AStA Wirtsch Sozialstat Arch (2020) 14:267–304
https://doi.org/10.1007/s11943-020-00275-8

ORIGINALVERÖFFENTLICHUNG

Treatment of sample under-representation and skewed heavy-tailed distributions in survey-based microsimulation: An analysis of redistribution effects in compulsory health care insurance in Switzerland

Tobias Schoch · **André Müller**

Received: 30 December 2019 / Accepted: 5 August 2020 / Published online: 1 September 2020

Abstract The credibility of microsimulation modeling with the research community and policymakers depends on high-quality baseline surveys. Quality problems with the baseline survey tend to impair the quality of microsimulation built on top of the survey data. We address two potential issues that both relate to skewed and heavy-tailed distributions.

First, we find that ultra-high-income households are under-represented in the baseline household survey. Moreover, the sample estimate of average income underestimates the known population average. Although the Deville–Särndal calibration method corrects the under-representation, it cannot achieve alignment of estimated average income in the right tail of the distribution with known population values without distorting the empirical income distribution. To overcome the problem, we introduce a Pareto tail model. With the help of the tail model, we can adjust the sample income distribution in the tail to meet the alignment targets. Our method can be a useful tool for microsimulation modelers working with survey income data.

The second contribution refers to the treatment of an outlier-prone variable that has been added to the survey by record linkage (our empirical example is health care cost). The nature of the baseline survey is not affected by record linkage, that is, the baseline survey still covers only a small part of the population. Hence, the sampling weights are relatively large. An outlying observation together with a high sampling weight can heavily influence or even ruin an estimate of a population characteristic. Thus, we argue that it is beneficial—in terms of mean square error—to use robust

T. Schoch (✉)
School of Business, Institute ICC, University of Applied Sciences Northwestern Switzerland, Riggenbachstrasse 16, 4600 Olten, Switzerland
E-Mail: tobias.schoch@fhnw.ch

A. Müller
Ecoplan AG – Research in Economics and Policy Consultancy, Monbijoustrasse 14, 3011 Bern, Switzerland

estimation and alignment methods, because robust methods are less affected by the presence of outliers.

Keywords Simulation · Pareto distribution · Representative outliers · Nonresponse · Calibration · Imputation

JEL classification C15 · C54 · C63 · C83 · I13 · I18

Methoden zur Behandlung von Unterrepräsentation bei schiefen Verteilungen für die stichprobenbasierte Mikrosimulation: Eine Analyse zu den Umverteilungseffekten in der obligatorischen Krankenversicherung der Schweiz

Zusammenfassung Eine qualitativ hochstehende Stichprobenerhebungen ist eine wesentliche Voraussetzung, um gültige und zuverlässige Aussagen mit stichprobenbasierten Mikrosimulationsstudien zu tätigen. Sind die stichprobenbasierten Schätzer massiv verzerrt (z. B. infolge von Antwortausfällen), so ist damit zu rechnen, dass die auf dem Basisdatensatz aufbauende Mikrosimulation ebenfalls zu verzerrten Umverteilungseffekten führt. Wir befassen uns mit zwei potenziellen Schätz-/bzw. Simulationsproblemen, die bei schiefen Verteilungen mit schweren Rändern (heavy tails) in personenrepräsentativen Haushaltserhebungen auftreten.

(1) Am Beispiel des Haushaltseinkommens zeigen wir auf, dass Haushalte mit einem sehr hohen Einkommen in der Erhebung deutlich unterrepräsentiert sind (hochgerechnet auf die Grundgesamtheit). Des Weiteren unterschätzt die Hochrechnung des arithmetischen Mittels das bekannte Populationsmittel in der Gruppe der Haushalte mit einem Einkommen von über 200.000 CHF deutlich. Der Einsatz der Kalibrierungsmethode nach Deville-Särndal vermag zwar die Unterrepräsentation der Haushalt zu korrigieren, ist jedoch nicht in der Lage, die Hochrechnungsgewichte so anzupassen, dass das geschätzte Mittel mit dem bekannten Populationsmittel übereinstimmt, ohne gleichzeitig die empirische Einkommensverteilung zu verfälschen. Wir schlagen darum eine alternative Methode vor, die eine Pareto-Verteilung für den rechten Rand der Verteilung (tail model) unterstellt. Mit Hilfe des Modells kann die Einkommensverteilung der Stichprobe am rechten Rand derart anpasst werden, dass das gewichtete arithmetische Stichprobenmittel mit dem bekannten Populationsmittel übereinstimmt.

(2) Der zweite Beitrag bezieht sich auf die Behandlung von ausreißeranfälligen Variablen mit schiefer Verteilung. Als Beispiel dienen uns die in einem Jahr aufgelaufenen Gesundheitskosten pro Person (Registerdaten). Diese Variable wurde mittels Datensatzverknüpfung in den Datensatz der Stichprobenerhebung integriert. Bei einer kleinen Gruppe von Personen waren die Gesundheitskosten außerordentlich hoch, so dass sie klar als Ausreisser zu bezeichnen sind. Besitzt ein Ausreisser überdies ein grosses Hochrechnungsgewicht, so sprechen wir von einer einflussreichen Beobachtung, die einen erheblichen Einfluss auf die Schätzer ausüben kann (verzerrte Schätzung und/oder aufgeblähte Varianz des Schätzers). Wir zeigen auf, dass es in solchen Situationen vorteilhaft ist, robuste Schätz- und Anpassungsme-

thoden (alignment/calibration methods) zu verwenden, weil sie durch einflussreiche Beobachtungen weniger stark beeinträchtigt werden.

1 Introduction

Health care policymakers have long been concerned with health care financing arrangements (e.g., per capita premiums, taxes, contributions from social security) and the effect of these arrangements on the receiver of health care. In the light of rising health care expenditures, the effects of financing arrangements on income distribution have recently attracted attention from policymakers. Because various financial arrangements have different implications for an individual's balance between payments made to and health care received from health care insurance, analysis of redistributive effects is worthwhile.

The early literature on redistribution in health care has mainly focused on individual financing arrangements and whether their redistributive effect is progressive or regressive with respect to income (e.g., Doorslaer et al. 1999). A limitation of this approach is how to aggregate the redistributive effects of separate financial arrangements to obtain the overall effect. Simply aggregating separate effects is not sensible because the financial arrangements are interdependent. Another weakness is the reliance on mainly aggregate-level data, which makes an examination of redistribution effects for subpopulations impossible.

