–2001), which included residents from the southwestern part of Copenhagen County	6563 (3372 Fs; 3191 Ms) 30–60 (mean: 46.3 ± 7.9) Follow-up: none	FFQ: 1 mo; SA; NA reproducibility, valid; 198 FIs (34 FGs)

(*Continued*)

mo. Reproducibility and/or relative validity of the FFQs were directly assessed within the 9 validation studies included in the review (9, 11, 30–34, 37, 40); in addition, 14 articles reported on a previous assessment of FFQ reproducibility and/or relative validity (10, 13, 24, 25, 27, 28, 40, 41, 48, 49, 51–53, 55), whereas 9 articles did not report any information (29, 36, 39, 43–45, 47, 56) or declared that they did not test for FFQ reproducibility and/or relative validity (54).

A different dietary assessment tool was used in 16 articles, including the 9 articles based on validation studies of FFQs (9, 11, 30–34, 37, 40). In 7 articles, information from 1 (35)

TABLE 1 (*Continued*)

Reference, location, title of study, quality	Study design	Subjects: *n*, age (y), and follow-up	Questionnaire
Liu et al., 2015, China (32) NA Poor quality	Validation study of a new FFQ developed from an NCI FFQ to capture DPs of rural Chinese population; random sample of subjects from an underdeveloped rural area of southwest China, free of chronic malignant diseases	179 40–70 at baseline in 2012 (mean: 55 ± 8.2) Follow-up: 1 y	FFQ: 1 y; IA; reproducibility and validity to be assessed in this study; 131 FIs FFQ1: completed at baseline FFQ2: completed 1 y after FFQ1 Six 3-d 24HRs completed between the 2 FFQs (eighteen 24HRs in 1 y, three 24HRs every 2 mo, on consecutive days, given by 2 weekdays and 1 weekend day) 18 FGs common to all dietary sources
Lo Siou et al., 2011, Canada (43) Tomorrow Project Fair quality	Tomorrow Project: longitudinal cohort study with 2-stage random sampling design including Albertans Ms and Fs with no personal history of cancer recruited between 2001 and 2007; subset of participants with complete data by November 2007	16,674 (6445 Ms; 10,229 Fs) 35–69 (mean: 50.5 ± 9.1 for Ms, and 50.5 ± 9.2 for Fs) Follow-up: none	FFQ: 1 y; SA; NA reproducibility and validity; 284 FIs (55 FGs)
Loy and Jan Mohamed, 2013, Malaysia (37) USM Birth Cohort Study Good quality	Validation study of the FFQ from USM Birth Cohort Study, based on a convenience sample of pregnant healthy Fs from the northeast of Peninsular Malaysia	162 pregnant Fs 19–40 (mean: 28.67) Follow-up: mid-pregnancy to late pregnancy	FFQ; 6 mo of pregnancy; IA; validity to be assessed in this study; 82 FIs; FFQ conducted immediately after completing the 24HRs in late pregnancy Six 24HRs, three 24HRs in mid (mean gestation: 15.6 wk) and late (mean gestation: 34.3 wk) pregnancy (2 weekdays and 1 weekend dietary intake)23 FGs common to all dietary sources
Maskarinec et al., 2000, USA (Hawaii) (52) NA Fair quality	Cross-sectional study based on an ethnically diverse population, with recruitment at different mammography facilities on Oahu	514 Fs 35–85 (mean: 53.9 ± 10.1) Follow-up: not applicable	FFQ: NA reference period; SA; valid; ∼209 FIs (39 FGs, but final analysis on 23 FGs due to skewness in FG distributions)
McCann et al., 2001, USA (New York) (45) Western New York Diet Study Fair quality	Western New York Diet Study: case-control study on endometrial cancer with population-based controls frequency-matched to cases on age and county of residence, conducted between October 1986 and March 1991 in the Buffalo area	1095 (232 cases; 863 controls) 40–85 for cases Follow-up: not applicable	FFQ: 2 y; IA; NA reproducibility and validity; 190 FIs (different numbers of FGs in the analysis corresponding to 3 different food grouping schemes: 168 FGs, as to useable information from FFQ, 56 FGs, as to nutrient content and use, and 36 FGs, as to USDA suggestions)
McNaughton et al., 2005, UK (35) Medical Research Council National Survey of Health and Development (1946 British Birth Cohort) Good quality	1946 British Birth Cohort: longitudinal study based on a social class–stratified, random sample of 5362 singleton births in England, Scotland, or Wales during the first week of March, 1946, with 21 occasions for collecting information throughout the life course until current article; data from 1989 interview	2265 subjects who completed the 48HR recall and the DR in 1989 43 in 1989 Follow-up: none	One 48HR at interview; one 5-d DR completed in the 5 d following the 48HR collection; one 24HR recall relative to the 24-h period preceding the interview 56 FGs common to all dietary sources

(*Continued*)

or multiple administrations of the same 24HR format was collected, with the number of collecting occasions ranging from 2 (50) to 18 (6 × 3 consecutive-day 24HRs) (32) and completion of the form in different combinations of time occasions and consecutive/nonconsecutive days; a DR was used in 10 articles, with reference time periods varying from 3 d (13, 40) to 7 d (9, 31, 33, 47), weighing system adopted (30, 33, 38, 47) or not, and single (13, 30, 35, 39, 40, 47) or multiple (9, 31, 33, 38) administrations of the same tools provided.

TABLE 1 (*Continued*)

Reference, location, title of study, quality	Study design	Subjects: *n*, age (y), and follow-up	Questionnaire
Nanri et al., 2012, Japan (31) JPHC Poor quality	Validation study of JPHC study FFQ; subsample of married couples from 5-y follow-up survey of the JPHC study (cohort 1: baseline 1990, and cohort 2: baseline 1993) who provided complete information on 2 FFQs and DRs	498 (244 Ms and 254 Fs, 290 in cohort 1 and 289 in cohort 2) Cohort 1: 40–59 at baseline; cohort 2: 40–69 at baseline Follow-up: 1 y	FFQ: 1 y; SA; reproducibility and validity to be assessed in this study; 147 FIs, but 134 FIs used for the final analysis FFQ_R: completed 1 y after or before FFQ_V FFQ_V: completed after DRs, and compared with DR 28-d or 14-d DRs completed during the course of 1 y [i.e., 7-d DRs collected 4 (or 2) times at 3-mo (or 6-mo) intervals during the years]; 558 DR FIs matched to 134 FFQ FIs 48 FGs common to all dietary sources
Newby et al., 2006, Sweden (10) SMC Good quality	SMC: population-based cohort based on a mammography screening in 2 counties in central Sweden from 1987 to 1990; subsample of SMC including healthy Fs at baseline with complete information on FFQ1 and FFQ2	33,840 Fs Mean: 52 at baseline (all Fs born between 1914 and 1948) Follow-up: from 1987–1990 to 1997–onwards	FFQ1 (1987–1990): 6 mo; SA; reproducible and valid; 67 FIs (29 FGs) FFQ2 (1997): 1 y; SA; based on the 1987 reproducible and valid FFQ; 97 FIs (32 FGs); mean time interval between FFQs: 8.8 y
Newby et al., 2006, Sweden (27) SMC Good quality	SMC: population-based cohort based on a mammography screening in 2 counties in central Sweden from 1987 to 1990; subsample of SMC including healthy Fs at baseline with complete information on FFQ1 and FFQ2	33,840 Fs Mean: 52 at baseline (all Fs born between 1914 and 1948) Follow-up: from 1987–1990 to 1997, 9 y of follow-up	FFQ1 (1987–1990): 6 mo; SA; reproducible and valid; 67 FIs (29 FGs FFQ2 (1997): 1 y; SA; based on the 1987 reproducible and valid FFQ; 97 FIs (32 FGs)
Northstone et al., 2008, UK (36) ALSPAC Fair quality	ALSPAC: longitudinal cohort study including a sample of pregnant Fs resident in the former Avon Health Authority with expected delivery date between April 1, 1991 and December 31, 1992; subset of ALSPAC study including Fs during pregnancy (1 wave)	12,053 pregnant Fs Age: NA Follow-up: NA	FFQ: NA reference period; SA; NA reproducibility and validity; NA FIs (44 FGs)
Okubo et al., 2010, Japan (38) NA Good quality	Cross-sectional study including apparently healthy volunteer Fs and their husbands from 3 areas of Japan [rural and urban Osaka (urban), Nagano (rural inland) and Tottori (rural coastal)]; Fs of 30–69 y, such that 8 Fs were equally distributed in each 10-y age stratum, but no age requirement for Ms	184 (92 Fs; 92 Ms) 31–69 for Fs (mean: 49.6 ± 11.4); 32–76 for Ms (mean: 52.8 ± 12.1) Follow-up: not applicable	DHQ; 1 mo; SA, valid; 150 FIs (145 effective FIs); DHQ administered 4 times (1 for each season over 1 y), 2 d before the start of the DRs Four 4-d weighed DRs (1 in each season over 1 y); 3 weekdays and 1 weekend day; 1299 FIs (1259 FIs used) 30 FGs common to all dietary sources

(*Continued*)

Regardless of the dietary assessment tool used, the number of FGs defined from the available food items ranged from 15 (56) to 56 (24, 35), with a median value of 30.5 FGs included in the statistical analysis. When information from >1 dietary source was available, the same food grouping scheme was adopted across the different sources in all the articles (9, 11, 13, 30–35, 37–40, 47).

Among the selected articles, 11 (29%) were based on studies of "good" quality, 17 (45%) on studies of "fair" quality, and 10 (26%) on studies of "poor" quality.

Tables 2–4 present details of DP identification method, methods for assessing DP reproducibility and/or validity, and main results on reproducibility and validity. Details of DP composition are presented in **Supplemental Tables 2**–

TABLE 1 (*Continued*)

Reference, location, title of study, quality	Study design	Subjects: *n*, age (y), and follow-up	Questionnaire
Park et al., 2005, USA (Hawaii and Los Angeles) (53) Hawaii-Los Angeles Multiethnic Cohort Study Poor quality	Baseline wave of the Multiethnic Cohort Study including the 5 principal ethnic groups (African Americans, Hawaiians, Japanese Americans, Latinos, and Whites) who lived in Hawaii and Los Angeles	195,298 45–75 Follow-up: none	FFQ: NA reference period; SA; valid; NA FIs (30 FGs, but final analysis on 20 FGs due to null values and nonnormality in FG distributions)
Ryman et al., 2015, USA (southwest Alaska) (54) CANHR Fair quality	CANHR: cohort study based on a convenience sampling of Alaska native (Yup'ik or Cup'ik) adults participating in CANHR study and completing at least 1 FFQ between September 2009 and May 2013	358 for EFA (1st FFQ, September 2009 to August 2011), 272 for CFA (1st FFQ, September 2011 to May 2013), 113 for test-retest (2nd FFQ, September 2009 to May 2013) >18 (median: 37; IQR: 23–54, in September 2009) Follow-up: September 2009 to May 2013	CANHR FFQ: 1 y; IA; 163 FIs (22 FGs, but final CFA on 18 FGs); not tested for reproducibility and validity FFQ1 in September 2009 to August 2011 for EFA (358 subjects) FFQ1 in September 2011 to May 2013 for CFA (272 subjects) FFQ2 in September 2009 to May 2013 for test-retest (113 subjects)
Sauvageot et al., 2017 Luxembourg, Belgium, and France (44) NESCaV Good quality	NESCaV: cross-border cardiovascular health population–based cross-sectional study, based on a stratified random sample of 3133 subjects recruited from 3 neighboring regions (Grand Duchy of Luxembourg, Wallonia in Belgium, and Lorraine in France) from the greater region	2298 18–69 Follow-up: not applicable	FFQ: 2 y; NA SA; NA reproducibility and validity; 134 FIs (45 FGs)
Schulze et al., 2003, Germany (55) EPIC-Potsdam Good quality	EPIC-Potsdam: cohort study including 27,548 Ms and Fs; Fs without a previous diagnosis of hypertension or intake of antihypertensive medication within a 4-wk period prior to the baseline examination were included at baseline, between August 1994 and September 1998	10,489 Fs, divided into learning (1937 Fs with normal blood pressure) and study (8552 Fs followed for 2–4 y for incident hypertension, and including 123 incident verified cases) samples 35–64 at baseline Follow-up: 2–4 y (until May 15, 2002)	FFQ: 1 y; SA; reproducible and valid; 148 FIs (44 FGs)
Togo et al., 2004, Denmark (29) MONICA Poor quality	Three consecutive surveys from MONICA project, including at baseline (M-82) a random sample of Danish citizens who lived in the western part of the Copenhagen County and had 30, 40, 50, and 60 y at baseline and further reexamined in 1987–1988 (M-87) and 1993–1994 (M-93)	2436 subjects participating in all 3 surveys, including 1806 subjects in M-82 30, or 40, or 50, or 60 at baseline in 1982–1984 Follow-up: at 5 y (1987–1988) and 11 y (1993–1994)	FFQ: 1 y; NA SA; NA reproducibility and validity; 26 FIs (21 FGs)
Togo et al., 2003, Denmark (47) MONICA Poor quality	Danish part of MONICA 1 (1982–1984) survey, including a random sample of Danish citizens who lived in the western part of the Copenhagen County and had 30, 40, 50, and 60 y at baseline	3785 (879 Ms and 927 Fs) 30, or 40, or 50, or 60 at baseline in 1982–1984 Follow-up: none	FFQ: 1 y; NA SA; NA reproducibility and validity; 26 FIs 7-d weighted DR completed in a normal week within 3 wk following the baseline investigation; 111 FIs 21 FGs common to all dietary sources

(*Continued*)

4. Among the 38 articles included, 30 performed PCA, EFA, or CFA and 6 performed CA (25, 42–44, 46, 56), whereas 2 articles carried out both EFA/CFA and CA (40, 51). In addition, 7 (22%) of the articles that carried out EFA or PCA assessed matrix factorability before starting the statistical analysis (30, 32, 34, 37, 40, 41, 50) (data not shown).

Table 2 concerns reproducibility of DPs derived from different statistical solutions, with 8 articles considering PCA/EFA (26, 36, 41, 45, 48, 50, 51, 57) and 6 considering CA (42–44, 46, 51, 56). The proposed research questions dealt with: *1*) input variable preprocessing—that is, adjustment by energy intake (26, 36, 42), standardization (46), and

TABLE 1 (*Continued*)

Reference, location, title of study, quality	Study design	Subjects: *n*, age (y), and follow-up	Questionnaire
Varraso et al., 2012, France and Spain (57) EGEA2-France, Spanish PAC-COPD Poor quality	EGEA2-France: cross-sectional study, 2003–2007 (12-y follow-up of EGEA study, which is a case-control and family asthma study); Spanish PAC-COPD, 2004–2007: cross-sectional study of patients hospitalized for the first time for a COPD exacerbation between 2004 and 2006	EGEA2-France: 1236; Spanish PAC-COPD: 274 EGEA2-France: mean: 43 ± 16; Spanish PAC-COPD: mean: 68 ± 8 Follow-up: not applicable	EGEA2-France: FFQ: 1 y; SA; based on a validated FFQ; 118 FIs (46 FGs); Spanish PAC-COPD: FFQ: 2 y; IA; NA reproducible and valid; 122 FIs (43 FGs all shared with EGEA2-France FGs)
Weismayer et al., 2006, Sweden (28) SMC Poor quality	SMC: population-based cohort based on a mammography screening in 2 counties in central Sweden from 1987 to 1990; subsample of SMC including 4 randomly selected subsamples of 1000 Fs each (giving a total of 4000 Fs), who completed 2 identical FFQs, to avoid survey learning effects	3606 Fs (871, 864, 887, and 967, at 4, 5, 6, 7 y after baseline) 49–70 Follow-up: 4, 5, 6, 7 y after baseline depending on the subsample	FFQ (1987–1990): 6 mo; SA; reproducible and valid; 67 FIs (25 FGs) FFQ completed at baseline and after 4, 5, 6, or 7 y depending on the subsample
Wirfalt et al., 2000, Sweden (46) MDC Fair quality	MDC: population-based prospective cohort study in Malmo, with baseline examinations conducted from March 1991 to October 1996; subset of participants with complete dietary data belonging to a substudy of the MDC study	5357 50–73 for Ms and 45–73 for Fs Follow-up: none	Modified DHQ combining a 7-d menu book with a 168-item FFQ: NA reference period; IA; reproducibility and validity assessed; 48 original FGs, but 43 FGs used in the final analysis due to negligible energy contribution and nonconsumption

[1] ALSPAC, Avon Longitudinal Study of Parents and Children; ATBC, Alpha-Tocopherol Beta-Carotene Cancer Prevention Study; CANHR, Center for Alaska Native Health Research study; CFA, confirmatory factor analysis; COPD, chronic obstructive pulmonary disease; DHQ, diet history questionnaire; DIETSCAN, DIETary patternS and CANcer in four European countries project; DR, dietary record; E3N, Mutuelle Generale de l'Education Nationale (EPIC-France); EFA, exploratory factor analysis; EGEA2-France, Epidemiological Study on the Genetics and Environment of Asthma 2—France; EPIC-NL, European Prospective Investigation into Cancer and Nutrition—the Netherlands; EPIC-Potsdam, European Prospective Investigation into Cancer and Nutrition—Potsdam; FFQ_R, food-frequency questionnaire from the reproducibility study; FFQ_V, food-frequency questionnaire from the relative validity study; FFQ1/FFQ2/FFQ3, food-frequency questionnaire at time 1, 2, or 3; FG, food group; FI, food item; HPFS, Health Professionals Follow-up Study; IA, interviewer-administered; IDEFICS, Identification and Prevention of Dietary and Lifestyle-induced Health Effects in Children and Infants; JPHC, Japan Public Health Center-Based Prospective Study; m24HR, mean 24-h recall; MDC, Malmo Diet and Cancer study; MONICA, MONItoring of trends and determinants in CArdiovascular disease; NA, not available; NCI, National Cancer Institute; NESCaV, Nutrition, Environment and Cardiovascular Health; NLCS, Netherlands Cohort Study on Diet and Cancer; ORDET, Ormoni e Dieta nella Eziologia dei Tumori in Italy; PAC-COPD, Phenotype and Course of Chronic Obstructive Pulmonary Disease study—Spain; REGARDS, Reasons for Geographic and Racial Differences in Stroke; SA, self-administered; SMC, Swedish Mammography Cohort; TLGS, Teheran Lipid and Glucose Study; USM, Universiti Sains Malaysia; 24HR, 24-h recall; 48HR, 48-h recall.

dichotomization (26); *2*) number of input variables (45) and subjects (57) to be included in the analysis; *3*) solution method for CA (43, 44, 56); *4*) rotation method for PCA/EFA (41, 48) and CFA (50); *5*) number of DPs to retain (25, 43, 44, 51); and *6*) score calculation—natural compared with applied (i.e., calculated using loadings from a separate PCA on subsample 1 and data from subsample 2) scores—in PCA (48). One article (25) proposed the comparison of different statistical solutions within the assessment of DP stability over time.