To overcome these limitations, researchers have adopted a microsimulation approach; e.g., Grabka (2004) and Drabinski (2004). Microsimulation models are useful in redistribution analysis because they enable the simulation of policy effects on a sample of economic agents (e.g., individual or households) at the individual level. The overall analysis then comprises an evaluation of the consequences induced by a policy or a policy reform on indicators of the activity or welfare for each individual agent in the underlying microdata (Bourguignon and Spadaro 2006; Spielauer 2011).[1] For a recent survey on the application of microsimulation models in health care research, we refer to Schofield et al. (2018).

Microsimulation studies are typically based on survey samples (e.g., households) on top of which the simulation runs. We are interested to study sample averages of the redistribution effects by breakdown variables (gender, age, income group, etc.). Thus, a qualitatively good baseline survey (i.e., absence of outliers and measurement errors) is indispensable to obtain reliable simulations because outliers and other data imperfections tend to bias the estimates.

In the early days of microsimulation, researchers have often been satisfied if their simulation model runs and approximately tracks observed data. Data quality and sound statistical inference have received microsimulation modelers' attention—at least— since the paper of Klevmarken (2002). Much of the research in this area has been devoted to alignment (also known as calibration or benchmarking) methods that attempt to align estimated characteristics (e.g., mean or total) with known population values; see Creedy and Tuckwell (2004) and references therein. These

[1] See Hannappel and Troitzsch (2015) for a recent survey article on microsimulation in German.

alignment methods do not explicitly address survey errors such as systematic under-representation of particular groups of agents or individuals; instead, they correct discrepancies between the baseline survey and known population values by reweighting the survey data. In general, the methods proved successful in improving simulation accuracy for a wide range of applications.

When alignment cannot be achieved with standard methods, Myck and Najsztub (2015) show that calibrating or reweighting the data sequentially over several stages may be beneficial. The authors prove the effectiveness of sequential calibration for a household survey that suffers from under-representation of high-income groups. In our application, we encounter the same problem: High-income households are under-represented in the baseline survey, compared with data from tax registers. Although calibration corrects the under-representation problem, it cannot achieve alignment of estimated average income in the right tail of the distribution with known population values *without* distorting the empirical distribution. Thus, there is a tradeoff: Either average income is aligned but the empirical distribution is severely distorted or vice versa. The problem is rooted in the inability of the (standard) calibration method to cope with skewed heavy-tailed distributions.

The first purpose of this paper is the introduction of a parametric Pareto model to describe the right tail of the income distribution. With the help of the tail model, we adjust the sample distribution such that average income in the top income bracket is aligned with known values from tax data. Our key contribution is a new method based on order statistics from the Pareto model; this contribution is an extension of our earlier model (Schoch et al. 2013; Müller and Schoch 2014a).

The second goal of the paper also refers to the treatment of skewed heavy-tailed, outlier-prone distributions. However, in this case, the baseline survey has fortunately been enriched with individual data on health care costs through record linkage. Thus, modeling cost data is unnecessary because the true cost data are available. Unfortunately, the heavy-tailed population distribution in conjunction with the baseline survey's small sampling fraction make standard estimation procedures very unreliable. We address this problem and propose robust estimating and alignment methods to cope with skewed heavy-tailed distributions. Although the combination of survey data with other sources through record linkage has been investigated, for example, Lohr and Raghunathan (2017) and Thompson (2019), the topic of this study has not been addressed.

To facilitate the methodological discussion, we apply the methods and techniques to our microsimulation model on compulsory health care insurance in Switzerland.[2]

The remainder of the paper is organized as follows. In Sect. 2, we provide background information on compulsory health care insurance in Switzerland. In Sect. 3, we explain the microsimulation model. In Sect. 4, we discuss how the Pareto tail model for income can correct the under-representation of low- and high-income groups and adjust for nonresponse bias. In Sect. 5, we study the problem of outliers that result from record linkage of heavy-tailed population distributions to the baseline survey. Finally, in Sect. 6, we conclude by discussing the major findings.

[2] All computations in this article were made with the R statistical software; see R Core Team (2019).

In Appendix A, we describe the microsimulation model briefly; Appendix B provides an introduction to the Deville–Särndal calibration method.

2 Institutional setting of compulsory health insurance

Basic compulsory health care insurance (CHI) in Switzerland is a package of insurance benefits that must be offered by any insurance provider to *any person without selection*.[3] In particular, all insurance contracts that qualify for CHI must not be subject to health assessments or similar gatekeepers to inhibit enrolment in an insurance plan. Any type of price discrimination or positive risk selection with respect to an individual's age, gender, or health condition is prohibited.[4] CHI is compulsory for all permanent residents in Switzerland. Hence, each individual is obliged to purchase a CHI contract from one of the 56 insurance providers who qualified for CHI in 2016 (BAG 2018). Family members are insured individually. CHI is not sponsored by employers. Individuals are free to choose and change their insurer and/or insurance contract once per year, but they must sign on with an insurer operating in their canton.[5] As a result, the provision of health care is heavily decentralized, and cantons exercise great control over health care (Crivelli et al. 2006).[6]

The benefits of CHI are *identical* for all insured persons throughout the country in the event of illness, accident, and maternity. Although the benefits are identical for all insured persons, CHI offers a set of insurance plans—among which individuals are free to choose—with different financing. The set of plans consist of a heavily regulated basic insurance (franchise ordinaire, CHF 300 deductible) and five special insurance plans that rebate the premium in exchange for greater financial liability (higher degree of cost sharing through higher deductibles when individuals first incur costs; Table 1) or for accepting a limited choice of providers (managed-care arrangements).

CHI premiums are unrelated to earnings but are raised as per capita premiums. To mitigate the regressive effect of the premiums, eligible low-income individuals are entitled to premium reductions or subsidies (individuelle Prämienverbilligung). The subsidies are co-financed by the cantons and the federal government but eligibility criteria, subsidy amount, and payout procedures differ by canton.

[3] With regard to the total healthcare expenditures of CHF 80.5 billion in 2016, the costs covered by CHI were CHF 43.9 billion in 2016 or approximately 55% (BAG 2018). The remaining CHF 36.9 billion is funded through supplementary insurance contracts, which *complement* CHI (e.g., insurance plans that cover alternative medicine, specialized inpatient hospital care plans, dental insurance). Supplementary insurance contracts are subject to a risk assessment and are available only to persons who undergo medical examination.

[4] A risk adjustment scheme among insurance providers (Risikoausgleich) reduces insurance companies' incentives to select positive risks.

[5] A CHI contract entitles the insured to visit any healthcare provider in their canton, and if they prefer, treatment outside their canton is available by paying for the difference between the prices charged in outside hospitals and reimbursements available in their canton.

[6] For a comparative review of the Swiss healthcare system, see OECD (2011).

Table 1 Franchise, premium rebates, and maximum deductible for adults (in CHF; 2016)

Franchise	300	500	1000	1500	2000	2500
Premium rebates[a]	0%	8%	20%	30%	40%	50%
Max. deductibles	700	1200	1700	2200	2700	3200

Source BAG (2017)
[a]Rebates refer to the CHF 300 franchise ordinaire

2.1 Redistributive effects in the system of compulsory health care

Compulsory health care is financed through mixed sources. From the perspective of individuals seeking health care, insurance providers are the main provider of reimbursement for basic health care expenditures. Reimbursements cover a portion of health care costs, and the insured pays the remainder of the incurred cost through cost sharing (the amount depends on the insurance policy) and out-of-pocket payments (OOP).

From a citizen's point of view, individuals contribute to the total health care expenditures in two ways: As health care insurance holders, they finance the system through premium payments (and cost sharing); as taxpayers, they establish the financial basis for health care providers in the cantons (e.g., hospitals) and social insurances (old-age and invalidity, means-tested supplementary benefits, and premium reductions). Fig. 1 shows the major financial flows in the system.

Contributions to and financial aids from the system differ greatly in order of magnitude (see balance sheet representation of CHI in Table 2). More importantly, the various financial sources have different implications for a household's balance between payments made to and financial support received from the system and hence for redistributive effects. Based on theoretical reasoning, we know that the (flat-rate) premium payments exercise a fairly strong regressive effect (i.e., the financial

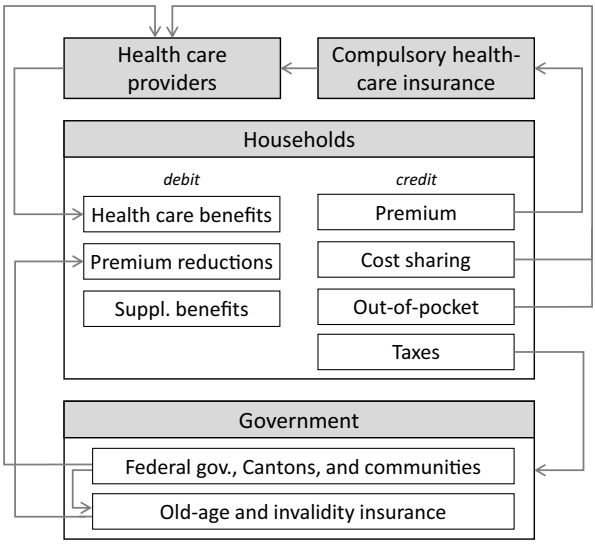

Fig. 1 Major financial flows in compulsory health care (source: Schoch et al. 2013)

Table 2 Balance sheet representation of compulsory health insurance (in CHF; 2016)

Items	Debit ($\times 10^9$)	Credit ($\times 10^9$)
Premium for compulsory health insurance[a]	28.6	
Cost sharing (deductible and co-payments)	4.3	
Out-of-pocket payments[b]	0.2	
Taxes (federal, cantonal, and municipality level)	10.7	
Health care benefits received by insured persons		39.4
Premium reductions paid to low-income people		2.2
Premium reductions through means-tested benefits[c]		2.3
Total	43.9	43.9

Source BAG (2018)
[a]This amount corresponds to the payments made by the health insurers to the system.
[b]Estimated amount; the methodological details are outlined in Schoch et al. (2013).
[c]Supplementary, means-tested benefits can be claimed when old-age or invalidity pensions fall below specified limits.

burden decreases in relative terms with growing income). The regressive effect is somewhat mitigated for low-income adults by premium reductions but tends to affect middle-class families. Taxes, by contrast, exert a progressive effect with respect to income: Households in high-income brackets contribute a disproportionally large share to total health care expenditures. Although theoretical reasoning may provide crude insights into the redistributive effect of a single financial element (e.g., taxes), it cannot demonstrate how the different financing elements interact and what redistributive *net* effect results. Notably, an analysis of aggregate data also cannot accomplish this. The availability of micro-level household (and personal) data is indispensable for studying redistributive effects in detail.

3 Microsimulation model

A household- or person-level dataset that contains data on all relevant financial elements of the CHI system (i.e., taxes, insurance premiums, etc.) is not available. Therefore, we must use simulation-based approaches or data combination techniques (e.g., record linkage) to study redistributive effects.

Our baseline survey dataset is the 2016 edition of the European Statistics on Income and Living conditions (SILC; Swiss Federal Statistical Office), which refers to the permanent resident population living in private households.[7] SILC is designed as a household survey and provides a rich set of sociodemographic and income-related variables.[8] The sampling design of SILC 2016 is a stratified random sample with proportional allocation; stratification is along the seven major regions (BFS 2016). SILC 2016 has a sample size of 17 880 individuals who live in 7761 households. In relative terms, the sample covers a sampling fraction of approximately 0.2% of the

[7] Individuals with permanent residence in collective households (e.g., nursing homes or prisons) are not included in the population definition.

[8] SILC 2016 is organized as a rotational panel survey. We do not make use of this property.

Swiss resident population. As a consequence of proportional sample allocation, the realized sample sizes for small cantons are small (e.g., 20 households in Uri and in 14 Appenzell Innerrhoden). Because of the very small sample sizes, canton-specific investigations require the application of small-area estimation methods. We do not address canton-level estimation; see Schoch et al. (2013) for further details.

3.1 A static microsimulation approach

In this paper, we focus on a static microsimulation model. However, the survey-related methodological issues we address concern dynamic simulation models to the same extent. The main purpose of static analysis is to simulate the distributional incidence of current policies and the impact on individuals and households of policy changes. Static models have no temporal dimension; instead, they focus on distributions and outcomes for a particular point in time (in our case, the year 2016). Moreover—and in contrast to dynamical models—individual and household characteristics and behaviors are considered exogenous in static microsimulation (Li et al. 2014).[9]

From a methodological perspective, the following two techniques are available to enrich the baseline survey data with supplementary individual- or household-level data:

(i) microsimulation,
(ii) record linkage (at the level of individuals).

When studying only the distributional incidence of *current* policies, technique (ii) is preferred because it augments the baseline data with observed data. However, linking data from auxiliary sources to survey data presents methodological difficulties (see Sect. 5). Moreover, in the vast majority of incidence analyses, record linkage is technically infeasible or prohibited by data-protection laws or both. In these cases, or when we want to investigate policy *changes* or counterfactual policy scenarios, microsimulation is the only feasible technique.