Concerning input variable preprocessing, 2 articles considered adjustment by energy intake with the residual method (26, 36) in PCA/EFA, whereas a third (42) considered percentage daily energy contribution compared with number of servings in CA; in the comparison between unadjusted and energy-adjusted solutions, 1 article used the correlation coefficient (36) and another (26) the Procrustes rotation method. Independently of the statistical approach and type of adjustment used, the conclusions on the comparison between energy-adjusted and unadjusted solutions were similar across articles (Supplemental Table 2): *1*) with PCA/EFA, the DPs extracted were generally similar (in terms of loadings and percentages of explained variances); *2*) with CA, the DPs were similar (in terms of higher/lower mean intakes of the FGs characterizing the clusters) and subgroups with high-energy contribution were consistently clustered across solutions; *3*) when available, correlation coefficients were >0.8 between similar DPs under the 2 solutions; *4*) DPs with high loadings on energy-contributing FGs were lost with energy adjustment (36); and *5*) the ability of CA to differentiate FGs with higher-than-mean intakes seemed higher with number of servings variables (42).

In addition, 2 articles considered the effect of standardizing or not FG intakes (expressed as percentage of daily energy intake) in CA (46) and of dichotomizing FGs with >75% of nonusers (26). In the former case (46), both the approaches led to well-separated and interpretable 6-cluster solutions that were stable and equivalent as to discriminant analysis; however, composition and number of subjects per cluster were different. An unstandardized solution was suggested

TABLE 2 Reproducibility of a posteriori dietary patterns across statistical solutions[1]

Reference, location, title of study	DP identification methods	Percent explained variance (no. of factors) or CFA/CA model	Assessment of reproducibility/validity	Main results
Bailey et al., 2006 USA (Pennsylvania) (42) Geisinger Rural Aging Study	Separate CAs using either number of servings or percent daily energy contribution from the same FG and according to the same CA approach; NA algorithm (PROC FASTCLUS); Euclidean distance; varying number of cluster from 2 to 6; screeplot of eigenvalues and within-cluster sum of squares plot to choose the optimal number of clusters	Not applicable, 2-cluster solution chosen examining screeplot of eigenvalues and within-cluster sum of squares plot	*Reproducibility*: No formal assessment	*Reproducibility*: Both methods consistently clustered subgroups with high energy contribution (e.g., fats and oils and dairy desserts); clusters resulting from the percent energy method were less likely to discern differences between FG and in particular to differentiate fruit and vegetable subgroups, as compared to number of servings method
Balder et al., 2003 Netherlands, Sweden, Finland, and Italy (26) DIETSCAN (NLCS, SMC, ATBC, ORDET)	Separate PCFAs on each of the 4 studies: standardization and separate analysis by sex; within each study, sensitivity analyses assessing the effect of: 1. untransformed vs. dichotomized variables (for FG with >75% of nonusers); 2. unadjusted vs. energy-adjusted variables using residual method; 3. solutions with 2–6 factors; 4. split-half analysis using the Procrustes rotation to compare different solutions; scree test to assess the final number of factors to retain in a range from 2 to 6 factors; varimax rotation; loading ≥|0.35| cut-off	NLCS: 23 (5) with Ms, 23.2 (5) with Fs; ORDET: 28.5 (4); SMC: 21.8 (4); ATBC: 20.3 (3); final results based on unadjusted variables for energy	*Reproducibility*: comparison of different scenarios within each study with Procrustes rotation; *cross-study reproducibility*: no formal assessment	*Reproducibility*: 1. Dichotomization: no effect (correlations of 0.98–1.00 on the diagonal of the Procrustes rotation matrix and low mutual correlations between factors); 2. Energy-adjustment: when using the energy-adjusted FG, the factor solutions were mostly comparable with the unadjusted factor solutions; mainly the DPs with high loadings on energy-contributing FG changed; by using energy-adjusted food variables, substitution of foods such as brown vs. white bread and low fat vs. medium and full-fat dairy products became more important, but other DPs unaffected by adjustment for energy (high correlations on the diagonal of the Procrustes rotation matrix); 3. Solutions with 2–6 factors: use of the Procrustes rotation matrix to track similar DPs across solutions with different number of factors: study-specific numbers of factors described with percentages of explained variance; 4. Split-half analysis: very similar results on the 2 subsamples *Cross-study reproducibility*: Two of the identified DPs were qualitatively similar across studies and between Ms and Fs
Bountziouka and Panagiotakos, 2012 Greece (41) NA	Separate PCAs conducted on the 2 administrations of the FFQ with different rotation methods; EIG >1; varimax and quartimax rotation among the orthogonal rotations and promax and oblimin rotation among the non-ortoghonal rotations; loading >|0.30| cut-off	Unrotated: 38 (4) with FFQ1 data and 40 (4) with FFQ2 data; Varimax rotation: 32.5 (4) with FFQ1 data and 35.6 (4) with FFQ2 data; Quartimax rotation: 32.8 (4) with FFQ1 data and 38.7 (4) with FFQ2 data; Promax rotation: NA (3); Oblimin rotation: NA (3)	*Reproducibility*: Kendall tau-b correlation coefficient between corresponding scores derived from solutions at different time-points with no rotation and with different rotation methods; Bland–Altman method (with 95% LOA) between scores from solutions at different time-points with no rotation and with different rotation methods	*Reproducibility*: 1. Unrotated solutions: All the 4 identified DPs were qualitatively similar and the following measures witnessed a good agreement between scores at the 2 time-points; Kendall tau-b correlation coefficient between FFQ1 and FFQ2 scores ranged from 0.50 to 0.63 (all $P < 0.0001$); Bland–Altman method: mean differences were equal to 0 but wide LOA especially for the LOW-FAT DP; 2. Orthogonal rotation solutions: 3 DPs were qualitatively similar across the 2 orthogonal solutions, but the agreement was low-to-moderate between scores at the 2 time-points; Kendall tau-b correlation coefficient between FFQ1 and FFQ2 scores ranged from 0.15 to 0.44 for the varimax (all $P < 0.0001$) and from 0.28 to 0.46 for the quartimax rotation method (all $P < 0.0001$); Bland–Altman method: mean differences were equal to 0, but wider LOA than with unrotated solutions; from both approaches, better agreement with quartimax (than varimax) rotation; 3. Non-orthogonal rotation solutions: 3 DPs were qualitatively similar, but the agreement was low-to-moderate between scores at the 2 time-points; Kendall tau-b correlation coefficient between FFQ1 and FFQ2 scores ranged from 0.21 to 0.41 for the promax (all $P < 0.0001$) and from 0.31 to 0.46 for the oblimin rotation method (all $P < 0.0001$); Bland–Altman method: mean differences were equal to 0 but wider LOA than with unrotated solution; from both approaches, better agreement with oblimin (than promax) rotation

(*Continued*)

TABLE 2 (*Continued*)

Reference, location, title of study	DP identification methods	Percent explained variance (no. of factors) or CFA/CA model	Assessment of reproducibility/validity	Main results
Castro et al., 2015 Brazil (50)Health Survey of the City of São Paulo	EFA: adjustment for within-person variation via Multiple Source Method; robust maximum likelihood estimation; EIG >1, scree test, interpretability; varimax among the orthogonal rotations and promax (power = 4) and oblimin rotation among the non-orthogonal rotations; alphanumeric labelling; CFA: loading ≥|0.20| or |0.25| cut-offs on EFA results based on different rotation methods; robust maximum likelihood estimation; adjusted chi-squared test, CFI, NNFI, RMSEA (90% CI), and SRMR	EFA: ∼10 with any rotation method used (2); CFA: 2-factor model with |0.25| cut-off and promax rotation method	*Reproducibility and validity*: CFA; different cut-off for FG inclusion; within CFA with and without different cut-offs for FG inclusion, comparison of rotation methods	*Validity*: 1. CFA with |0.20| cut-off: regardless of rotation method, factor loadings were statistically significant for all DPs ($P < 0.05$) and similar to those from EFA; [Reproducibility: promax and oblimin produced DPs with small but significant correlations ($r = 0.17$, $P < 0.01$); irrespective of rotation method, unacceptable model fits except for SRMR (SRMR < 0.08)]; 2. CFA with |0.25| cut-off: regardless of rotation method, factor loadings were statistically significant for all DPs ($P < 0.05$) and similar to those from EFA; [Reproducibility: better model fit with promax (best values of CFI, NNFI, RMSEA, and SRMR) and then varimax, and last oblimin rotation solution (CFI and NNFI <0.90); small but significant correlations between factors, with both promax ($r = 0.19$, $P < 0.01$) and oblimin rotations ($r = 0.18$, $P < 0.01$)]
Dekker et al., 2013 Netherlands (25) Doetinchem Cohort Study	CA: percentage energy contributed variables (nutrient density); *k*-means algorithm; bootstrap and internal cluster validity indexes (Calinski–Harabasz index, Davies–Bouldin index, and prediction-strength method) to assess the optimal number of clusters to retain between 2 and 6 clusters; labelling based on FG that contributed the highest percentage of total energy compared with other DPs within the same survey (≥ 40% higher energy indicated an important FG); robustness analysis with partitioning around medoids method	Not applicable, 2-cluster solution chosen according to Jaccard similarity indexes and internal cluster validity indexes	*Reproducibility*: internal cluster validity and stability (Jaccard indexes with 0.85 cut-off) indexes; *Stability over time*: 1. Stability of DPs over time in terms of contribution of a FG to total energy between the 2 clusters within the same survey (*t*-test, 99% CI, highly important FG were those with >1.4 time the percentage of total energy contributed for one compared to the other cluster by any FG) and comparison of the differences across surveys with a 5% cut-off; 2. Transitions of individuals between DPs over time: proportion of stable eaters (those assigned to the same cluster) and transitioners (those assigned to different clusters) in all 3 surveys and in survey 2 and 4 (over the higher 10-year period); relative change in mean percentage of total energy a specific FG contributed from survey 2 to survey 4 between individuals with stable and unstable behavior	*Reproducibility*: 1. Internal cluster stability: highly stable clusters, with Jaccard indexes >0.85 for most cluster numbers from 2 to 6, but highest stability for the 2-cluster solution; 2. Internal cluster validity: indexes pointing to 2-cluster solution, although with some exceptions; *Stability over time*: 1. Stability of DPs over time in terms of contribution of a FG to total energy: the 2 DPs were similar in all 3 surveys in terms of percentages of total energy contributed by relevant FG within each survey, although with small differences in FG composition across surveys (i.e. soft drinks with sugar and high-fiber cereals); the 2 DPs retained their relative difference in FG intake at each of the surveys, with FG relative intakes in each DP not changing >5% per survey; low-fiber bread was the only exception, with relative differences being equal to −7.06, −13.1, and −4.56 percentage of total energy contributed in surveys 2, 3, and 4 respectively, so 2 changes were >5%, but the third was not; 2. Transitions of individuals between DPs over time: 30.7% of the 4007 subjects with complete FFQ information were stable eaters assigned to HIGH-FIBER BREAD DP in all 3 surveys and 11.1% were stable eaters assigned to LOW-FIBER BREAD DP in all 3 surveys, giving a total of 41.8%; when comparing surveys 2 and 4 on the the longest time frame (10 y), 57.8% of participants assigned to HIGH-FIBER BREAD DP in both surveys, 15.2% assigned to LOW-FIBER BREAD DP at both surveys, 18.7% went from the HIGH- to LOW-FIBER BREAD DP, and 9.6% went from the LOW- to HIGH-FIBER BREAD DP; among stable eaters over time, no significant differences in percentage energy contributed by important FG was found during the 10-y period; transitioners had higher relative differences in percentage of energy intake for important FG than stable eaters (0.27–3.01 as compared to 0.86–1.88)

(*Continued*)

TABLE 2 (*Continued*)