Regarding CHI simulation, Fig. 2 shows all variables that must be included into the SILC baseline survey for distributional analysis. In our earlier incidence analysis (Schoch et al. 2013), all listed variables were simulated for each individual or household in the sample (see Appendix A for a model overview). In the current model, the insurance-related variables (e.g., premium; see variables left of arrow "A" in Fig. 2) could be taken from a recently established register on compulsory health care, maintained by the Swiss Federal Office of Public Health.[10] The remaining variables (see arrow "B") are subject to microsimulation at the level of individuals or households.

[9] Bourguignon and Spadaro (2006) distinguish between arithmetic and behavioral models. The latter type of models include a detailed representation of the behavioral response of individuals and households to changes, whereas arithmetic models ignore behavioral responses.

[10] The BAGSAN register is a compilation of individual-level administrative records collected by insurance providers. The record linkage to SILC 2016 was effected by the Swiss Federal Statistical Office. Linkage is based on a unique person-specific identifier (AHV-Nummer = Sozialversicherungsnummer). The match was successful in 99.2% of all cases.

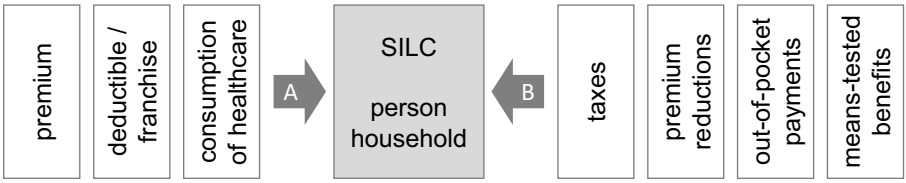

Fig. 2 Simulation or data combination strategies to enrich the SILC 2016 survey with financial data related to compulsory health care insurance

3.2 Finite population inference

Inference in microsimulation models is in principle no different from (ordinary) inferential statistics, but inference aspects have often been neglected. Researchers have often been satisfied if their simulation model runs and approximately tracks observed data (Klevmarken 2002). The insufficient attention to statistical inference is undesirable and unjustified because standard software allows for the routine computation of sampling variances (Figari et al. 2014).

To address statistical inference in microsimulation models, we first note that design-based inference[11] is the relevant mode of inference for survey-based microsimulation because the model builds on top of baseline survey data. Second, we follow Klevmarken (2002) to distinguish between two *modes of simulation*:

(i) simulation based on a set of deterministic rules,
(ii) model-based simulation (stochastic).

Simulation based on set a of deterministic rules is nonstochastic by design; here, stochastic refers to the notion of a super-population model, in the sense of Godambe and Thompson (1986). That is, we assume—in principle—that we can perform simulation by applying deterministic rules to observed variables. For illustration purposes, we consider the following example: Given pre-tax income and relevant socioeconomic variables, tax payments can be computed for each individual in the sample by using a set of rules that describe the taxation regime. From the perspective of sampling theory, the simulated tax payments are regarded as constants. The only stochastic element is induced by the sampling design, which is not affected by simulation. Consequently, we can estimate the total of a simulated variable by the Horvitz–Thompson estimator.[12] Statistical inference then refers to the sampling distribution of the estimated total under the sampling design in use. This approach to inference is certainly useful when the rules underlying simulation are assumed to be deterministic or at least predominantly deterministic.

Inference for model-based simulation is far more intricate because the (super population) model induces an additional stochastic element to the stimulation. This additional randomness accounts for uncertainty that is integral to the statistical

[11] Design-based inference is also known as finite population sampling or randomization inference; see e.g., Särndal et al. (1992, Chap. 2).

[12] Likewise, we can estimate other design-based characteristics, e.g., population mean by the Hajek estimator; cf. Särndal et al. (1992, Chap. 5.7).

model. When the parameters that characterize the simulation model can be estimated from a different dataset, we may attempt to incorporate model uncertainty from the estimation exercise into our simulations.[13]

Regarding our CHI microsimulation model, the two most important variables for CHI financing volume—and subject to microsimulation—are taxes (24.4% share of total finances) and premium reductions (5.0% share). These variables are of a predominantly deterministic nature in the aforementioned sense. Hence, standard sampling inference applies. Regarding means-tested benefits, our earlier model (Schoch et al. 2013) included means-tested benefits as a separate financing instrument. In the current model, only *premium reductions financed through* means-tested supplementary benefits are considered.[14] Their share is 5.3% of total financial flows in CHI. More importantly, the simulated values are of predominantly deterministic nature.

The last financial element subject to simulation is OOP, which cannot be deduced from a set of deterministic rules. Instead, OOP depend on individual behavior, perception of health risks (e.g., self-assessed health condition, prevalence of chronic conditions), household composition, endowment of resources, and limitations because of financial constraints. Therefore, stochastic models or heuristics[15] must be applied for simulation purposes, which implies that inferential statistics cannot relate only to randomization inference. However, because the contribution of OOP to the system is virtually negligible (share of 0.5%), we neglect the model-based contribution to statistical uncertainty. This approach incurs some error; however, the amount of uncertainty not properly accounted for is negligible.

3.3 Unbiasedness of estimates from the baseline survey

So far we have *implicitly* assumed that the baseline survey dataset provides unbiased (or nearly so) estimates of population characteristics. Under a broader perspective, we define the total survey error as the difference between the population characteristic and the sample-based estimate of that characteristic. The total survey error is a measure of quality and can be further subdivided into sampling error and nonsampling error. The sampling error is under the control of the survey statistician. Nonsampling errors are virtually unpredictable and difficult to control. They refer to the entire survey process and comprise the following types or errors: specification errors, measurement errors, sampling frame errors, nonresponse errors, and processing errors; see e.g., Biemer and Lyberg (2003, Chap. 2).

Next, we assume that the specification, measurement, sampling frame, and processing errors are negligible. Therefore, the nonresponse error becomes the focus. We do not claim that all error components other than nonresponse errors are absent;

[13] In our earlier model (Schoch et al. 2013), we modeled the insurance-related variables with data from the Swiss Health Survey. Next, we applied the estimated models to the baseline survey data for predictions and simulations.

[14] The current simulation deviates from earlier versions in terms of scope. It restricts attention exclusively to CHI-related policies and does not incorporate social welfare in a broader sense (e.g., means-tested benefits and allowances from social security).

[15] Viable empirical information on the distribution of OOP is scarce. Therefore, we used primarily heuristic models in the modeling process; see Schoch et al. (2013) for further details.

we only point out that nonresponse dominates total survey error. Regarding our baseline survey, SILC 2016, we can provide verified reasons that substantiate the negligibility assumption.[16]

In the presence of nonresponse, survey estimates tend to be biased. As a direct consequence, all simulation models built on top of the baseline survey data are—as a rule—at risk of generating simulated values whose estimated population characteristics are also biased; cf. Myck and Najsztub (2015).[17] How can we tell that the baseline survey is at risk of producing biased results? Although we cannot answer this question for the entire dataset, we can study individual variables. Of particular importance are variables that serve as inputs for the simulation, for instance, household income. For each variable, we can check whether certain estimated characteristics (e.g., mean or total) are aligned or benchmarked with their known population values. When the characteristics are not properly aligned, we may calibrate the sampling weights such that alignment with the population is achieved using the calibration method of Deville and Särndal (1992); see Klevmarken (2002) or Creedy and Tuckwell (2004) for a discussion of the method in microsimulation.[18]

Two further points are notable. First, statistical inference for a variable of interest is considerably more difficult if that variable has been directly subject to calibration. The original Deville–Särndal method only covers the case, where calibration is conducted with respect to auxiliary variables but not the variable of interest. Second, calibration or reweighting cannot always completely remove the bias, as we explore in Sect. 4.

4 Correcting for nonresponse bias in the baseline survey

In survey research, we distinguish between unit and item nonresponse. Unit nonresponse refers to households (or individuals) who do not participate in the survey because of explicit refusal or unavailability. Item nonresponse occurs when some of the sampled households who agreed to participate in the survey refuse to answer specific questions (see e.g., Groves and Couper 1998, Chap. 1).

When considering income-related nonresponse, strong empirical evidence has been presented that item nonresponse is more accentuated for households in the tails of the income distribution (Biewen 2001). Frick and Grabka (2005) draw the same

[16] SILC is the reference source for comparative statistics on income distribution and social inclusion in the European Union and associated countries. It is more an entire framework than a just a common survey because it is based on a harmonized (among European countries) set of variables and common concepts, guidelines, and procedures (cf. specification, measurement, and processing errors). The sampling frame of Swiss SILC is derived directly from population registers, which are far less prone to coverage errors than establishing the frame on grounds of phone directories; for further details see the SILC quality report, BFS (2017).

[17] For example, suppose that the sample estimate of the pre-tax income distribution is heavily biased. Under that circumstance, a rule-based tax simulation would inevitably produce a faulty distribution of tax payments.

[18] Unless otherwise indicated, we write "Deville–Särndal calibration" for ease of simplicity to mean the calibration under the chi-squared distance measure (possibly imposing some bounds on the weights); see also Appendix B.

conclusion for the German Socio-Economic Panel. They demonstrate that households' propensity to not answering income-related questions is nearly twice as high in the top income decile compared with a median-income household. Consequently, differential or selective item nonresponse can lead to biased estimates.

Although survey teams undertake great efforts to avoid or correct for *unit* nonresponse, it is practically unavoidable. More importantly, when survey compliance is correlated with the variables of interest, there are serious concerns about biases in survey-based inference for these variables, as demonstrated in a theoretical model by Korinek et al. (2006). The authors also provide substantial empirical evidence that unit nonresponse is indeed income dependent. That is, Korinek et al. (2006) find a significantly negative income effect on survey compliance: survey response probability decreases with increasing income. Thus, sample estimates of income characteristics tend to be heavily downward biased. Consequently, we must reckon with biased simulation results because income and other variables possibly affected by nonresponse enter microsimulation as model input.

4.1 Empirical evidence of under-representation in the tails

Unlike surveys, tax registers are not limited by under- or over-representation. Thus, tax register data are a trusted benchmark against which we can compare proportions, means, or totals estimated from surveys, to detect potential nonresponse bias and other survey-related errors.

When comparing estimated shares of households from the 2016 SILC survey against aggregated tax data (ESTV 2017) by income brackets, we find that the estimated shares of households in both tails of the income distribution are noticeably under-represented.[19] Fig. 3 illustrates this finding for the case of married couples; similar patterns of under-representation are found for taxable entities other than married couples (not shown). Also in Fig. 3, the estimated shares of the households in the lower half of the income distribution tend to be slightly over-represented in SILC. The under-representation of the top income bracket is not only a problem observed in the Swiss SILC data, but also in other countries; see e.g. Törmälehto (2017) who presents empirical evidence of under-representation in EU-SILC 2012 for all European and associated countries. Though, the degree of under-representation varies considerably between countries.

4.1.1 Alignment by calibration and reweighting

We attempt to calibrate the weights of the SILC sample data such that the frequency distribution of households by income brackets is aligned with the income distribution resulting from the tax register. We easily achieve this objective when the household shares by income bracket are considered calibration targets (among other totals and proportions); see Schoch et al. (2013) for more details. However, this approach makes the sampling weights dependent on the income variable, which implies that

[19] Empirical evidence of under-representation of wealthy households was found in Müller and Schoch (2014b), who studied asset data in SILC.

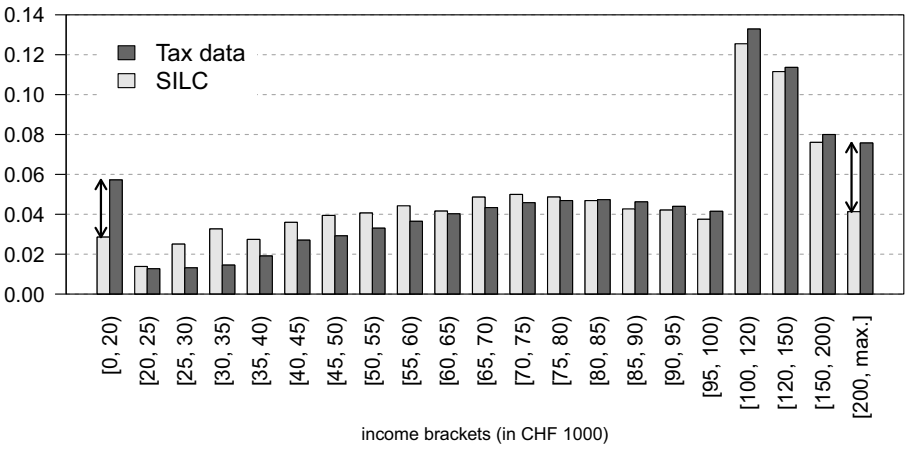

Fig. 3 Share of households by income brackets (in CHF 1000) for tax data and the SILC 2016 survey (source: SILC and ESTV (2017), Normal- und Sonderfälle)

statistical inference becomes technically much more challenging (unless we neglect the dependency introduced through calibration). Myck and Najsztub (2015) propose a different but closely related approach: Calibrate the weights over several stages on variables from administrative records to correct the under-representation of high-income groups. In our application, the indirect calibration method of Myck and Najsztub (2015) was inferior.[20]

Although calibration aligns the estimated household shares by income brackets with known values, we continue to observe an anomaly in the top income bracket: Estimated average income in the top income bracket is only CHF 315 096 (after calibration) and therefore too small by approximately 25% compared with the value reported in the tax register data (i.e., CHF 394 370; ESTV 2017). Consequently, this underestimate of average income implies—based on progressive taxation—downward biased results for the simulation of taxes (and other simulated variables). What can we do to rectify the anomaly? Does it help to run another round of calibration, but with average income as the calibration target?