Reference, location, title of study	DP identification methods	Percent explained variance (no. of factors) or CFA/CA model	Assessment of reproducibility/validity	Main results
Fransen et al., 2014 Netherlands (51) EPIC-NL	PCA: percentage energy contributed variables from both subsamples and the whole study population based on varying number of factors retained from 2 to 6; EIG >1, scree test, scree test optimal coordinate, interpretability; varimax rotation; alphanumeric labelling; EFA: percentage energy contributed variables from both subsamples and the whole study population based on varying number of factors retained from 2 to 6; EIG >1, scree test, scree test optimal coordinate, interpretability; varimax rotation; alphanumeric labelling; CA: top-coding of percentage energy contributed variables from both subsamples and the whole study population; *k*-means algorithm; Calinski–Harabasz and Davies–Bouldin indexes to assess the number of clusters to retain; CFA: loading ≥|0.25| cut-offs on PCA results (with a different number of DPs) for variables in the replication sample; loading ≥|0.20| cut-offs to name DPs	PCA/EFA: NA (2); CA: 2-cluster solution according to Calinski–Harabasz and Davies–Bouldin indexes; CFA: 3-factor model chosen according to confirmation success measure	*Reproducibility*: 1. Comparison of results from either PCA/EFA or CA on derivation and replication samples; 2. Comparison of results from either PCA/EFA or CA on derivation and whole samples; 3. Cluster stability with Jaccard similaries; 4. Internal validity indexes for PCA/EFA (EIG>1, scree test, scree test optimal coordinate, interpretability) and CA (Calinski–Harabasz and Davies–Bouldin) to identify the number of DPs to retain; *Validity*: CFA on replication sample starting from PCA/EFA on derivation sample with indexes of confirmation success (ratio of FG not confirmed to the total number of FGs and deviations in factor loadings between PCA/EFA and CFA)	*Reproducibility*: 1. Comparison between derivation and replication samples: PCA/EFA: good reproducibility; CA: good reproducibility (small deviations between the 2 subsamples, although increasing with increasing number of clusters); 2. Comparison between derivation and whole samples: PCA/EFA: almost identical DPs on the subsamples and whole population study; CA: almost identical clusters on the subsamples and whole population study; 3. Cluster stability: highly stable cluster solutions (Jaccard similaries for all solutions >0.85), with the best solution given by 2 clusters; 4. Internal validity indexes: PCA/EFA: no optimal number of DPs to retain common to all indexes (EIG >1: 11 DPs, scree test: 3 DPs, scree test optimal coordinate: 8 DPs); CA: 2-cluster solution was optimal according to the Calinski–Harabasz and Davies–Bouldin indexes; *Validity*: CFA on replication sample starting from PCA/EFA on derivation sample: high concordance between confirmation success measures; different confirmation success indexes between DPs within the same solution; all solutions contained 1 or more poorly confirmed DP (deviation >30%); 3-component solution was better confirmed that the others
Greve et al., 2016 Germany (56) IDEFICS	CA: rescaled relative frequencies (variances equal to 1); *k*-means (10,000 starting values), Ward's method and Gaussian mixture models (with automatic model selection via the Bayesian Information Criterion) in comparison; varying number of clusters to retain between 2 and 6; labelling based on the difference between the cluster-specific mean consumption frequency and the overall mean consumption frequency measured in units of overall SDs for the FG	Not applicable, 3-cluster solution chosen because of the highest similarities of the cluster solutions derived with each method	*Reproducibility*: ARI to assess pairwise agreement between clustering solutions	*Reproducibility*: Very little agreement between the 3 clustering methods; for all possible numbers of solutions, the Gaussian mixture model solution was constantly more similar to the *k*-means solution than to the Ward's solution; the best fitting Gaussian mixture model was the one that allowed the variances of the food consumption frequencies to vary within and between clusters; comparing the 3 clustering methods, the solutions with 3 clusters were most similar to each other (ARI equal to 0.47 comparing Gaussian mixture model vs. *k*-means, 0.23 for Gaussian mixture model vs. Ward's method, and 0.20 for *k*-means vs. Ward's method)
Lau et al., 2008 Denmark (48)Inter99 Study	Subsample 1: PCA 1: overall analysis and separate analyses by sex; PCFA; scree test, interpretability; varimax and promax rotations compared; loading ≥|0.40| cut-off; Subsample 1: PCA 2: as PCA 1 but including only FI whose loading was ≥|0.40| cut-off; Subsample 2: PCA 3: overall analysis and separate analyses by sex; same criteria of PCA 1; natural scores; Subsample 2: PCA 4: overall analysis and separate analyses by sex; same criteria of PCA 1; applied scores with PCA 1-based loadings; Subsample 1: CFA: loading ≥|0.40| cut-off on PCA 1 results; RMSEA	PCA 1: 17.1 (2) for entire subsample 1, 17.0 (2) for Ms, and 15.4 (2) for Fs; PCA 2, 3, and 4: NA (2) CFA: No model selection	*Reproducibility*: Pearson correlation coefficient between scores based on PCA 1 and PCA 2 in subsample 1; Pearson correlation coefficient between scores based on PCA 3 and PCA 4 in subsample 2; Bland–Altman plot between scores based on PCA 1 and PCA 2 in subsample 1, RV (95% CI of the difference of factor scores/95% CI of the average of factor scores) measure; Bland–Altman plot between scores based on PCA 3 and PCA 4 in subsample 2, with RV; *Validity*: CFA	*Reproducibility*: Rotation method on PCA 1: no significant differences in the final DPs derived from varimax vs. promax transformation, so promax rotation used for the PCA 1 analysis; Pearson correlation coefficient between scores based on PCA 1 and PCA 2 in subsample 1 was equal to 0.93 ($P < 0.0001$) for TRADITIONAL and MODERN DPs; Pearson correlation coefficient between scores based on PCA 3 (natural scores) and PCA 4 (applied scores) in subsample 2 was equal to 0.89, 0.98, and 0.90 ($P < 0.0001$) for the TRADITIONAL DP in all, Fs and Ms, respectively, and 0.89, 0.99, and 0.93 ($P < 0.0001$) for MODERN DP in all, Fs and Ms, respectively; Bland–Altman method: no systematic bias between scores based on PCA 1 and PCA 2 in subsample 1; relatively poor agreement (RV = 39.9% for TRADITIONAL DP and 37.6% for MODERN DP and PCA 1 and PCA 2 scores); no systematic bias between scores based on PCA 3 and PCA 4 in subsample 2; relatively poor agreement (RV = 47.5% for TRADITIONAL DP and 47.7% for MODERN DP and PCA 3 and PCA 4 scores); for Fs acceptable RV, whereas for Ms larger variations than for Fs; *Validity*: CFA: good fit (RMSEA equal to 0.008 < 0.10)

(*Continued*)

TABLE 2 (*Continued*)

Reference, location, title of study	DP identification methods	Percent explained variance (no. of factors) or CFA/CA model	Assessment of reproducibility/validity	Main results
Lo Siou et al., 2011 Canada (43) Tomorrow Project	Separate CAs using 3 different algorithms (*k*-means, Ward's minimum variance, and flexible beta with beta equal to −0.25 and −0.50): standardization [(value − minimum) divided by range] of the percentage of daily total energy intake; varying number of clusters from 2 to 7; between- versus within-cluster variance criterion to choose the optimal number of clusters; checking of potential outliers but no need to remove them	Not applicable, 4-cluster solution chosen for Ms according to median (natural) log-transformed ratios of between- versus within-cluster variances for the 55 FGs (best cluster had many FGs with large ratios) and with number of clusters varying from 2 to 7 obtained from applying the *k*-means method, and 3-cluster solution chosen for Fs according to interpretability of the results	*Reproducibility*: 1. Optimal clustering method: separately for Ms and Fs, average values over 20 repetitions for Hubert and Arabie's ARI and kappa and Cramer's V statistics to identify the optimal clustering method based on a split-half cross validation approach considering the different numbers of clusters; 2. Optimal number of clusters: median log-ratio value of between- versus within-cluster variances for the 55 FGs (best cluster had many FGs with large ratios) and with number of clusters varying from 2 to 7 obtained from applying the *k*-means method	*Reproducibility*: 1. Optimal clustering method: for Ms, as the number of clusters increased, the agreement and association between cluster assignments decreased when the *k*-means and Ward's methods were applied; a similar pattern was observed for Fs with the *k*-means method; agreement and association between cluster assignments remained low when applying the flexible-beta method; compared with the other 2 clustering methods, the *k*-means method had the highest reproducibility of the cluster solutions for Ms and Fs and with all different numbers of clusters; 2. Optimal number of clusters: in Ms, the median log-ratio value jumped from −3.45 to −3.03 between the 3-cluster and 4-cluster solutions, suggesting the optimal number of clusters for Ms was 4; in Fs, the median log-ratio values varied little across different numbers of clusters, suggesting no clear choice for the number of clusters
McCann et al., 2001 USA (New York) (45) Western New York Diet Study	Separate PCAs for each of the 3 food classification methods: controls-only PCA; percentage of variance explained by each factor, interpretability; varimax rotation; descriptive labelling; loading ≥0.30 or ≤ −0.20 cut-offs used for the calculation of factor scores	7.7 (2) with 168 FGs data, 13.4 (2) with 56 FGs data, and 16.9 (2) with 36 FGs data	*Reproducibility*: percentage exact agreement of classification along the diagonal for tertiles of DP scores by the 3 food classification schemes	*Reproducibility*: Food classification method affected neither the number nor character of the DPs identified, although total variance explained in food use increased as the detail included in the PCA decreased (∼8%, with 168 FGs to ∼17%, with 36 FGs); Percentage exact agreement: for both DPs, exact agreement in tertile classification decreased as the difference in the number of items used for PCA increased; for the HEALTHY DPs, almost half the subjects were misclassified on DP score by the broader food-use classification method including 36 FGs, as compared to 168 FGs; for the HIGH FAT DPs, the effect was similar but less dramatic, with percentage exact agreement decreasing from 81% (168 FGs vs. 56 FGs) to 76% (168 FGs vs. 36 FGs)
Northstone et al., 2008 UK (36) ALSPAC	Separate PCAs on unadjusted (weekly frequency of consumption) and adjusted (residual method) dietary variables: standardization; scree test, interpretability; varimax rotation; loading >|0.3| cut-off	32.4 (5) with unadjusted data and 26.9 (4) with energy-adjusted data	*Reproducibility*: Pearson correlation coefficient between scores from similar DPs on the unadjusted and energy-adjusted data	*Reproducibility*: Slight differences seen in terms of components extracted and factor loadings obtained; strong correlations (all >0.8) between scores from analogous unadjusted and energy-adjusted DPs; PROCESSED DP obtained from the unadjusted data was negatively correlated with both HEALTH-CONSCIOUS and CONFECTIONERY DPs obtained using the energy-adjusted data (−0.538 and −0.492, respectively)
Sauvageot et al., 2017 Luxembourg, Belgium, and France (44)NESCaV	Separate CAs using 3 different algorithms (*k*-means, *k*-medians, and Ward's minimum variance): standardization [(value − minimum) divided by range] of the residuals calculated according to Willett method; varying number of clusters from 2 to 6; 20 repetitions of the algorithm; for *k*-means and *k*-medians, 1000 runs with different random starting seeds, and solution that had the minimum total within-cluster sum of squares distances was selected; stability measure representing empirical misclassification rate across solutions on training and test samples (with *k*-nearest-means classifier for *k*-means and Ward's method and *k*-nearest-medians classifier for the *k*-medians algorithm), Cramer's V and ARI to choose the optimal combination of clustering method and number of clusters; truncation of >6 SDs values	Not applicable, 3-cluster solution with *k*-means chosen according to Cramer's V and ARI	*Reproducibility*: Optimal clustering method and number of clusters: box-plots and average values over 20 repetitions of each algorithm for each index	*Reproducibility*: Regardless of stability indices and number of clusters, more stable solutions were obtained with *k*-means; the most stable solution was obtained with 3 clusters

(*Continued*)

TABLE 2 (*Continued*)

Reference, location, title of study	DP identification methods	Percent explained variance (no. of factors) or CFA/CA model	Assessment of reproducibility/validity	Main results
Varraso et al., 2012 France and Spain (57)EGEA2-France, Spanish PAC-COPD	PCA and CFA used as equivalent approaches on 1000 randomly selected samples from 4 different set-ups: 1. 100% of EGEA2-France study; 2. 50% of EGEA2-France study; 3. 25% of EGEA2-France study; 4. 100% of Spanish PAC-COPD study; PCA: scree-plot, interpretability; varimax rotation; distribution of the factor loading of FG to each DP represented via box-plot and median loading >|0.30| as cut-off; CFA: not based on previous EFA; 4 different models tested (3-factor and 2-factor models with correlated latent variables, 3-factor and 2-factor models with independent latent variables); chi-squared test, GFI, and RMSEA; distribution of the factor loading of FG to each DP represented via box-plot and median loading >|0.30| as cut-off	PCA: NA (3); CFA: 3-factor model with no correlation among latent variables (highest GFI and lowest RMSEA)	*Reproducibility and validity*: statistical properties (min, quartile 1, median, quartile 3, max) of the distribution of the factor loading of each FG to each DP in each of the 4 subsamples considered	*Reproducibility and validity*: Two consistent DPs were identified by CFA in each of the subsamples, whereas PCA led to less interpretable (smaller median of factor loadings and higher dispersion) DPs, especially for the smallest sample
Wirfalt et al., 2000 Sweden (46)MDC	Separate CAs using 2 different input variable formats: standardization or not of the percentage of daily total energy intake; *k*-means algorithm; varying number of clusters from 2 to 10; interpretability (cluster size and ability to differentiate FG intakes) to choose the optimal number of clusters	Not applicable, 6-cluster solution chosen according to interpretability of results	*Reproducibility*: 1. Optimal number of clusters: no formal assessment; 2. Choice of the optimal input variable format: for each set of input variables, discriminant analysis after assuming the optimal 6-cluster solution (discriminant function chosen with all 43 FGs and with stepwise regression to identify FGs contributing significantly to the formation of clusters)	*Reproducibility*: 1. Optimal number of clusters: the 6-cluster solution produced reasonably sized and well-separated clusters for both input variable formats considered; 2. Choice of the optimal input variable format: the 6-cluster solution identified for each set of input variables was reproducible: with standard discriminant analysis, the agreement between actual and predicted cluster allocation ranged between 91.0% and 95.2% for the unstandardized variables, and between 91.1% and 100% for the *z*-scored variables; when using the stepwise function of the discriminant analysis, 18 unstandardized variables and 31 *z*-scored variables contributed significantly to the predicted cluster allocations

[1] ALSPAC, Avon Longitudinal Study of Parents and Children; ARI, adjusted Rand index; ATBC, Alpha-Tocopherol Beta-Carotene Cancer Prevention Study; CA, cluster analysis; CFA, confirmatory factor analysis; CFI, comparative fit index; CI, confidence interval; DIETSCAN, DIETary patternS and CANcer in four European countries project; DP, dietary pattern; EFA, exploratory factor analysis; EGEA2-France, Epidemiological Study on the Genetics and Environment of Asthma 2—France; EIG, Eigenvalue; EPIC-NL, European Prospective Investigation into Cancer and Nutrition—the Netherlands; F, female; FFQ, food-frequency questionnaire; FFQ1/FFQ2/FFQ3, food-frequency questionnaire at time 1, 2, or 3; FG, food groups; FI, food items; GFI: goodness of fit index; IDEFICS, Identification and Prevention of Dietary and Lifestyle-induced Health Effects in Children and Infants; LOA, limits of agreement; M, male; MDC, Malmo Diet and Cancer study; NA, not available; NESCaV, Nutrition, Environment and Cardiovascular Health; NLCS, Netherlands Cohort Study on diet and cancer; NNFI, non-normed fit index or Tucker-Lewis index; ORDET, Ormoni e Dieta nella Eziologia dei Tumori in Italy; PAC-COPD, Phenotype and Course of Chronic Obstructive Pulmonary Disease study–Spain; PCA, principal component analysis; PCFA, principal component factor analysis; RMSEA, root mean square error of approximation; RV, relative variation; SMC, Swedish Mammography Cohort; SRMR, standardized root mean square residual.

because standardized variables just allowed isolation of 1 or a few clusters including extreme individuals, whereas the remaining clusters were all very similar. In the latter case (26), the Procrustes rotation method confirmed that dichotomizing variables with a high percentage of nonusers did not affect the FGs with significant factor loadings, the magnitude of the factor loadings, or the explained variance, and thus the order of the extracted DPs.

Two articles assessed the effect of different numbers of: *1*) input variables (from different food grouping schemes) in PCA-derived DPs (45); or *2*) subjects to include in PCA and CFA (not based on previous EFA) in a study combining 2 studies from France and Spain (57). In the former case-control study on endometrial cancer (45), the DPs identified according to 3 food grouping schemes (168 usable FFQ items, or 56 FGs from nutrient content or use classification, or 36 FGs from the USDA suggestions) were not materially different except for the total variance explained in food use, which increased as the detail included in the PCA decreased (up to ∼17% with 36 FGs). However, for both DPs, exact agreement in tertile classification decreased as the difference in the number of items used for PCA increased, and misclassification rates were higher for the "healthy" DP. In the latter article (57), PCA and CFA were carried out on 1000 randomly selected samples from 4 different setups—100%, 50%, or 25% of the French study (1236 subjects) and 100% of the Spanish study (274 subjects). From the bootstrap-based distributions of the factor loadings to each FG for each DP, a more consistent set of CFA-based, rather than PCA-based, DPs was identified across the setups. CFA-based DPs outperformed PCA-based ones, especially when smaller sample sizes were considered.

Three articles (43, 44, 56) were concerned with the choice of the optimal algorithm for performing CA and compared the mostly used *k*-means and Ward minimum variance algorithms with flexible beta (43), with *k*-medians (44), or with Gaussian mixture models (56), in a complex setup of varying number of clusters. Together with them (43, 44, 56), another 2 articles assessed the simpler issue of the optimal number of clusters to retain when a *k*-means algorithm was carried out (25, 51). Finally, Fransen et al. (51) considered the same research question for PCA and EFA too. In the comparison of clustering algorithms (43, 44, 56), the *k*-means provided the highest reproducibility of the cluster solutions with all different numbers of clusters, compared with the Ward minimum variance (43, 44), flexible beta algorithm (43), and *k*-medians (44). For all possible numbers of solutions, the Gaussian mixture model was more similar to the *k*-means algorithm than to the Ward algorithm; however, the best Gaussian mixture model identified from the data implied FG variances to vary within and between clusters and it was therefore more general than the equivalent model subsumed by the *k*-means algorithm (56). Regarding the choice of the optimal number of clusters, 1 article (43) adopted a split-half cross-validation approach and used the median log-ratio value of between- compared with within-cluster variances of the available FGs, after having previously identified the optimal algorithm as the *k*-means algorithm—with the Hubert–Arabie adjusted Rand index (ARI), kappa, and Cramer V statistics; a similar article (44) identified the optimal combination of clustering method and number of clusters by using the box-plot and average value (over 20 repetitions of each algorithm) of the distribution of Cramer V statistic and ARI; the article by Greve et al. (56) assumed that the optimal number of clusters was the one that provided more similar solutions across the different algorithms, based on pairwise comparisons of ARI values.