The calibration method is not appropriate to overcome the underlying problem, which manifested itself because of an underestimate of average income. The underlying problem is that too few high- and ultrahigh-income households were included in the sample for mainly two reasons: (i) these households are rare, and the SILC sampling design did not oversample this special group; and (ii) survey compliance decreases with increasing income (see the aforementioned discussion). These findings are substantiated when we compare the estimated frequencies of ultra-high-income households in SILC with the results in Foellmi and Martínez (2017). Thus, we should—loosely speaking—add some high- and ultra-high-income households to the sample to correct for the deficiency. Calibration and similar reweighting

[20] The results of applying the calibration method of Myck and Najsztub (2015) were inferior in terms of achieving the alignment targets and preventing excessively large weights.

techniques do not succeed because they only modify the "importance" of existing households in the sample. Even worse, these methods can distort the observed income distribution in their attempt to align the estimated sample mean in the top income bracket with the known population value.

Indeed, our numerical analysis (Schoch et al. 2013) shows that calibration tends to increase the weights of observations with high incomes in the top income bracket. Although weight adjustment ensures that the sample average in the top income fulfills the benchmark, it overemphasizes high-income households whose income is still small compared with the households that should have been in the sample in greater number. Since our primary interest is not average income but the entire income distribution (for simulation purposes), any distortion of the distribution is problematic; hence, reweighting methods are not a viable option.

4.2 Pareto tail modeling

The complete tax-data distribution of income is unavailable to us. Therefore, we cannot use it to adjust the SILC income distribution in the right tail. In the absence of empirical data, we thus assume that the right tail of the income distribution can be described by a parametric Pareto distribution. With the help of the tail model, we adjust the sample distribution such that average income in the top income bracket (i.e., above CHF 200 000) is aligned with the known value from the tax data.

Pareto tail models have been a productive assumption in many applications, for example, Dell et al. (2007) show that a Pareto tail model describes top incomes in Switzerland well; see also Foellmi and Martínez (2017). The assumption has also been beneficial in robust statistics; see e.g., Cowell and Victoria-Feser (2007) and Alfons et al. (2013).

To fix notation, we let the income of household i be represented by random variable X_i $(i = 1, ..., n)$, which is defined on the positive real line. Let $\{X_i, i \geq 0\}$ denote a sequence of independent and identically distributed random variables with cumulative distribution function F. Many of the empirically studied parametric income distributions (e.g., Singh–Maddala, Dagum and Generalized Beta) have heavy tails. In particular, their tail decay as a power law $F(X \geq x) \sim L(x) \cdot x^{-(\theta+1)}$ as $x \to \infty$, where $\theta > 0$ is a parameter and $L(y)$ denotes a regularly varying function (Kleiber and Kotz 2003, Chap. 3.3). The tail behavior of such income distributions can be described by a Pareto distribution

$$F_\theta(x) = 1 - \left(\frac{x}{x_0}\right)^{-\theta} \qquad (x \geq x_0), \tag{1}$$

where $x_0 > 0$ is a threshold and $\theta > 0$ is the shape parameter of the Pareto distribution. The corresponding density function is given by $f_\theta(x) = \theta x_0^\theta / x^{\theta+1}$ (for $x > x_0$) and is shown in Fig. 4 for some values of the parameter θ (the threshold x_0 is kept fixed at $x_0 = 1$ for the sake of comparison). We observe that smaller values of θ decrease the density at x_0 and simultaneously imply a heavier tail.

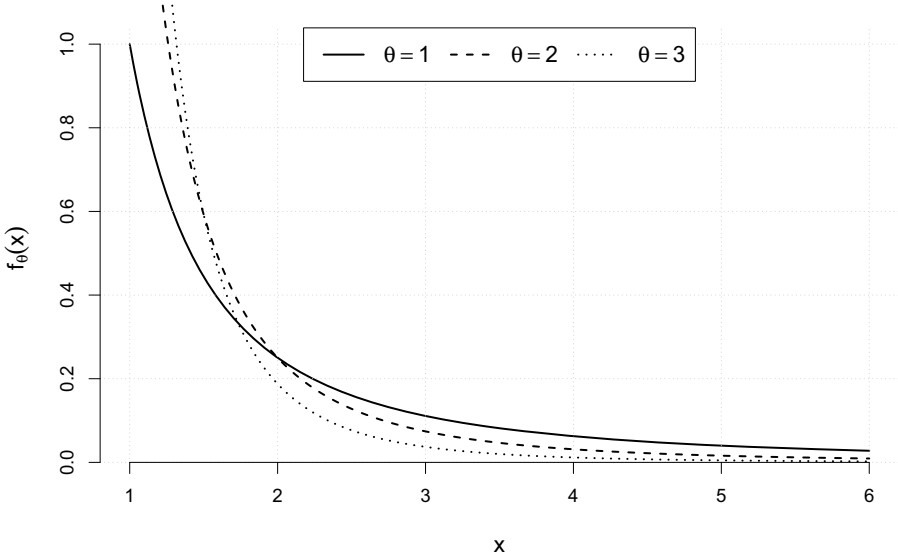

Fig. 4 Pareto density function as a function of variable x for three values of the shape parameter θ (the threshold x_0 is kept fixed at $x_0 = 1$)

4.2.1 Parametrization of the Pareto tail model

To use the Pareto tail model, we must determine or estimate the model parameters from tax data. The threshold x_0 is fixed at CHF 200 000 because this marks the beginning of the top income bracket. To determine the shape parameter θ, we use average income as published by the federal tax authority; see ESTV (2017). Next, we relate the empirical average to the expected value of a Pareto random variable X with law $X \sim F_\theta(x)$. Under this law, the expected value conditional on θ is (Kleiber and Kotz 2003, p. 71)

$$\mathbb{E}_\theta(X) = \frac{\theta x_0}{\theta - 1} \qquad (\text{for } \theta > 1). \tag{2}$$

Putting the empirical average in place of the expected value and substituting CHF 200 000 for x_0, we can solve Equation (2) for θ. Furthermore, since average income in the top income bracket (from tax register data) is known for each canton, we compute canton-specific parameter estimates (the threshold x_0 is the same for all cantons; Table 3). The estimated shape parameters show great variation among the cantons: from 1.42 (canton SZ) to 2.72 (canton JU). For these cantons, the 99% income quantile under the Pareto tail assumption is, respectively, CHF 5.12 million (SZ) and CHF 1.09 million (JU).