Finally, when no algorithm choice was allowed and the *k*-means algorithm was carried out (25, 51), the optimal number of clusters to retain was identified with internal cluster validity (e.g., Calinski–Harabasz index, Davies–Bouldin index, and prediction-strength method) and stability (e.g., Jaccard) indexes (25, 51); for PCA/EFA the usual criteria for identifying the optimal number of factors to retain were adopted (51).

Three articles were concerned with the choice of the optimal rotation method in EFA (41) or PCA (48) and of a combination of rotation method and cutoff for FG inclusion in EFA and CFA (50). Based on 2 close administrations (at 15 d apart) of the same FFQ, the first article (41) assessed the effect on DP repeatability of 2 orthogonal (varimax and quartimax) and 2 nonorthogonal (promax and oblimin) rotations, compared with an unrotated solution. The main conclusions were: *1*) in the unrotated solutions, the identified DPs were similar over the 2 FFQ administrations, although the limits of agreement were wide; *2*) for either orthogonal or nonorthogonal rotation, the agreement was poorer between corresponding DPs at the 2 time points, compared with the unrotated solution; *3*) between the orthogonal rotations, a better agreement was found for the quartimax rotation; and *4*) between the nonorthogonal rotations, a better agreement was found for the oblimin rotation (41). Based on the baseline data from a population survey, the second article (48) concluded that DPs derived from varimax and promax rotations were qualitatively similar and opted for the promax solution, which allows correlations between DPs. Based on another population-based survey, the third article (50) assessed the effect on DP reproducibility of different cutoffs (i.e., |0.20| or |0.25|) for FG inclusion and rotation method (i.e., varimax, promax, and oblimin), with the following conclusions: *1*) a |0.25| cutoff for FG inclusion in EFA provided reproducible results for any rotation method; *2*) a |0.25| cutoff for FG inclusion in CFA defined a valid CFA model; and *3*) a better model fit was observed for CFA with promax and then varimax, and last oblimin rotation solution, with small but significant correlations between factors.

Finally, 1 article (48) assessed the difference between using natural and applied PCA-based scores. It concluded that: *1*) correlation coefficients between natural and applied scores for the same DP were high (≥0.89) and significant; *2*) no systematic bias was found in the Bland–Altman plot comparing natural and applied scores; and *3*) for both DPs, the agreement was relatively weak in men and only acceptable

in women, as indicated by the relative variation measure (48).

Table 3 concerns reproducibility and/or relative validity of DPs, with 7 articles assessing DP reproducibility and relative validity together (9, 11, 30–34), 7 articles assessing relative validity of DPs only (13, 35, 37–40, 47), and 1 article assessing DP reliability (54). All the articles derived DPs from PCA or EFA and 1 article additionally derived DPs with CA (40). Dietary patterns were separately identified on FFQ data at time 1 and 2 (9, 11, 30–34, 54), and/or on mean intakes from multiple administrations of the gold standard dietary assessment tool—mean 24HR (m24HR) or mean DR (9, 11, 31–34, 37, 38). The DP identification process was similar in all the articles and generally included a combination of eigenvalue >1, scree test, and interpretability to choose the number of DPs to retain, a varimax rotation to improve DP interpretation, and descriptive labeling for naming the identified DPs. Five articles (31, 38, 39, 47, 54) proposed standardization—with (47) or without Kaiser normalization (39)—or log-transformation of input variables (31, 38, 54) alone (54) or with an adjustment by energy intake through the residual method for either input variables (38) or DP scores (31).

The number of described DPs ranged from 2 to 5, with 47% of the articles naming and describing 2 DPs; however, 7 articles (9, 13, 32–34, 37, 39) reported the existence of additional DPs not common to all dietary sources (Supplemental Table 3). The described DPs were generally similar across different dietary sources (in terms of factor loadings and percentages of explained variance) and their names reflected these similarities; some variation in DP composition was reported across available dietary sources or different time points in 1 article (35), whereas in another article (40), additional DPs were identified for FFQ data only (Supplemental Table 3). The described DPs generally included a "healthy"/"health-aware"/"fruits" and "vegetables"/"prudent"/"Mediterranean" profile, and a "less healthy"/"Western"/"processed food(s)" pattern, but we also identified variants of "traditional" (11, 31, 34, 35, 38, 47, 54), "sweet-based" (34, 40, 47), "sandwich-based" (30, 35), or "alcohol-based" DPs (33, 34, 40) (Supplemental Table 3). Reproducibility of DPs was assessed with 1 (9, 11, 31, 33) or >1 statistical approaches (30, 32, 34); similarly, relative validity was assessed with 1 (9, 31, 33, 35, 47) or more (11, 13, 30, 32, 34, 37–40) approaches, and reliability was assessed with >1 statistical method (54). The intraclass correlation coefficient (11, 32, 34, 54), the (Pearson, Spearman, or Kendall) correlation coefficient (9, 11, 13, 30–35, 37–40, 47), the Bland–Altman method (11, 13, 30, 32, 34, 37–40), the proportions of subjects classified into the same, adjacent, opposite quantiles, and the weighted kappa coefficient (30, 32, 34, 37) were used alone or in combination for the assessment of reproducibility and/or relative validity. Partial, deattenuated or corrected correlation coefficients were also introduced in some articles to account for the effect of energy intake, and/or of repeated administrations of the gold standard dietary assessment tool (9, 11, 32, 33).

Among the 7 articles simultaneously assessing reproducibility and relative validity of DPs (9, 11, 30–34), the main results were: *1*) the different statistical approaches used led to concordant results, except for 1 article (30) where only the Bland–Altman approach consistently highlighted increasing differences in DP scores with increasing scores; *2*) under the same statistical approach, the assessment of DP reproducibility provided generally stronger results than relative validity (9, 11, 30, 31, 33, 34); and *3*) well-characterized DPs based on a few identifiable FGs were more likely to be reproducible and valid than DPs including different aspects of the diet simultaneously (Supplemental Table 3); for example, the "sandwich and drinks" DP (30), the "animal and plant protein" DP (34), and the "drinker" DP (33) had higher reproducibility and relative validity than others from the same articles.

Among the 7 articles assessing relative validity of DPs only (13, 35, 37–40, 47), we distinguished between those comparing FFQs and DRs (13, 39, 40, 47), the one comparing the FFQ with a 24HR (37), and those studies not based on FFQ data (35, 38). In the first group (13, 39, 40, 47), the relative validity of all DPs was questionable with any approach in 1 article (40) and it was poor for the "Western" DP in another article (39); however, the "healthy"/"prudent"/"green" DPs showed a higher degree of relative validity, compared with the corresponding "Western"/"Western/traditional" DPs in 3 articles (13, 39, 47). In contrast, when comparing FFQ-based DPs with those on m24HR (37), the "less-healthy" DP was found to be more valid than the "healthy" DP in pregnant women, although results for both DPs were stronger than in previous articles. When 24HR or 48-h recall (48HR) were compared with DR data (35), relative validity was moderate to good with 48HR-based DPs, but less strong with 24HR-based DPs; the "health-aware" DP showed the highest validity on the 48HR-based comparison. Finally, when a diet history questionnaire was compared with a DR (38), the "healthy" DP was found to be valid, but not the other 2 DPs, which showed wider limits of agreement in women, based on DR data.

When the reliability of CFA-based DPs was evaluated by Ryman et al. (54), composite reliability of DPs was good and similar across DPs, but test-retest reliability of DPs was moderate. In addition, indicator and test-retest reliabilities of CFA-based FGs were similar and poor to fair. The "processed foods" and the "fruits-and-vegetables" DPs showed better reliability overall.

Table 4 provides details on the 15 articles assessing construct validity of DPs through the application of CFA (10, 24, 27–29, 47, 48, 50–55) to validate previous EFA-based DPs or as an alternative 1-step approach to be compared with PCA/EFA (49, 57). Some of them used CFA-based DPs for assessing more general research questions on relative validity of DPs (47), DP reproducibility (50, 57) or reliability (54), DP stability over time (10, 27–29), or cross-study reproducibility (24); other studies simply used CFA to represent DPs of a population of interest in a more ideal way (48, 49, 51–53, 55).

TABLE 3 Reproducibility and/or relative validity of a posteriori dietary patterns[1]

Reference, location, title of study	DP identification methods	Percent explained variance (no. of factors) or CFA/CA model	Assessment of reproducibility/validity	Main results
Ambrosini et al., 2011, Australia (13) Western Australian pregnancy cohort (Raine) study	Separate EFAs (maximum likelihood method) conducted on the FFQ and DR data with all available information used (1613 subjects for FFQ and 822 subjects for DR data): EIG >1 on FFQ data only, scree test; varimax rotation; loading >|0.20| cutoff; 4 FGs removed from the final analysis due to small loadings on all factors	84 (2) with FFQ data and 53 (2) with DR data	*Relative validity*: Spearman correlation coefficient (crude and partial, with adjustment by total energy intake) and Bland–Altman method (with 95% LOA) between scores from FFQ and DR data	*Relative validity*: The identified DPs were similar although not identical in terms of loadings; modest Spearman correlation coefficient between DP scores from FFQ and DR given by 0.43 (crude) and 0.45 (partial and corrected) ($P < 0.001$) for HEALTHY DP and 0.27 (crude) and 0.36 (partial and corrected) ($P < 0.001$) for WESTERN DP; correlations improved after adjustment for energy intake Bland–Altman method: acceptable (not significantly different from 0) mean agreement for both DP scores; 95% LOA given by (−1.69, 1.65) for HEALTHY DP and (−1.89, 1.82) for WESTERN DP, so slightly narrower for HEALTHY DP; minor differences between girls and boys in all previous analyses
Asghari et al., 2012, Iran (11) TLGS	Separate PCFAs on FFQ1, FFQ2, FFQ3, and m24HR: scree test and interpretability; varimax rotation; descriptive labeling; applied scores from previous EFAs to data from FFQ3 were reported but their use was not clear	27.4 (2) with FFQ1 data, 31.6 (2) with FFQ2 data, 39.0 (3) with FFQ3 data, and 32.0 (2) with m24HR data	*Reproducibility*: intraclass correlation coefficient between scores from FFQ1 and FFQ2 data *Relative validity*: Spearman correlation coefficient, and Bland–Altman method (with 95% LOA) between scores from FFQ2 and scores from m24HR data, deattenuated correlation coefficient (Rosner and Willett formula) between each DP score to reduce the random within-person month-to-month variability in 24HR-based DPs *Stability over time*: intraclass correlation coefficient between continuous scores from FFQ2 and FFQ3 data, weighted kappa coefficient and proportions of subjects at the same quintile, adjacent quintile, and opposite quintile when comparing quintiles classification of factor scores between baseline and follow-up data	*Reproducibility*: intraclass correlation coefficients between FFQ1- and FFQ2-based scores equal to 0.72 ($P < 0.001$) for the IRANIAN TRADITIONAL DP, and 0.80 ($P < 0.001$) for the WESTERN DP *Relative validity*: crude and corrected Spearman correlation coefficients between FFQ2 and m24HR similar and equal to 0.48 for the IRANIAN TRADITIONAL and 0.75 for the WESTERN DPs Bland–Altman plot: 95% LOA for the difference between factor scores from FFQ2 and m24HR lay between −1.58 and 1.58 for the IRANIAN TRADITIONAL and between −1.33 and 1.33 for the WESTERN DP *Stability over time*: intraclass coefficients between FFQ2- and FFQ3-based scores equal to −0.09 ($P = 0.653$) for the IRANIAN TRADITIONAL and 0.49 ($P < 0.001$) for the WESTERN DPs; percentage of subjects at the same quintile higher for the WESTERN DP vs. the IRANIAN TRADITIONAL DP (27.1% vs. 20.2%); proportion of individuals at the opposite quintile reversed (35.8% vs. 41.5%) Weighted kappa coefficient: 0.09 (95% CI: −0.05, 0.23) for the IRANIAN TRADITIONAL and 0.20 (95 % CI: 0.05, 0.34) for the WESTERN DP

(*Continued*)

TABLE 3 (*Continued*)

Reference, location, title of study	DP identification methods	Percent explained variance (no. of factors) or CFA/CA model	Assessment of reproducibility/validity	Main results
Beck et al., 2012, New Zealand (30) NA	Separate PCFAs on FFQ1, FFQ2, and DR: EIG >1, scree test, interpretability; varimax rotation; descriptive labeling	~20 (2) with each of the 3 dietary sources	*Reproducibility*: Pearson correlation coefficient and Bland–Altman method (with 95% LOA) between scores from FFQ1 and FFQ2 data; weighted kappa coefficient and proportions of subjects at the same third, or the opposite third when comparing tertiles classification of factor scores between FFQ1 and FFQ2 data*Relative validity*: Pearson correlation coefficient and Bland–Altman method (with 95% LOA) between scores from FFQ1 and DR data; weighted kappa coefficient and proportions of subjects at the same third, or the opposite third when comparing tertiles classification of factor scores between FFQ1 and DR data	*Reproducibility*: good Pearson correlation coefficients between FFQ1 and FFQ2 DP scores (0.76 for the HEALTHY DP and 0.76 for the SANDWICH AND DRINKS DP; $P < 0.001$) Bland–Altman method: the difference between DP scores from FFQ1 and FFQ2 increased with increasing scores for both DPsCross-classification of DP scores: >50% of participants classified in the same third and <10% misclassified into the opposite third for both the DPs between FFQ1 and FFQ2; Weighted kappa coefficient between FFQ1 and FFQ2: moderate (HEALTHY) and good (SANDWICH AND DRINKS DP) *Relative validity*: reasonable Pearson correlation coefficients between FFQ1 and DR DP scores (0.34 for the HEALTHY DP and 0.62 for the SANDWICH AND DRINKS DP; ($P < 0.001$) Bland–Altman method: the difference between DP scores from FFQ1 and DR increased with increasing scores for both DPs Cross-classification of DP scores: >50% of participants classified in the same third and <10% misclassified into the opposite third for both the DPs between FFQ1 and DR;Weighted kappa coefficient between FFQ1 and DR DP scores fair (HEALTHY) and moderate (SANDWICH AND DRINKS)
Bountziouka et al., 2011, Greece (40) NA	Separate PCAs conducted on FFQ and DR data: EIG >1.4, scree test; varimax rotation; loading >|0.30| cut-off Separate CAs conducted on FFQ and DR data: *k*-means method; Euclidean and Mahalanobis distances; maximum achieved distances between cluster's centers; 2-, 3-, and 5-cluster solutions considered	PCA: 35 (4) with FFQ data and 29 (4) with DR dataCA: not applicable, 2-cluster solution chosen according to maximum achieved distances between cluster's centers	*Relative validity*: Kendall tau-b correlation coefficient between scores from FFQ and DR; Bland–Altman method (with 95% LOA) between scores from FFQ and DR; Kendall tau-b correlation coefficient and exact classification rate for CA	*Relative validity*: PCA: Kendall tau-b correlation coefficient: significant but low correlation coefficient, equal to 0.22 for the WESTERN and 0.23 for the MEDITERRANEAN DPs ($P < 0.001$ for both DPs) Bland–Altman method: 95% LOA given by −2.35, 2.30 for WESTERN and −2.23, 2.26 for MEDITERRANEAN DPCA: Kendall tau-b correlation coefficient: very good agreement between clusters derived from FFQ and DR (0.81, $P < 0.001$)Exact classification rate: 48% and 59% depending on the distance used
Crozier et al., 2008, UK (39) NA	Separate PCFAs conducted on FFQ and DR data: standardization; NA criteria for choosing the number of factors; NA rotation; descriptive labelling; Fisher–Yates transformation of scores to improve adherence to normality	15.9 (2) with FFQ data and 14.3 (2) with DR data	*Relative validity*: Pearson correlation coefficient between scores from FFQ and DR; Bland–Altman method (with 95% LOA) between scores from FFQ and DR	*Relative validity*: the corresponding DPs from FFQ and DR data were strikingly similar, especially the PRUDENT DP; Pearson correlation coefficients between FFQ and DR scores were 0.67 ($P < 0.001$) for PRUDENT DP and 0.35 ($P < 0.001$) for WESTERN DP Bland–Altman method: good agreement between scores from FFQ and DR for PRUDENT DP (95% of the differences lying within −1.58 and +1.58 SDs), but less good for WESTERN DP (95% of the differences lying within −2.22 and +2.22 SDs); consistently wider limits for the WESTERN DP with generally similar variations across characteristics

(*Continued*)

TABLE 3 (*Continued*)

Reference, location, title of study	DP identification methods	Percent explained variance (no. of factors) or CFA/CA model	Assessment of reproducibility/validity	Main results
Hong et al., 2016, China (34) NA	Separate EFAs on FFQ1, FFQ2, and m24HR: EIG, scree test, interpretability; varimax rotation; loading >0.30 cutoff	40.0 (4) for FFQ1 data, 44.9 (4) for FFQ2 data, and 32.4 (4) for m24HR data	*Reproducibility*: intraclass correlation coefficient between DP scores from FFQ1 and FFQ2 data Cross-classification: range of agreement rates for the same or adjacent quartile classifications and misclassification into opposite quartiles; kappa coefficient *Relative validity*: Pearson correlation coefficient between DP scores from FFQ1 and FFQ2, respectively, and m24HR data, after adjusting for energy intake using the residual method Cross-classification: range of agreement rates for the same or adjacent quartile classifications and misclassification into opposite quartiles; kappa coefficient Bland–Altman method and 95% LOA considering mFFQ, in comparison with m24HR scores	*Reproducibility*: the 4 derived DPs were qualitatively similar across the 3 sources of dietary data, although loadings were partly different; good intraclass correlation coefficient between DP scores from FFQ1 and FFQ2 data (>0.6 for all DPs, all $P < 0.001$)Cross-classification: range of agreement rates for the same or adjacent quartile classifications equal to 29.2–66.3% (both for ANIMAL AND PLANT PROTEIN DP, with adjacent and same quartile, respectively) and misclassification into opposite quartiles was <5% for all DPs Kappa coefficient: fair-to-moderate (range: 34–68% with minimum for NUTS AND SWEETS and maximum for ANIMAL AND PLANT PROTEIN DPs, respectively) *Relative validity*: reasonable adjusted Pearson correlation coefficient between DP scores from FFQ and m24HR data (range of adjusted values: 0.387–0.838 with minimum for CHINESE TRADITIONAL DP and maximum for ANIMAL AND PLANT PROTEIN DP) Cross-classification: range of agreement rates for the same or adjacent quartile classifications equal to 32.4 (for CHINESE TRADITIONAL DP, same quartile, FFQ1) to 47.0% (for ANIMAL AND PLANT PROTEIN DP, same quartile, FFQ1) and misclassification into opposite quartiles was <5% for all DPs Kappa coefficient: fair-to-moderate (range: 25.9–48.1% for BEVERAGE AND ALCOHOL DP with FFQ1 and ANIMAL AND PLANT PROTEIN with FFQ1, respectively) Bland–Altman method: mean agreement between DP scores derived from mFFQ and m24HR were not significantly different from 0 in all comparisons; mean differences were 0.0 (95% LOA: −1.03 to 1.04) for ANIMAL AND PLANT PROTEIN DP, 0.0 (95% LOA: −1.7 to 1.6) for NUTS AND SWEETS DP, −0.1 (95% LOA: −2.0 to 1.8) for CHINESE TRADITIONAL DP, and −0.2 (95% LOA: −1.9 to 1.5) for BEVERAGE AND ALCOHOL DP. From all statistical analyses, ANIMAL AND PLANT PROTEIN DP had better performance than the other DPs
Hu et al., 1999, USA (Massachusetts) (9) HPFS	Separate PCFAs on FFQ1, FFQ2, and mDRs: EIG >1, scree test, interpretability; varimax rotation; descriptive labeling	20 (2) with each of the dietary sources	*Reproducibility*: crude Pearson correlation coefficient between DP scores from FFQ1 and FFQ2 *Relative validity*: crude and corrected (for week-to-week variation in DRs) Pearson correlation coefficient between DP scores from either FFQ1 or FFQ2 and DR	*Reproducibility*: good crude Pearson correlation coefficient between DP scores from FFQ1 and FFQ2 (0.70 for the PRUDENT and 0.67 for the WESTERN DPs) *Relative validity*: (crude and corrected) Pearson correlation coefficients between DP scores from either FFQ1 or FFQ2 and DR ranged from 0.34 to 0.74

(*Continued*)

TABLE 3 (*Continued*)

Reference, location, title of study	DP identification methods	Percent explained variance (no. of factors) or CFA/CA model	Assessment of reproducibility/validity	Main results
Khani et al., 2004, Sweden (33) SMC	Separate PCFAs on FFQ1 and FFQ2 within the reproducibility sample, and on FFQ and mDRs within the validity sample: EIG >1.8; varimax rotation; descriptive labeling	Within the reproducibility sample: 29 (3) for FFQ1 data and 30 (3) for FFQ2 data Within the validity sample: 30 (3) for DR data and 34 (3) for FFQ data	*Reproducibility*: crude Spearman correlation coefficient between DP scores from FFQ1 and FFQ2 data *Relative validity*: crude and corrected (for unreproducibility of the FFQ) Spearman correlation coefficient between DP scores from FFQ and DR	*Reproducibility*: good crude Spearman correlation coefficient between DP scores from FFQ1 and FFQ2 data (range: 0.63–0.73 across DPs), with highest results for the DRINKER DP *Relative validity*: reasonable (crude and corrected) Spearman correlation coefficient between DP scores from FFQ and DR (range of crude values: 0.41–0.73; range of corrected values: 0.50–0.85), with highest results for the DRINKER DP
Liu et al., 2015, China (32) NA	Separate PCFAs on FFQ1, FFQ2, and m24HR: EIG >1.5, scree test, interpretability; varimax rotation; loading >|0.4| cutoff	30 (2)	*Reproducibility*: Pearson correlation coefficient (crude and partial, with adjustment for log10-transformation of total energy intake), intraclass correlation coefficient (to adjust for the effect of different scales of measures), and Bland–Altman method (with 95% LOA) between scores from FFQ1 and FFQ2 data; weighted kappa coefficient and proportions of subjects at the same third, or the opposite third when comparing tertiles classification of factor scores between FFQ1 and FFQ2 data *Relative validity*: Pearson correlation coefficient [crude, partial (with adjustment for log10-transformation of total energy intake), and deattenuated, to correct monthly and seasonal variation] and Bland–Altman method (with 95% LOA) between scores from either FFQ1 or FFQ2 and 24HR data; weighted kappa coefficient and proportions of subjects at the same third, or the opposite third when comparing tertiles classification of factor scores between either FFQ1 or FFQ2 and 24HR data	*Reproducibility*: between FFQ1 and FFQ2, crude Pearson correlation coefficients equal to 0.58 for the PRUDENT DP and 0.60 for the PROCESSED FOOD DP, partial Pearson correlation coefficient equal to 0.51 for PRUDENT DP and 0.56 for PROCESSED FOOD DP, intraclass correlation coefficient equal to 0.57 for PRUDENT DP and 0.55 for PROCESSED FOOD DP Bland–Altman method: divergence not obvious between DP scores on FFQ1 and FFQ2 Cross-classification analysis: >54% of the participants correctly classified into the same tertile and <9% misclassified into an opposite tertile for both DPs when 2 FFQs compared; moderate weighted kappa coefficient (0.45 for PRUDENT and 0.56 for PROCESSED FOOD) between the 2 FFQs *Relative validity*: between FFQs and 24HRs, crude Pearson correlation coefficients ranged from 0.45 to 0.64 for PRUDENT DP and from 0.46 to 0.50 for PROCESSED FOOD DP, deattenuated correlation coefficients ranged from 0.54 to 0.78 for the PRUDENT DP and from 0.55 to 0.61 for the PROCESSED FOOD DP; partial Pearson correlation coefficients ranged from 0.41 to 0.56 for the PRUDENT DP and from 0.42 to 0.44 for the PROCESSED FOOD DP Bland–Altman method: divergence not obvious between DP scores on FFQ1 or FFQ2 and 24HR data Cross-classification analysis: >54% of the participants correctly classified into the same tertile and <9% misclassified into an opposite tertile for both DPs when FFQs and 24HRs compared Moderate weighted kappa coefficient (range: 0.42–0.60 across the 2 DPs and FFQs)
Loy and Jan Mohamed, 2013, Malaysia (37) USM Birth Cohort Study	Separate PCAs on FFQ and m24HR: EIG >1, scree test, interpretability; varimax rotation; descriptive labeling	22.4 (2) with FFQ data and 20.7 (2) with m24HR data	*Relative validity*: Pearson correlation coefficient and Bland–Altman method (with 95% LOA) between scores from FFQ and 24HR data; weighted kappa coefficient and proportions of subjects at the same third, or the opposite third when comparing tertiles classification of factor scores between FFQ and m24HR data	*Relative validity*: relatively high Spearman correlation coefficient between DP scores from FFQ and m24HR data given by 0.59 (HEALTHY DP) and 0.63 (LESS-HEALTHY DP) ($P < 0.001$) Bland–Altman method: good agreement for both DPs, with 95% of the differences within ±1.87 SD (HEALTHY DP) and ±1.69 SD (LESS-HEALTHY DP); no association between the difference and the average for both DPs Cross-classification: acceptable (<10%) degrees of misclassification and lower than recommended percentage of classified in the same third (~50% or more) for both DPs: moderate (0.56) and good (0.72) agreement from weighted kappa coefficient for the HEALTHY and LESS-HEALTHY DPs, respectively; from all criteria, LESS-HEALTHY DP more valid than HEALTHY DP

(*Continued*)

TABLE 3 (*Continued*)

Reference, location, title of study	DP identification methods	Percent explained variance (no. of factors) or CFA/CA model	Assessment of reproducibility/validity	Main results
McNaughton et al., 2005, UK (35) Medical Research Council National Survey of Health and Development (1946 British Birth Cohort)	Separate PFCAs on 24HR recall, 48HR recall, and DR data: separate analyses by sex; EIG >1, scree test; varimax rotation; loading >0.3 cutoff	Range: 19 (5) with 24HR data to 22 (5) with DR data	*Relative validity*: correlation coefficient between scores from similar DPs across dietary assessment tools	*Relative validity*: five distinct DPs were identified using the DR and 48HR, but were less consistent on the 24HR data; moderate-to-good correlations between factor scores on 48HR and DR data (0.13–0.67, all $P < 0.001$), with the highest values for the HEALTH-AWARE DP in both Ms and Fs; correlations with 48HR data were higher than those between the 24HR and DR data (−0.01 to 0.59, with most P values < 0.001)
Nanri et al., 2012, Japan (31) JPHC	Separate PCAs on log-transformed data from FFQ_R, FFQ_V, and mDR data: separate analyses by sex; EIG >1, scree test, interpretability; varimax rotation; descriptive labeling; energy-adjusted scores using the residual method	In Ms: 23.9 for mDR data, 29.4 for FFQ_R data, and 26.5 for FFQ_V data (3); in Fs: 23.0 for mDR data, 24.9 for FFQ_R data, and 32.9 for FFQ_V data (3)	*Reproducibility*: Spearman correlation coefficient between DP scores from the FFQ_R and FFQ_V data in both Ms and Fs *Relative validity*: Spearman correlation coefficient between DP scores from mDR and FFQ_V data	*Reproducibility*: acceptable Spearman correlation coefficients between DP scores from the FFQ_R and FFQ_V data in both Ms and Fs for the 3 DPs (TRADITIONAL JAPANESE DP in Ms and WESTERNIZED JAPANESE DP in Fs given by 0.77 and 0.71, respectively, range of correlation coefficients: 0.55–0.77 across DPs) *Relative validity*: acceptable Spearman correlation coefficients between DP scores from mDR and FFQ_V (TRADITIONAL JAPANESE DP in Ms and in Fs given by 0.49 and 0.63, respectively; range of correlation coefficients: 0.32–0.63 across DPs)
Okubo et al., 2010, Japan (38) NA	Separate PCFAs conducted on DHQ1, mDHQ, and mDR data: log-transformation and adjustment by energy intake with residual method; separate analyses by sex; scree test, interpretability; varimax rotation; descriptive labeling	In Fs, 30.1 (3) with DHQ1 data, 31.2 (3) with mDHQ data, and 30.8 (3) with mDR data; in Ms, 21.5 (2) with DHQ1 data, 24.4 (2) with mDHQ data, and 25.8 (2) with mDR data	*Relative validity*: Pearson correlation coefficient between DHQ1 and mDR data and between mDHQ and mDR data; Bland–Altman method (with 95% LOA) between scores from DHQ1 and mDRs	*Relative validity*: the identified factor loadings were similar in magnitude and direction across DHQ1, mDHQ, and mDR data; Pearson correlation coefficients for the HEALTHY, WESTERN, and JAPANESE TRADITIONAL DPs in Fs were equal to 0.57, 0.36, and 0.44, and for the HEALTHY and WESTERN DPs in Ms were 0.62 and 0.56; when mDHQ was examined, correlation coefficients improved for Fs (0.45–0.69)Bland–Altman method: for both Ms and Fs, mean differences between scores derived from DHQ1 and DR were 0; 95% LOA for the difference between factor scores derived from DHQ1 and DR lay within −1.81 and 1.81 for HEALTHY, within −2.22 and 2.22 for WESTERN, and within −2.08 and 2.08 for JAPANESE TRADITIONAL DPs in Fs; and within −1.83 and 1.83 for the HEALTHY and within −1.71 and 1.71 for the WESTERN DPs in Ms; agreements generally improved between mDHQ and DR data

(*Continued*)

TABLE 3 (*Continued*)

Reference, location, title of study	DP identification methods	Percent explained variance (no. of factors) or CFA/CA model	Assessment of reproducibility/validity	Main results
Ryman et al., 2015 USA (southwest Alaska) (54) CANHR	EFA: log-transformation (base e) on 358 subjects; NA criteria for choosing the number of factors; NA rotation; loading ≥|0.60| cutoff CFA: loading ≥|0.35| cutoffs (and a priori knowledge of Alaska native diet) on EFA results on 272 subjects; 3-factor model with correlated factors; GFI, AGFI, RMSEA, CFI, and NNFI	EFA: NA (3); CFA: 3-factor model with correlated factors	*Validity*: CFA *Reliability*: composite reliability of DPs with CFA (squared standardized loadings and sum of the error variances), test-retest reliability of DPs with intraclass correlation coefficient (FFQ1 and FFQ2), indicator reliability of individual FGs included in the CFA-based DPs with CFA (square of the standardized factor loadings for each FG), and test-retest reliability of individual FGs included in the CFA-based DPs with intraclass correlation coefficient (FFQ1 and FFQ2)	*Validity*: CFA: significant and high (>0.40) standardized coefficients of FGs on the given factor, except for 1 FG; satisfactory goodness of fit indexes (GFI, AGFI, CFI, and NNFI values were 0.93, 0.91, 0.92, and 0.91, respectively, all >0.90, and RMSEA was equal to 0.004 < 0.005) *Reliability*: composite reliability of DPs: ranged from 0.56 to 0.73; test-retest reliability of DPs: ranged from 0.34 to 0.66; indicator reliability of individual FGs included in CFA-based DPs: ranged from 0.07 to 0.46; test-retest reliability of individual FGs included in CFA-based DPs: ranged from 0.11 to 0.50, with better reliability for market-based FGs
Togo et al., 2003, Denmark (47) MONICA	Separate PCFAs on FFQ and DR data (in octiles): separate analyses by sex; standardization with Kaiser normalization; scree test, interpretability; promax rotation; loading >|0.30| cutoff Separate CFAs on FFQ and DR data: loading ≥|0.30| cutoff on EFA results; polychoric correlation matrix; RMSEA; weighted least square variable estimates with robust SEs and mean- and variance-adjusted chi-square test statistic	In Ms: 30.5 (3) with FFQ data and 26.2 (3) with DR data; in Fs: 23.8 (2) with FFQ data and 19.8 (2) with DR data	*Relative validity*: Pearson correlation coefficient between scores based on FFQ and DR data *Validity*: CFA; Pearson correlation coefficient between scores based on EFA and CFA	*Relative validity*: EFA on FFQ and DR data: the identified DPs were very similar, although the percentages of explained variance were lower on DR data; Pearson correlation coefficient between FFQ-based and DR-based scores ranged between 0.34 (TRADITIONAL DP among Ms) and 0.61 (both GREEN DPs, among Ms and Fs) CFA on FFQ and DR data: Pearson correlation coefficient between FFQ-based and DR-based scores ranged between 0.37 (TRADITIONAL DP among Ms) and 0.64 (GREEN DP among Fs); higher correlations with CFA than with EFA *Validity* (EFA vs. CFA with the same dietary source): CFA-based DPs were similar across dietary sources and came from models with reassuring model fit (RMSEA <0.10 regardless of the dietary source and sex) FFQ data: Pearson correlation coefficient between EFA-based and CFA-based scores ranged between 0.91 (TRADITIONAL DP among Ms) and 0.96 (SWEET-TRADITIONAL DP among Fs) DR data: Pearson correlation coefficient between EFA-based and CFA-based scores ranged between 0.82 (GREEN DP among Ms) and 0.94 (both GREEN and SWEET-TRADITIONAL DPs among Fs); higher correlations were found when using the same dietary data

[1] AGFI, adjusted goodness-of-fit index; CA, cluster analysis; CANHR, Center for Alaska Native Health Research study; CFA, confirmatory factor analysis; CFI, comparative fit index; DHQ1/DHQ2/DHQ3, diet history questionnaire at time 1, 2, or 3; DP, dietary pattern; DR, dietary record; EFA, exploratory factor analysis; EIG, eigenvalue; FFQ_R, food-frequency questionnaire from the reproducibility study; FFQ_V, food-frequency questionnaire from the relative validity study; FFQ1/FFQ2/FFQ3, food-frequency questionnaire at time 1, 2, or 3; FG, food group; GFI, goodness-of-fit index; HPFS, Health Professionals Follow-Up Study; JPHC, Japan Public Health Center-based Prospective study; LOA, limits of agreement; m24HR, mean 24-h recall; m48HR, mean 48-h recall; mDHQ, mean diet history questionnaire; mDR, mean dietary record; mFFQ, mean food frequency questionnaire; MONICA, MONItoring of trends and determinants in CArdiovascular Disease; NA, not available; NNFI, nonnormed fit index or Tucker–Lewis index; PCA, principal component analysis; PCFA, principal component factor analysis; RMSEA, root mean square error of approximation; SMC, Swedish Mammography Cohort; TLGS, Teheran Lipid and Glucose Study; USM, Universiti Sains Malaysia; 24HR, 24-h recall; 48HR, 48-h recall.