4.2.2 Incorporating the Pareto tail assumption into the simulation model

Because the Pareto model is only used for tail modeling, all incomes below the threshold of CHF 200 000 are not affected, and estimation for the lower part of

Table 3 Estimated shape parameters θ by canton

ZH	BE	VD	AG	SG	GE	LU	BL	TI	VS	FR	SO	TG
2.18	2.46	2.42	2.62	2.27	2.34	2.18	2.35	1.98	2.15	2.55	2.34	2.25
(continued)												
NE	BS	GR	SZ	ZG	SH	JU	AR	GL	NW	OW	UR	AI
2.63	1.81	2.27	1.42	1.59	2.50	2.72	2.21	2.53	1.56	1.79	2.56	1.77

Computation based on ESTV (2017)

the distribution refers to the empirical distribution. Regarding the right tail of the distribution, three approaches are worth considering:

(i) imputation of randomly drawn observations from the Pareto model;
(ii) semi-parametric estimation; and
(iii) imputation of expected order statistics from the Pareto model.

In approach i), we replace the observed incomes above the threshold x_0 *with randomly drawn values* from the Pareto tail model F_θ in (1). This approach has been used by, for example, Alfons et al. (2013), in robust statistics; see also Törmälehto (2017) for an application to EU-SILC. Usually, the empirical mean of the imputed observations is not perfectly aligned with the expected value. However, alignment can be achieved by scaling the values slightly. In our earlier model, we used this method with canton-specific parameter values; see Schoch et al. (2013). The major advantage of method (i) is that it generates a corrected income variable that can then be used in the simulation as if it were the original variable. A disadvantage is that the households are assigned randomly drawn income values that may not be related to their originally observed income. Thus, a relatively poor household can be turned into a high-income household (and vice versa). This is normally not an issue, unless the simulated results are to be studied for fine-grained subpopulations.

The second approach is a semi-parametric estimating method and is inspired by Cowell and Victoria-Feser (2007). This approach directly sets in at the stage of estimation and—so to speak—skips the imputation stage. Denote by $F(x)$ the entire income distribution, which is defined as a mixture distribution,

$$F(x) = \begin{cases} F_n(x) & \text{if } x < x_0, \\ F_n(x_0) + \{1 - F_n(x_0)\} \cdot F_\theta(x) & \text{if } x \geq x_0, \end{cases} \tag{3}$$

with the empirical distribution function $F_n(x) = \sum_{i \in s} w_i \mathbb{1}\{x_i \leq x\} / \sum_{i \in s} w_i$, where summation is over all elements in the sample s, w_i is the sampling weight, and $\mathbb{1}\{\cdot\}$ denotes the indicator function. Any characteristic of interest (e.g., arithmetic mean) that can be expressed as a statistical functional $T : G \to \mathbb{R}_+$ of a distribution function G can be computed at the distribution defined in (3). For instance, the (weighted) sample mean—computed at an arbitrary distribution G—can

be expressed as a statistical functional $T(G) = \int x \mathrm{d}G(x)$ where integration is over the positive real line. When T is computed at F defined in (3), we obtain

$$T(F) = \frac{1}{\sum_{i \in s} w_i} \left(\sum_{i \in s} w_i x_i \mathbb{1}\{x_i \leq x_0\} + \frac{\theta x_0}{\theta - 1} \sum_{i \in s} w_i \mathbb{1}\{x_i > x_0\} \right), \tag{4}$$

which highlights that $T(F)$ is a weighted average of the empirical mean for incomes below threshold x_0 and the expected value in the right tail under the Pareto model. We observe that this method does not explicitly replace or impute incomes in the baseline dataset.[21]

The third method is new, according to our review of the literature. For ease of discussion, we neglect the canton-specific tail models and work with a nation-wide model only. Let $X_{1:n}, ..., X_{n:n}$ denote the n order statistics (i.e., observations sorted in ascending order) of the observed income variable in the right tail (i.e., for $x > x_0$). Under the Pareto model in (1), the expected value of the k-th order statistic is (David and Nagaraja 2003)

$$\mathbb{E}_\theta(X_{k:n}) = \frac{x_0 n!}{(n-k)!} \cdot \frac{\Gamma(n - k + 1 - \theta^{-1})}{\Gamma(n + 1 - \theta^{-1})} =: \mu_{k:n} \qquad (\text{for } 1 \leq k \leq n), \tag{5}$$

where Γ denotes the Gamma function.[22] For the imputation approach, we replace all empirical income order statistics $X_{1:n}, ..., X_{n:n}$ in the baseline data by the expected values $\mu_{1:n}, ..., \mu_{n:n}$ (under the Pareto tail model). This method has several advantages over the other two approaches. First, the arithmetic mean of the imputed $\mu_{i:n}$'s ($i = 1, ..., n$) is equal to the (overall) expected value under the Pareto model defined in (1), that is, the mean of the imputed observations is automatically aligned with the benchmark from tax data.[23] Second, the imputation strategy preserves the households' income ranks. A relatively poor household is not turned into a high-income household (and vice versa). Lastly, the changes in income generated by imputation are small; to observe this, we computed the percentage change in income between the empirical and the imputed value for all 271 households in the top income bracket (Fig. 5). Observe that the changes are displayed by income quantiles. For incomes below the third quartile, the changes are less than 21.5 percentage points. The largest change in income for an individual household is an increase of 239.4%, which reflects the fact that households with especially high incomes were under-represented in the original data.

[21] The weighted average in (4) can be easily computed. When the functional of interest is more complicated, notably, if it depends on a known function $h : \mathbb{R} \to \mathbb{R}$ (e.g., average tax payments depend on income and other variables), we have $T_h(F) = \int h(x) \mathrm{d}F(x)$ and the functional may not have a closed-form solution. In this case, we may use numerical integration methods to evaluate T_h.