TABLE 4 Construct validity of a posteriori dietary patterns[1]

Reference, location, title of study	DP identification methods	Percent explained variance (no. of factors) or CFA/CA model	Assessment of reproducibility/validity	Main results
Bedard et al., 2015, France (49) E3N (EPIC-France)	PCA: scree test, interpretability; varimax rotation; descriptive labeling CFA: not based on previous EFA; 4 different models tested (3-factor and 2-factor models with correlated latent variables, 3-factor and 2-factor models with independent latent variables); overall chi-square test of fit, GFI, and RMSEA with 90% CI	PCA: 24 (3); CFA: 3-factor model with no correlation among latent variables (highest GFI and lowest RMSEA)	*Validity*: CFA; Pearson correlation coefficient between corresponding scores from PCA and CFA	*Validity*: CFA: good fitting of selected model; Pearson correlation coefficients between corresponding scores from EFA and CFA ranged from 0.83 to 0.87
Castro et al., 2015, Brazil (50) Health Survey of the City of São Paulo	EFA: adjustment for within-person variation via multiple source method; robust maximum likelihood estimation; EIG >1, scree test, interpretability; varimax among the orthogonal rotations and promax (power = 4) and oblimin rotation among the nonorthogonal rotations; alphanumeric labelling CFA: loading ≥|0.20| or |0.25|cutoffs on EFA results based on different rotation methods; robust maximum likelihood estimation; adjusted chi-square test, CFI, NNFI, RMSEA (90% CI), and SRMR	EFA: ~10 with any rotation method used (2) CFA: 2-factor model with |0.25| cutoff and promax rotation method	*Reproducibility and validity*: CFA; different cutoff for FG inclusion; within CFA with and without different cutoffs for FG inclusion, comparison of rotation methods	*Validity*: 1. CFA with |0.20| cutoff: regardless of rotation method, factor loadings were statistically significant for all DPs ($P < 0.05$) and similar to those from EFA [reproducibility: promax and oblimin produced DPs with small but significant correlations ($r = 0.17$; $P < 0.01$); irrespective of rotation method, unacceptable model fits except for SRMR (SRMR <0.08)] 2. CFA with |0.25| cutoff: regardless of rotation method, factor loadings were statistically significant for all DPs ($P < 0.05$) and similar to those from EFA [reproducibility: better model fit with promax (best values of CFI, NNFI, RMSEA, and SRMR) and then varimax, and last oblimin rotation solution (CFI and NNFI <0.90); small but significant correlations between factors, with both promax ($r = 0.19$; $P < 0.01$) and oblimin rotations ($r = 0.18$; $P < 0.01$)]
Fransen et al., 2014, Netherlands (51) EPIC-NL	PCA: percentage energy contributed variables from both subsamples and the whole study population based on varying number of factors retained from 2 to 6; EIG >1, scree test, scree test optimal coordinate, interpretability; varimax rotation; alphanumeric labelling EFA: percentage energy contributed variables from both subsamples and the whole study population based on varying number of factors retained from 2 to 6; EIG >1, scree test, scree test optimal coordinate, interpretability; varimax rotation; alphanumeric labelling CA: top-coding of percentage energy contributed variables from both subsamples and the whole study population; *k*-means algorithm; Calinski–Harabasz and Davies–Bouldin indexes to assess the number of clusters to retain CFA: loading ≥|0.25| cutoffs on PCA results (with a different number of DPs) for variables in the replication sample; loading ≥|0.20| cutoffs to name DPs	PCA/EFA: NA (2); CA: 2-cluster solution according to Calinski–Harabasz and Davies–Bouldin indexes; CFA: 3-factor model chosen according to confirmation success measure	*Reproducibility*: 1. Comparison of results from either PCA/EFA or CA on derivation and replication samples 2. Comparison of results from either PCA/EFA or CA on derivation and whole samples 3. Cluster stability with Jaccard similarities 4. Internal validity indexes for PCA/EFA (EIG >1, scree test, scree test optimal coordinate, interpretability) and CA (Calinski–Harabasz and Davies–Bouldin) to identify the number of DPs to retain *Validity*: CFA on replication sample starting from PCA/EFA on derivation sample with indexes of confirmation success (ratio of FGs not confirmed to the total number of FGs and deviations in factor loadings between PCA/EFA and CFA)	*Reproducibility*: 1. Comparison between derivation and replication samples: PCA/EFA: good reproducibility; CA: good reproducibility (small deviations between the 2 subsamples, although increasing with increasing number of clusters) 2. Comparison between derivation and whole samples: PCA/EFA: almost identical DPs on the subsamples and whole population study; CA: almost identical clusters on the subsamples and whole population study 3. Cluster stability: highly stable cluster solutions (Jaccard similarities for all solutions >0.85), with the best solution given by 2 clusters 4. Internal validity indexes: PCA/EFA: no optimal number of DPs to retain common to all indexes (EIG >1: 11 DPs; scree test: 3 DPs; scree test optimal coordinate: 8 DPs); CA: 2-cluster solution was optimal according to the Calinski–Harabasz and Davies–Bouldin indexes *Validity*: CFA on replication sample starting from PCA/EFA on derivation sample: high concordance between confirmation success measures; different confirmation success indexes between DPs within the same solution; all solutions contained ≥1 poorly confirmed DP (deviation >30%); 3-component solution was better confirmed than the others

(*Continued*)

TABLE 4 (*Continued*)

Reference, location, title of study	DP identification methods	Percent explained variance (no. of factors) or CFA/CA model	Assessment of reproducibility/validity	Main results
Judd et al., 2014, USA (24) REGARDS	EFA on the first split-sample, CFA on the second split-sample, and final PCA on the whole sample as far as the model is correctly identified EFA: 3 separate PCAs by population subgroups [region (southeastern US stroke belt/nonbelt), sex (male/female), and race (black/white)] to identify the optimal number of factors in a range from 3 to 6 factors; EIG >1.5, scree test, interpretability of results from stratified PCAs; varimax rotation; descriptive labeling CFA: loading >|0.20| cutoff on EFA results; no different correlation structures specified; RMSEA and CFI	NA (5)	*Cross-study reproducibility*: CC determined for each stratification pair for each of the factor number solutions ("excellent" when the smallest coefficient was >0.8, "good" between 0.65 and 0.8, "acceptable" between 0.5 and 0.65, and "poor" <0.5) *Validity*: CFA	*Cross-study reproducibility*: PCA stratified by region of residence on the first half-sample: excellent CC for the 4- and 5-factor solutions, and acceptable CC for the 3- and 6-factor solutions PCA stratified by gender: good CC for the 5- and 6-factor solutions and poor CC for the 3- and 4-factor solutions PCA stratified by race: acceptable CC in the 5-factor solution, but poor CC for the other 3; the 5-factor solution had an acceptable congruence in all stratified analyses and it was interpretable, so this was the final model selected for CFA CFA on the second half-sample using the 5-factor solution: very good results, even when removing FGs with low factor loadings (RMSEA values <0.05)
Lau et al., 2008, Denmark (48) Inter99 Study	Subsample 1: PCA 1: overall analysis and separate analyses by sex; PCFA; scree test, interpretability; varimax and promax rotations compared; loading ≥|0.40| cutoff Subsample 1: PCA 2: as PCA 1 but including only FIs whose loading was ≥|0.40| cutoff Subsample 2: PCA 3: overall analysis and separate analyses by sex; same criteria of PCA 1; natural scores Subsample 2: PCA 4: overall analysis and separate analyses by sex; same criteria of PCA 1; applied scores with PCA 1-based loadings Subsample 1: CFA: loading ≥|0.40| cutoff on PCA 1 results; RMSEA	PCA 1: 17.1 (2) for entire subsample 1, 17.0 (2) for Ms, and 15.4 (2) for Fs; PCA 2, 3, and 4: NA (2); CFA: no model selection	*Reproducibility*: Pearson correlation coefficient between scores based on PCA 1 and PCA 2 in subsample 1; Pearson correlation coefficient between scores based on PCA 3 and PCA 4 in subsample 2; Bland–Altman plot between scores based on PCA 1 and PCA 2 in subsample 1, RV (95% CI of the difference of factor scores/95% CI of the average of factor scores) measure; Bland–Altman plot between scores based on PCA 3 and PCA 4 in subsample 2, with RV *Validity*: CFA	*Reproducibility*: rotation method on PCA 1: no significant differences in the final DPs derived from varimax vs. promax transformation, so promax rotation used for the PCA 1 analysis; Pearson correlation coefficient between scores based on PCA 1 and PCA 2 in subsample 1 was equal to 0.93 ($P < 0.0001$) for TRADITIONAL and MODERN DPs; Pearson correlation coefficients between scores based on PCA 3 (natural scores) and PCA 4 (applied scores) in subsample 2 were equal to 0.89, 0.98, and 0.90 ($P < 0.0001$) for the TRADITIONAL DP in all, Fs and Ms, respectively, and 0.89, 0.99, and 0.93 ($P < 0.0001$) for MODERN DP in all, Fs and Ms, respectively Bland–Altman method: no systematic bias between scores based on PCA 1 and PCA 2 in subsample 1; relatively poor agreement (RV = 39.9% for TRADITIONAL DP and 37.6% for MODERN DP and PCA 1 and PCA 2 scores); no systematic bias between scores based on PCA 3 and PCA 4 in subsample 2; relatively poor agreement (RV = 47.5% for TRADITIONAL DP and 47.7% for MODERN DP and PCA 3 and PCA 4 scores); for Fs acceptable RV, whereas for Ms larger variations than for Fs *Validity*: CFA: good fit (RMSEA equal to 0.008 < 0.10)
Maskarinec et al., 2000, USA (Hawaii) (52) NA	EFA: log-transformation (base e) on the first half of the population; scree test, interpretability; varimax rotation; loading ≥|0.60| cutoff CFA: loading ≥|0.60| cutoffs on EFA results for variables in the second half of the population; chi-square test, RMSEA, CFI, NNFI, parsimonious NFI; *t* test on factor loadings; final CFA results applied on the whole sample	EFA: 93 (4); CFA: no model selection	*Validity*: CFA	*Validity*: CFA: significant standardized coefficients of FGs on the given factor, but goodness of fit indexes slightly inappropriate (significant chi-square test $P < 0.0001$; RMSEA equal to 0.14 > 0.10; CFI equal to 0.82 < 0.90; NNFI equal to 0.83 < 0.90; parsimonious NFI equal to 0.68 > 0.60)

(*Continued*)

TABLE 4 (*Continued*)

Reference, location, title of study	DP identification methods	Percent explained variance (no. of factors) or CFA/CA model	Assessment of reproducibility/validity	Main results
Newby et al., 2006, Sweden (10) SMC	Separate PCFAs at each time point: scree test, interpretability; varimax rotation; descriptive labeling Separate CFAs at each time point: loading ≥|0.15| cutoff based on loadings ≥|0.20| cutoff from EFA results and a priori knowledge	PCFA: 35.4 (6) with FFQ1 (1987) data, 32.4 (6) with FFQ2 (1997) data CFA: no model selection	*Validity*: CFA; *Stability over time*: mean and SD intakes of CFA-based FGs at both time points and Spearman correlation coefficient between CFA-based FGs; Pearson correlation coefficient between DP scores at 2 time points; Pearson correlation coefficient between DP scores from PCFA and CFA at fixed time points	*Validity*: CFA, but no goodness-of-fit assessment or formal comparison with EFA; *Stability over time*: intakes of vegetables, fruit, seafood, refined grains, soda, sugary foods, and sweet baked goods increased over the time period, whereas intakes of meat and whole grains decreased over the time period; Spearman correlation coefficient between CFA-based FGs ranged from 0.23 to 0.70 (all $P < 0.0001$); Pearson correlation coefficient between DP scores in 1987 and 1997 ranged from 0.27 (WESTERN/SWEDISH DP) to 0.54 (ALCOHOL DP) for CFA-based DPs (all $P < 0.0001$) and were similar for PCFA-based DPs; Pearson correlation coefficient between DP scores from PCFA and CFA at fixed time points were ≥0.90 (all $P < 0.0001$)
Newby et al., 2006, Sweden (27) SMC	Separate PCFAs at each time point: scree test, interpretability; varimax rotation; descriptive labeling Separate CFAs at each time point: loading ≥|0.15| cutoff based on loadings ≥|0.20| cutoff from EFA results and a priori knowledge	PCFA: 35.4 (6) with FFQ1 (1987) data, 32.4 (6) with FFQ2 (1997) dataCFA: no model selection	*Validity*: CFA *Stability over time*: no formal assessment	*Validity*: CFA, but no goodness-of-fit assessment or formal comparison with EFA *Stability over time*: similar FG and factor loadings for each DP were seen in 1987 and 1997; some variation was observed for HEALTHY DP (seafood, poultry, and eggs also contributed to HEALTHY DP in 1987, whereas legumes and soy products contributed to HEALTHY DP in 1997)
Park et al., 2005, USA (Hawaii and Los Angeles) (53) Hawaii-Los Angeles Multiethnic Cohort Study	PCFA: Box–Cox transformation on the first half of the population and in the 10 separate ethnic-gender groups defined on this first half of the sample; EIG >1.25, scree test; interpretability; varimax rotation; loading ≥|0.60| cutoff to exclude other 7 FGs from the analysis CFA: loading ≥|0.60| cutoff on PCFA results for variables in the second half of the population and in the 10 separate ethnic-gender groups defined on this second half of the sample; RMSEA, CFI, and NNFI; *t* test on factor loadings; final PCFA results applied on the whole sample	PCFA: 63.5 (3); CFA: no model selection	*Validity*: CFA	*Validity*: CFA: significant and high (>0.6) standardized loadings (all $P < 0.001$); acceptable goodness of fit indexes (RMSEA equal to $0.095 < 0.10$; CFI equal to $0.90 = 0.90$; NNFI equal to $0.88 < 0.90$)
Ryman et al., 2015, USA (southwest Alaska) (54) CANHR	EFA: log-transformation (base e) on 358 subjects; NA criteria for choosing the number of factors; NA rotation; loading ≥|0.60| cutoff CFA: loading ≥|0.35| cutoffs (and a priori knowledge of Alaska native diet) on EFA results on 272 subjects; 3-factor model with correlated factors; GFI, AGFI, RMSEA, CFI, and NNFI	EFA: NA (3); CFA: 3-factor model with correlated factors	*Validity*: CFA *Reliability*: composite reliability of DPs with CFA (squared standardized loadings and sum of the error variances), test-retest reliability of DPs with intraclass correlation coefficient (FFQ1 and FFQ2), indicator reliability of individual FGs included in the CFA-based DPs with CFA (square of the standardized factor loadings for each FG), and test-retest reliability of individual FGs included in the CFA-based DPs with intraclass correlation coefficient (FFQ1 and FFQ2)	*Validity*: CFA: significant and high (>0.40) standardized coefficients of FGs on the given factor, except for 1 FG; satisfactory goodness of fit indexes (GFI, AGFI, CFI, and NNFI values were 0.93, 0.91, 0.92, and 0.91, respectively, all >0.90, and RMSEA was equal to $0.004 < 0.005$) *Reliability*: composite reliability of DPs ranged from 0.56 to 0.73; test-retest reliability of DPs ranged from 0.34 to 0.66; indicator reliability of individual FGs included in CFA-based DPs ranged from 0.07 to 0.46; test-retest reliability of individual FGs included in CFA-based DPs ranged from 0.11 to 0.50, with better reliability for market-based FGs

(*Continued*)

TABLE 4 (*Continued*)

Reference, location, title of study	DP identification methods	Percent explained variance (no. of factors) or CFA/CA model	Assessment of reproducibility/validity	Main results
Schulze et al., 2003, Germany (55) EPIC-Potsdam	EFA: on the learning sample with following reanalyses limiting the number of included FGs until 8 FGs; EIG >1, scree test; no rotation; descriptive labellingCFA: loading ≥0.40 cutoff on EFA results using the sample study; CFA: 2-factor model with uncorrelated factors; GFI, RMSEA, CFI, NNFI; simplified scores	EFA: NA (2); CFA: no model selection	*Validity*: CFA	*Validity*: CFA: significant standardized loadings; acceptable goodness of fit indexes, except for borderline significance of NNFI (GFI equal to 0.98 > 0.90; RMSEA equal to 0.07 < 0.10; CFI equal to 0.93 > 0.90; NNFI equal to 0.90 = 0.90)
Togo et al., 2004, Denmark (29) MONICA	EFA: on a subsample of the M-82 data (who filled a DR too); separate analyses by sex; scree test, interpretability; varimax rotation; descriptive labeling CFA: loading ≥|0.30| cutoff on EFA results; CFA: 3-factor model with correlated factors; CFA performed on M-82 data (all M-82 participants) and on the subgroup including M-82–87 data; to include diet information at 5-y follow-up, CFA performed as a mean-structure factor analysis with group mean factor scores at baseline equal to 0 (but free to be estimated at M-87) and fixed loadings and factor-factor correlations over time; minimization technique to calculate factor scores	EFA: 30.5 (3) among Ms; 23.8 (3) among Fs; CFA: 3-factor model with correlated factors separately for Ms and Fs applied for the baseline cross-sectional analysis and as a mean-structure factor analysis	*Validity*: CFA at baseline *Stability over time*: CFA as mean-structure factor analysis on the subgroup with data at both time points (M82–87)	*Validity*: CFA, but no goodness-of-fit assessment or formal comparison with EFA *Stability over time*: CFA: by design, high correlations between corresponding DP scores at both time points (range: 0.88–0.95); between M-82 and M-87, the GREEN DP score mean increased to 0.30 for Ms and to 0.24 for Fs, the TRADITIONAL (men) and the SWEET-TRADITIONAL (women) DPs decreased to −0.27 and −0.18, and the SWEET DP (men) was virtually unchanged
Togo et al., 2003, Denmark (47) MONICA	Separate PCFAs on FFQ and DR data (in octiles): separate analyses by sex; standardization with Kaiser normalization; scree test, interpretability; promax rotation; loading >|0.30| cutoff Separate CFAs on FFQ and DR data: loading ≥|0.30| cutoff on EFA results; polychoric correlation matrix; RMSEA; weighted least-square variable estimates with robust SEs and mean- and variance-adjusted chi-square test statistic	In Ms: 30.5 (3) with FFQ data and 26.2 (3) with DR data; in Fs: 23.8 (2) with FFQ data and 19.8 (2) with DR data	*Relative validity*: Pearson correlation coefficient between scores based on FFQ and DR data *Validity*: CFA; Pearson correlation coefficient between scores based on EFA and CFA	*Relative validity*: EFA on FFQ and DR data: the identified DPs were very similar, although the percentages of explained variance were lower on DR data; Pearson correlation coefficients between FFQ-based and DR-based scores ranged between 0.34 (TRADITIONAL DP among Ms) and 0.61 (both GREEN DPs, among Ms and Fs) CFA on FFQ and DR data: Pearson correlation coefficient between FFQ-based and DR-based scores ranged between 0.37 (TRADITIONAL DP among Ms) and 0.64 (GREEN DP among Fs); higher correlations with CFA than with EFA *Validity* (EFA vs. CFA with the same dietary source): CFA-based DPs were similar across dietary sources and came from models with reassuring model fit (RMSEA <0.10 regardless of the dietary source or sex) FFQ data: Pearson correlation coefficient between EFA-based and CFA-based scores ranged between 0.91 (TRADITIONAL DP among Ms) and 0.96 (SWEET-TRADITIONAL DP among Fs) DR data: Pearson correlation coefficient between EFA-based and CFA-based scores ranged between 0.82 (GREEN DP among Ms) and 0.94 (both GREEN and SWEET-TRADITIONAL DPs among Fs); higher correlations were found when using the same dietary data

(*Continued*)

TABLE 4 (*Continued*)

Reference, location, title of study	DP identification methods	Percent explained variance (no. of factors) or CFA/CA model	Assessment of reproducibility/validity	Main results
Varraso et al., 2012, France and Spain (57) EGEA2-France, Spanish PAC-COPD	PCA and CFA used as equivalent approaches on 1000 randomly selected samples from each of 4 different setups: 1. 100% of EGEA2-France study 2. 50% of EGEA2-France study 3. 25% of EGEA2-France study 4. 100% of Spanish PAC-COPD study PCA: scree-test, interpretability; varimax rotation; distribution of the factor loading of FGs to each DP represented via box-plot and median loading >|0.30| as cutoff CFA: not based on previous EFA; 4 different models tested (3-factor and 2-factor models with correlated latent variables, 3-factor and 2-factor models with independent latent variables); chi-square test, GFI, and RMSEA; distribution of the factor loading of FGs to each DP represented via box-plot and median loading >|0.30| as cutoff	PCA: NA (3); CFA: 3-factor model with no correlation among latent variables (highest GFI and lowest RMSEA)	*Reproducibility and validity*: statistical properties (min, quartile 1, median, quartile 3, max) of the distribution of the factor loading of each FG to each DP in each of the 4 subsamples considered	*Reproducibility and validity*: two consistent DPs were identified by CFA in each of the subsamples, whereas PCA led to less interpretable (smaller median of factor loadings and higher dispersion) DPs, especially for the smallest sample
Weismayer et al., 2006, Sweden (28) SMC	Separate EFAs at baseline and at follow-up for each of the 4 subgroups: scree test, interpretability; varimax rotation; descriptive labeling Separate CFAs at baseline and at follow-up for each of the 4 subgroups: loading ≥|0.20| cutoff on EFA results	EFA: NA (3); CFA: no model selection	*Validity*: CFA *Stability over time*: 1. Spearman correlation coefficient between baseline and follow-up scores for each of the 4 groups and both EFA-based and CFA-based scores 2. *t* test of baseline and follow-up differences in mean intakes for the 18 CFA-based FGs with ≥1 loading >0.2 for any of the 3 DPs in any of the 4 subsamples 3. Spearman correlation coefficient between baseline and follow-up intakes of 18 CFA-based FGs with ≥1 loading >0.2 for any of the 3 DPs in any of the 4 subsamples Internal stability of DPs: test of significant changes in the covariance matrix for each confirmed DP at baseline and follow-up	*Validity*: CFA, but no goodness-of-fit assessment or formal comparison with EFA *Stability over time*: 1. Spearman correlation coefficients between EFA-based DP scores equal to 0.59, 0.57, 0.59, and 0.50 for HEALTHY DP, 0.47, 0.48, 0.51, and 0.39 for WESTERN DP, and 0.54, 0.66, 0.58, and 0.46 for ALCOHOL DP after 4, 5, 6, and 7 y, respectively; Spearman correlation coefficients between CFA-based DPs equal to 0.63, 0.63, 0.62, and 0.54 for HEALTHY DP, 0.60, 0.54, 0.56, and 0.57 for WESTERN DP, and 0.73, 0.76, 0.70, and 0.75 for ALCOHOL DP after 4, 5, 6, and 7 y, respectively 2. *t* Test: no evidence of a difference in the means for 10, 6, 6, and 2 of 25 FGs after 4, 5, 6, and 7 y, respectively, but evidence that 3, 7, 8, and 11 of the 18 FGs underwent significant changes after 4, 5, 6, and 7 y, respectively ($P \leq 0.01$) 3. Spearman correlation coefficients between baseline and follow-up intakes of FGs consistently decreasing in size over time (no correlation after 7 y exceeding the size of the correlations after 4 y) Internal stability of DPs: no significant instability after 4 and 5 y of follow-up; significant instabilities for WESTERN DP after 6 y ($P = 0.01$) and for WESTERN ($P = 0.02$) and ALCOHOL DPs ($P = 0.01$) after 7 y

[1] AGFI, adjusted goodness-of-fit index; CA, cluster analysis; CANHR, Center for Alaska Native Health Research study; CC, congruence coefficient; CFA, confirmatory factor analysis; CFI, comparative fit index; DP, dietary pattern; DR, dietary record; E3N, Mutuelle Generale de l'Education Nationale (EPIC-France); EFA, exploratory factor analysis; EGEA2-France, Epidemiological Study on the Genetics and Environment of Asthma 2—France; EIG, eigenvalue; EPIC-NL, European Prospective Investigation into Cancer and Nutrition—the Netherlands; EPIC-Potsdam, European Prospective Investigation into Cancer and Nutrition—Potsdam; FFQ1/FFQ2/FFQ3, food frequency questionnaire at time 1, 2, or 3; FG, food group; FI, food item; GFI, goodness-of-fit index; MONICA, MONItoring of trends and determinants in CArdiovascular Disease; NA, not available; NFI, normed fit index; NNFI, nonnormed fit index or Tucker–Lewis index; PAC-COPD, Phenotype and Course of Chronic Obstructive Pulmonary Disease study—Spain; PCA, principal component analysis; PCFA, principal component factor analysis; REGARDS, Reasons for Geographic and Racial Differences in Stroke; RMSEA, root mean square error of approximation; RV, relative variation; SMC, Swedish Mammography Cohort; SRMR, standardized root mean square residual.

When CFA was used after a previous EFA, the cutoffs for FG inclusion in the CFA models ranged from |0.20| (10, 24, 27, 28) to |0.60| (52, 53), and the CFA model was estimated on a different (validation) sample in 5 articles (24, 51–53, 55).

Among the 15 CFA-based articles, 4 (49–51, 57) provided a formal model selection procedure, where different numbers of DPs, cutoffs for FGs (and rotation methods), and/or correlation structures between DPs were considered. In addition, in 10 articles, the goodness of fit of the selected CFA model was formally tested according to 1 (47, 48) or more (24, 49–55) indexes, whereas 1 article (57) used descriptive statistics from the bootstrap-based distributions of the factor loadings of each FG to each DP. None of the 4 articles that assessed stability of CFA-based DPs over time (10, 27–29) gave details on model fitting. Finally, some articles provided results on values and statistical significance of standardized factor loadings (50, 52–55), and a few compared EFA- and CFA-based DPs with correlation coefficients between factor scores of similar DPs (47, 49).

Among the 10 articles using goodness-of-fit indexes (24, 47–55), the final CFA model was considered a good model in 8 articles and a slightly inappropriate model in 1 article (52), whereas in another article (50), a cutoff of |0.25| for FG inclusion provided a good model fitting, compared with a CFA with |0.20| cutoff. In general, FG standardized loadings were high and reached statistical significance (50, 52–55), and correlation coefficients between EFA- and CFA-based DP scores were very high (47, 49). In another article (57), CFA outperformed PCA in terms of DP interpretability on a bootstrap-based comparison. Overall, the different statistical criteria pointed to reassuring results: most CFA models confirmed their utility in identifying the minimal constructs characterizing the overall diet in the populations under consideration.

Concerning the quality assessment of the included studies, those of "good" quality consistently identified highly reproducible and/or valid DPs; studies of "poor" quality still tended to identify DPs with a fair-to-good reproducibility and/or validity. However, for some articles (10, 27–29) it was not possible to formally evaluate DP validity, in the absence of CFA goodness-of-fit statistics.

Conclusions

The concept of healthy eating patterns has been adopted by the Dietary Guidelines for Americans over time and there is an emerging body of evidence for the beneficial or detrimental effects of DPs on health. Nevertheless, the key issues of reproducibility and validity of DPs have been assessed by a limited number of articles (mostly based on a priori DPs) and using very different approaches. This review included 38 articles on a posteriori DPs, with ∼15 articles dealing with each research question. To our knowledge, this is the first attempt to collect the overall evidence on these issues and it is therefore valuable, yet it is still limited in its ability to draw strong conclusions.

The identification of DPs with PCA/EFA or CA has traditionally used standard statistical approaches and software. However, since 2011, 7 of our articles have assessed matrix factorability before starting PCA/EFA (30, 32, 34, 37, 40, 41, 50) and 3 articles (43, 44, 56) have proposed some innovation in CA procedures, with sound conclusions. Some novelties have therefore been introduced in the identification of a posteriori DPs over the last decade. However, there are essentially no specific investigations of fundamental questions that researchers should consider when using EFA or CA. For example, this happened for input variable format (e.g., nutrients or FGs, and, in the latter case, number of servings or percentage daily energy intake), transformation (e.g., log-transformation or not), and/or potential adjustment by energy intake (on input data or on DP scores, with the residual method or with other solutions), with only 4 articles (26, 36, 42, 46) included in the current review. Similarly, many other relevant topics were investigated in at most 3 or 4 articles, so evidence is too weak to draw any conclusions on reproducibility of DPs across different statistical solutions.

We found more convincing results from the assessment of reproducibility of DPs over short time periods and of relative validity of DPs. Before reporting the key findings, some general concerns have to be introduced. Firstly, during this review, it has often happened that the Results sections described those DPs that were similar across the available dietary datasets, whereas the Discussion sections were left with a short note on the presence of additional DPs that were not common to all dietary datasets (9, 13, 32–34, 37, 39). Secondly, DP similarities were defined qualitatively, looking at factor loading matrices and percentages of explained variances or at FGs that contributed higher-than-mean intakes for each cluster. Thirdly, when present, the quantification of similarities relied mostly on elementary statistics, with no statistical models assumed. Fourthly, the optimal number of DPs to retain was chosen separately for each dietary dataset. Any assessment of reproducibility or relative validity of DPs is based on these critical points.

An opposite solution to independent sets of DPs (to be later analyzed for reproducibility and validity) is to work on a merged data matrix and force the dietary data to express the same set of DPs across dietary datasets. We recently introduced multistudy factor analysis (58) to allow for the simultaneous identification of common and study-specific DPs across different studies, within a statistical model that includes a formal assessment of the number of shared and study-specific DPs. A similar idea of partial sharing of DPs could be applied in the assessment of DP reproducibility and relative validity, after multiple measures from each subject are taken into account. Use of a statistical model would solve most of the inherent limitations of correlation coefficients, cross-classification, and weighted kappa coefficients.

In the validation studies of FFQs that we analyzed, the assessment of reproducibility of DPs provided systematically better results than the corresponding assessment of relative validity, independently of the statistical approach used. This suggests that multiple administrations of the same dietary tool improve consistency of the corresponding DPs,

compared with having 2 different dietary sources. In the latter case, reference periods, number of collected food items, and the administration process are deeply different. An effort is generally made to create a common set of FGs that fits both the instruments; however, other differences cannot be eliminated and are reflected in the weaker agreement between corresponding DPs.

It is reassuring that results on DP relative validity were similar regardless of whether reproducibility was assessed in the same study design or not (13, 35, 37–40, 47). However, in articles assessing DP relative validity only, the presence of different study designs, dietary assessment tools (24HR or DR), reference period of collection, and timing of administration made the comparison of results even more difficult.

Reproducibility of DPs across multiple administrations of the same FFQ was good and the differences between corresponding factor scores were not systematically biased. However, we detected some variability in factor scores that was reflected in wider-than-expected limits of agreement. A 1-y (median) time interval between FFQ administrations across studies could be at the origin of this extra variability.

CFA should have a wider use in nutritional epidemiology, either for describing dietary habits of a population in a more ideal way or for assessing more general questions on reproducibility of DPs over time (10, 27–29), across populations (24), or dietary sources (47). The current review showed that, when used to identify synthetic dietary profiles from a previous EFA or as a 1-step approach, CFA provided models with good fit and interpretable DPs. Publication bias is likely to be present in this case, especially with those articles that simply confirm a previous EFA. Some caution is therefore needed before concluding on the effective power of CFA. However, we lacked information on model goodness of fit for most of the articles assessing more general research questions through CFA-based DPs (10, 27–29). Researchers should remember that using CFA to assess reproducibility of DPs in time or across studies requires giving details on CFA performance too.

We have speculated on the possibility that some DPs would have been more likely to be reproducible and valid than others across the articles included in the review. Unfortunately, CFA does not allow evaluation of the validity of single DPs. The goodness-of-fit measures represent global model fitting, whereas the significance tests on standardized CFA loadings are not informative, when based on highly selected FGs and a reasonable sample size. Regarding reproducibility and relative validity, there is some evidence that DPs built on a few characteristic FGs were more likely to be reproducible and valid; for example, "sandwich-based" (30) or "alcohol-based" (33) DPs gave better results on reproducibility and relative validity than other DPs presented in the same articles. Similarly, well-characterized traditional DPs (e.g., reference 31) could be more likely to be reproducible and valid, although this was not always true (e.g., reference 34). Western-like or prudent-like DPs were generally based on a higher number of dominating FGs, and those FGs represented different aspects of "Western" (e.g., processed food, red meat, sausages, butter, french fries, eggs, high-fat dairy products) or "prudent" (e.g., fruits, vegetables, fish, poultry, low-fat dairy products, nuts, and seeds) diets. These aspects could explain why these DPs reached only fair-to-moderate levels of agreement. A similar argument was already presented in a previous review of empirically derived DPs (6).

It is crucial to evaluate the quality of the original studies included in a systematic review using standardized and validated quality assessment tools, like that (23) we refer to in the current analysis. However, our topic did not fit well within the typical research question of a possible association between exposure and disease. In addition, any evaluation of reproducibility and/or validity of DPs depends strongly on how well DPs were originally identified in the sample under consideration. Finally, the way the assessment of reproducibility and validity of a posteriori DPs is carried out (e.g., how many criteria were considered and which criteria were used) should deserve additional attention. A standard quality assessment tool cannot capture all these aspects, which are fundamental in a systematic review on reproducibility and validity of a posteriori DPs. Nevertheless, we showed that better-designed studies were more likely to provide highly reproducible and/or valid DPs. This conclusion reflects the general idea that good results are more likely to come from well-designed and carefully implemented studies, based on a sound statistical analysis.

In conclusion, although some caution is worthy, this preliminary attempt to collect evidence on reproducibility and relative and construct validity of a posteriori DPs provides several reasonable conclusions on a topic that has not been fully considered so far. In addition, we provide those new to factor or cluster analyses with a small guide that summarizes evidence on several subjective decisions involved in the DP identification process.

Acknowledgments

The authors' responsibilities were as follows—VE, MF: designed research; VE, MD: collected the relevant articles and selected those to be included in the systematic review; VE, RDV, MD, LP, AS, MF: performed the quality assessment of the original studies included in the systematic review; MD, LP: prepared the first draft of Table 1 and of part of Tables 2–4; VE, RDV, MF: completed and refined Tables 2–4; AS: revised all the tables and checked their consistency with the text; AS: prepared Figure 1; RDV: prepared Figure 2; VE: wrote the manuscript and had primary responsibility for final content; and all authors: read and approved the final manuscript.

References

1. Hu FB. Dietary pattern analysis: a new direction in nutritional epidemiology. Curr Opin Lipidol 2002;13(1):3–9.
2. Liese AD, Krebs-Smith SM, Subar AF, George SM, Harmon BE, Neuhouser ML, Boushey CJ, Schap TE, Reedy J. The Dietary Patterns Methods Project: synthesis of findings across cohorts and relevance to dietary guidance. J Nutr 2015;145(3):393–402.

3. US Department of Health and Human Services and USDA. Dietary guidelines for Americans 2015–2020 [Internet]. 8th ed. [cited 2019 Sept 17]. Available from: https://health.gov/dietaryguidelines/2015/.
4. Moeller SM, Reedy J, Millen AE, Dixon LB, Newby PK, Tucker KL, Krebs-Smith SM, Guenther PM. Dietary patterns: challenges and opportunities in dietary patterns research. An Experimental Biology workshop, April 1, 2006. J Am Diet Assoc 2007;107(7):1233–9.
5. Tucker KL. Dietary patterns, approaches, and multicultural perspective. Appl Physiol Nutr Metab 2010;35(2):211–8.
6. Newby PK, Tucker KL. Empirically derived eating patterns using factor or cluster analysis: a review. Nutr Rev 2004;62(5):177–203.
7. Edefonti V, Randi G, La Vecchia C, Ferraroni M, Decarli A. Dietary patterns and breast cancer: a review with focus on methodological issues. Nutr Rev 2009;67(6):297–314.
8. Northstone K, Smith AD, Newby PK, Emmett PM. Longitudinal comparisons of dietary patterns derived by cluster analysis in 7- to 13-year-old children. Br J Nutr 2013;109(11):2050–8.
9. Hu FB, Rimm E, Smith-Warner SA, Feskanich D, Stampfer MJ, Ascherio A, Sampson L, Willett WC. Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire. Am J Clin Nutr 1999;69(2):243–9.
10. Newby PK, Weismayer C, Akesson A, Tucker KL, Wolk A. Long-term stability of food patterns identified by use of factor analysis among Swedish women. J Nutr 2006;136(3):626–33.
11. Asghari G, Rezazadeh A, Hosseini-Esfahani F, Mehrabi Y, Mirmiran P, Azizi F. Reliability, comparative validity and stability of dietary patterns derived from an FFQ in the Tehran Lipid and Glucose Study. Br J Nutr 2012;108(6):1109–17.
12. Northstone K, Emmett PM. A comparison of methods to assess changes in dietary patterns from pregnancy to 4 years post-partum obtained using principal components analysis. Br J Nutr 2008;99(5):1099–106.
13. Ambrosini GL, O'Sullivan TA, de Klerk NH, Mori TA, Beilin LJ, Oddy WH. Relative validity of adolescent dietary patterns: a comparison of a FFQ and 3 d food record. Br J Nutr 2011;105(4):625–33.
14. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA; PRISMA-P. Preferred Reporting Items for Systematic review and Meta-Analysis Protocols (PRISMA-P) 2015 statement. Syst Rev 2015;4:1.
15. Bamia C, Orfanos P, Ferrari P, Overvad K, Hundborg HH, Tjonneland A, Olsen A, Kesse E, Boutron-Ruault MC, Clavel-Chapelon F, et al. Dietary patterns among older Europeans: the EPIC-Elderly study. Br J Nutr 2005;94(1):100–13.
16. Peng RD. Reproducible research and biostatistics. Biostatistics 2009;10(3):405–8.
17. Broman K, Cetinkaya-Rundel M, Nussbaum A, Paciorek C, Peng R, Turek Dl, Wickham H. Recommendations to funding agencies for supporting reproducible research[Internet]. American Statistical Association; 2017[cited 2019, Sept 12]. Available from: http://www.amstat.org/asa/files/pdfs/pol-reproducibleresearchrecommendations.pdf.
18. Wirfalt AK, Jeffery RW. Using cluster analysis to examine dietary patterns: nutrient intakes, gender, and weight status differ across food pattern clusters. J Am Diet Assoc 1997;97(3):272–9.
19. Edefonti V, Hashibe M, Ambrogi F, Parpinel M, Bravi F, Talamini R, Levi F, Yu G, Morgenstern H, Kelsey K, et al. Nutrient-based dietary patterns and the risk of head and neck cancer: a pooled analysis in the International Head and Neck Cancer Epidemiology consortium. Ann Oncol 2012;23(7):1869–80.
20. Tseng M, Breslow RA, DeVellis RF, Ziegler RG. Dietary patterns and prostate cancer risk in the National Health and Nutrition Examination Survey Epidemiological Follow-up Study cohort. Cancer Epidemiol Biomarkers Prev 2004;13(1):71–7.
21. Edefonti V, Bravi F, Garavello W, La Vecchia C, Parpinel M, Franceschi S, Dal Maso L, Bosetti C, Boffetta P, Ferraroni M, et al. Nutrient-based dietary patterns and laryngeal cancer: evidence from an exploratory factor analysis. Cancer Epidemiol Biomarkers Prev 2010;19(1):18–27.
22. Sun J, Buys NJ, Hills AP. Dietary pattern and its association with the prevalence of obesity, hypertension and other cardiovascular risk factors among Chinese older adults. Int J Environ Res Public Health 2014;11(4):3956–71.
23. National Heart, Lung, and Blood Institute. Quality assessment tool for observational cohort and cross-sectional studies. Bethesda (MD): NIH, Department of Health and Human Services; 2014.
24. Judd SE, Letter AJ, Shikany JM, Roth DL, Newby PK. Dietary patterns derived using exploratory and confirmatory factor analysis are stable and generalizable across race, region, and gender subgroups in the REGARDS study. Front Nutr 2014;1:29.
25. Dekker LH, Boer JM, Stricker MD, Busschers WB, Snijder MB, Nicolaou M, Verschuren WM. Dietary patterns within a population are more reproducible than those of individuals. J Nutr 2013;143(11):1728–35.
26. Balder HF, Virtanen M, Brants HA, Krogh V, Dixon LB, Tan F, Mannisto S, Bellocco R, Pietinen P, Wolk A, et al. Common and country-specific dietary patterns in four European cohort studies. J Nutr 2003;133(12):4246–51.
27. Newby PK, Weismayer C, Akesson A, Tucker KL, Wolk A. Longitudinal changes in food patterns predict changes in weight and body mass index and the effects are greatest in obese women. J Nutr 2006;136(10):2580–7.
28. Weismayer C, Anderson JG, Wolk A. Changes in the stability of dietary patterns in a study of middle-aged Swedish women. J Nutr 2006;136(6):1582–7.
29. Togo P, Osler M, Sorensen TI, Heitmann BL. A longitudinal study of food intake patterns and obesity in adult Danish men and women. Int J Obes Relat Metab Disord 2004;28(4):583–93.
30. Beck KL, Kruger R, Conlon CA, Heath AL, Coad J, Matthys C, Jones B, Stonehouse W. The relative validity and reproducibility of an iron food frequency questionnaire for identifying iron-related dietary patterns in young women. J Acad Nutr Diet 2012;112(8):1177–87.
31. Nanri A, Shimazu T, Ishihara J, Takachi R, Mizoue T, Inoue M, Tsugane S; JPHC FFQ Validation Study Group. Reproducibility and validity of dietary patterns assessed by a food frequency questionnaire used in the 5-year follow-up survey of the Japan Public Health Center-Based Prospective Study. J Epidemiol 2012;22(3):205–15.
32. Liu X, Wang X, Lin S, Song Q, Lao X, Yu IT. Reproducibility and validity of a Food Frequency Questionnaire for assessing dietary consumption via the dietary pattern method in a Chinese rural population. PLoS One 2015;10(7):e0134627.
33. Khani BR, Ye W, Terry P, Wolk A. Reproducibility and validity of major dietary patterns among Swedish women assessed with a food-frequency questionnaire. J Nutr 2004;134(6):1541–5.
34. Hong X, Ye Q, Wang Z, Yang H, Chen X, Zhou H, Wang C, Chu W, Lai Y, Sun L, et al. Reproducibility and validity of dietary patterns identified using factor analysis among Chinese populations. Br J Nutr 2016;116(5):842–52.
35. McNaughton SA, Mishra GD, Bramwell G, Paul AA, Wadsworth ME. Comparability of dietary patterns assessed by multiple dietary assessment methods: results from the 1946 British Birth Cohort. Eur J Clin Nutr 2005;59(3):341–52.
36. Northstone K, Ness AR, Emmett PM, Rogers IS. Adjusting for energy intake in dietary pattern investigations using principal components analysis. Eur J Clin Nutr 2008;62(7):931–8.
37. Loy SL, Jan Mohamed HJ. Relative validity of dietary patterns during pregnancy assessed with a food frequency questionnaire. Int J Food Sci Nutr 2013;64(6):668–73.
38. Okubo H, Murakami K, Sasaki S, Kim MK, Hirota N, Notsu A, Fukui M, Date C. Relative validity of dietary patterns derived from a self-administered diet history questionnaire using factor analysis among Japanese adults. Public Health Nutr 2010;13(7):1080–9.
39. Crozier SR, Inskip HM, Godfrey KM, Robinson SM. Dietary patterns in pregnant women: a comparison of food-frequency questionnaires and 4 d prospective diaries. Br J Nutr 2008;99(4):869–75.

40. Bountziouka V, Tzavelas G, Polychronopoulos E, Constantinidis TC, Panagiotakos DB. Validity of dietary patterns derived in nutrition surveys using a priori and a posteriori multivariate statistical methods. Int J Food Sci Nutr 2011;62(6):617–27.
41. Bountziouka V, Panagiotakos DB. The role of rotation type used to extract dietary patterns through principal component analysis, on their short-term repeatability. J Data Sci 2012;10:19–36.
42. Bailey RL, Gutschall MD, Mitchell DC, Miller CK, Lawrence FR, Smiciklas-Wright H. Comparative strategies for using cluster analysis to assess dietary patterns. J Am Diet Assoc 2006;106(8):1194–200.
43. Lo Siou G, Yasui Y, Csizmadi I, McGregor SE, Robson PJ. Exploring statistical approaches to diminish subjectivity of cluster analysis to derive dietary patterns: the Tomorrow Project. Am J Epidemiol 2011;173(8):956–67.
44. Sauvageot N, Schritz A, Leite S, Alkerwi A, Stranges S, Zannad F, Streel S, Hoge A, Donneau AF, Albert A, et al. Stability-based validation of dietary patterns obtained by cluster analysis. Nutr J 2017;16(1):4.
45. McCann SE, Marshall JR, Brasure JR, Graham S, Freudenheim JL. Analysis of patterns of food intake in nutritional epidemiology: food classification in principal components analysis and the subsequent impact on estimates for endometrial cancer. Public Health Nutr 2001;4(5):989–97.
46. Wirfalt E, Mattisson I, Gullberg B, Berglund G. Food patterns defined by cluster analysis and their utility as dietary exposure variables: a report from the Malmo Diet and Cancer Study. Public Health Nutr 2000;3(2):159–73.
47. Togo P, Heitmann BL, Sorensen TI, Osler M. Consistency of food intake factors by different dietary assessment methods and population groups. Br J Nutr 2003;90(3):667–78.
48. Lau C, Glumer C, Toft U, Tetens I, Carstensen B, Jorgensen T, Borch-Johnsen K. Identification and reproducibility of dietary patterns in a Danish cohort: the Inter99 study. Br J Nutr 2008;99(5):1089–98.
49. Bedard A, Garcia-Aymerich J, Sanchez M, Le Moual N, Clavel-Chapelon F, Boutron-Ruault MC, Maccario J, Varraso R. Confirmatory factor analysis compared with principal component analysis to derive dietary patterns: a longitudinal study in adult women. J Nutr 2015;145(7):1559–68.
50. Castro MA, Baltar VT, Selem SS, Marchioni DM, Fisberg RM. Empirically derived dietary patterns: interpretability and construct validity according to different factor rotation methods. Cad Saude Publica 2015;31(2):298–310.
51. Fransen HP, May AM, Stricker MD, Boer JM, Hennig C, Rosseel Y, Ocke MC, Peeters PH, Beulens JW. A posteriori dietary patterns: how many patterns to retain? J Nutr 2014;144(8):1274–82.
52. Maskarinec G, Novotny R, Tasaki K. Dietary patterns are associated with body mass index in multiethnic women. J Nutr 2000;130(12):3068–72.
53. Park SY, Murphy SP, Wilkens LR, Yamamoto JF, Sharma S, Hankin JH, Henderson BE, Kolonel LN. Dietary patterns using the Food Guide Pyramid groups are associated with sociodemographic and lifestyle factors: the multiethnic cohort study. J Nutr 2005;135(4):843–9.
54. Ryman TK, Boyer BB, Hopkins S, Philip J, O'Brien D, Thummel K, Austin MA. Characterising the reproducibility and reliability of dietary patterns among Yup'ik Alaska Native people. Br J Nutr 2015;113(4):634–43.
55. Schulze MB, Hoffmann K, Kroke A, Boeing H. Risk of hypertension among women in the EPIC-Potsdam Study: comparison of relative risk estimates for exploratory and hypothesis-oriented dietary patterns. Am J Epidemiol 2003;158(4):365–73.
56. Greve B, Pigeot I, Huybrechts I, Pala V, Bornhorst C. A comparison of heuristic and model-based clustering methods for dietary pattern analysis. Public Health Nutr 2016;19(2):255–64.
57. Varraso R, Garcia-Aymerich J, Monier F, Le Moual N, De Batlle J, Miranda G, Pison C, Romieu I, Kauffmann F, Maccario J. Assessment of dietary patterns in nutritional epidemiology: principal component analysis compared with confirmatory factor analysis. Am J Clin Nutr 2012;96(5):1079–92.
58. De Vito R, Lee YCA, Parpinel M, Serraino D, Olshan AF, Zevallos JP, Levi F, Zhang ZF, Morgenstern H, Garavello W, et al. Shared and study-specific dietary patterns and head and neck cancer risk in an international consortium. Epidemiology 2019;30(1):93–102.