[22] When n is large, we can approximate the factorial and the (log) Gamma function using Stirling's approximation (or a more refined approximation).

[23] The alignment property only holds for the arithmetic mean but not necessarily for the weighted mean. However, this is not problematic and can be addressed by scaling the imputed values slightly such that the weighted mean is aligned with the benchmark. In our case, the scaling factor is 1.0023.

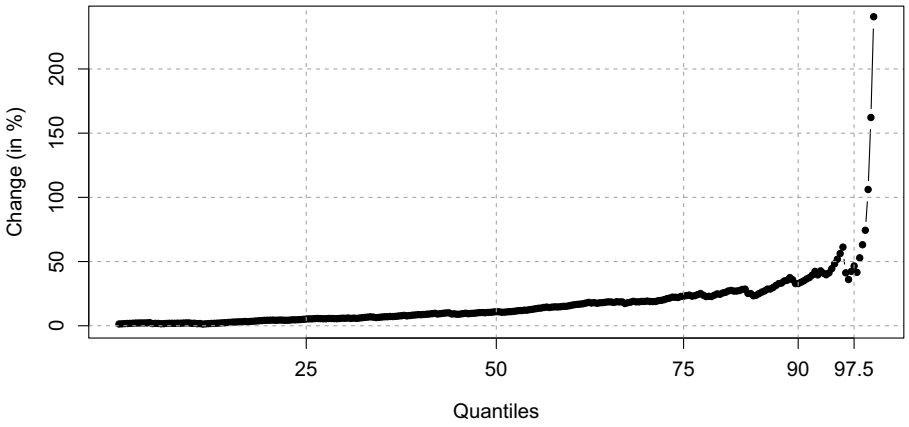

Fig. 5 Percentage change between observed income and imputed income (against household income quantiles; application of method iii)

4.3 Empirical illustration

As an illustration of the methods, we simulated average tax payments in favor of the system of CHI. Tax payments include federal, cantonal, and municipality taxes and are simulated from pre-tax income (and other variables). In Fig. 6, we show average tax payments in favor of CHI for households in different income brackets, once with and once without Pareto tail correction. Tax payments in the top income bracket are substantially underestimated when the correction is not considered. The correction method used in the display of Fig. 6 is based on method iii), that is, imputation by expected order statistics from the Pareto model. However, the two other methods yield similar results (not shown) because the display uses a rather

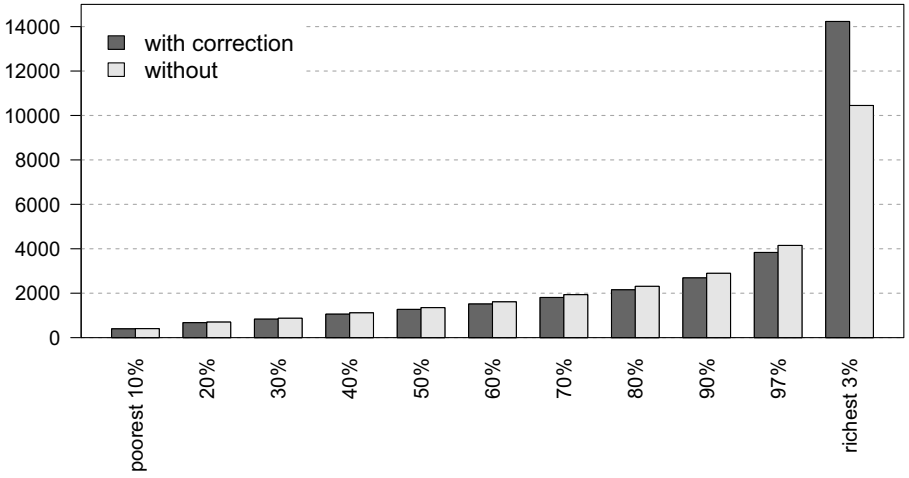

Fig. 6 Average tax payments in favor of the health care system by 11 types of households from different income brackets; tax payments are computed with and without Pareto correction

coarse income bracketing. When we specify smaller income brackets (e.g., in 0.5% steps), the imputations of method (i) show a nonsmooth behavior in the right tail, which is undesirable.

5 Register data and record linkage (with heavily skewed data)

Individual health care cost data are typically not available from household surveys because interviewees do not know the amount of costs they incurred in a calendar year. This was the case in our earlier simulation model (Schoch et al. 2013); therefore, we simulated individual health care costs.[24] The major difficulty in this modeling exercise was the replication of the outlier-prone, heavily right-skewed and zero-inflated distribution of the cost data. Zero-inflation occurs because the majority of individuals did not use any health care-related services; hence, no costs were incurred. By contrast, medical treatment for a few people incurred tremendous costs (outliers).

As we pointed out in Sect. 3, insurance-related data (i.e., premium, franchise, and health care costs) are now available from a register on compulsory health care. Moreover, the Swiss Federal Statistical Office linked the register data to the 2016 SILC survey through record linkage. Thus, we can avoid modeling the cost data because the true cost data are available. It cannot get any better than this, right?

Unfortunately, linking register data to an existing survey dataset is insufficient to guarantee good results. The nature of the baseline survey is not affected by record linkage, that is, the baseline survey still covers only a small, randomly selected part of the underlying population ($\approx 0.2\%$ sampling fraction). A sampling fraction of 0.2% implies—under the simplifying assumption of simple random sampling without replacement—that on average each person receives a sampling weight of approximately 476.[25] Thus, each sampled person is said to represent approximately 476 individuals in the population.

Moreover, we may think of linking health care costs to the survey as if we had directly sampled from the very skewed cost population distribution. As a result, we obtain a sample that shows high sampling variability. Even more problematic is the analysis of such data for breakdowns or domains of interest (e.g., breakdown by gender or age group) because outlying values tend to be more influential in smaller samples. For instance, if a person in a subpopulation has incurred a huge amount of health care costs (e.g., several hundred thousands of CHF), that individual's value represents (under our simplified calculations) the values of 476 individuals in the population and therefore exerts a tremendous influence on the subpopulation's distribution of health care costs. The compound effect of an outlying observation and a large weight can completely ruin an estimate. Thus, we clearly cannot let such extreme data or outliers be untreated.

[24] We modeled individual health care costs with the help of explanatory variables such as number of doctor visits, length of stay in hospital, etc. (conditional on, e.g., age, gender).

[25] In the calculation, we neglected stratification and assumed equal weighting (and absence of nonresponse adjustment).

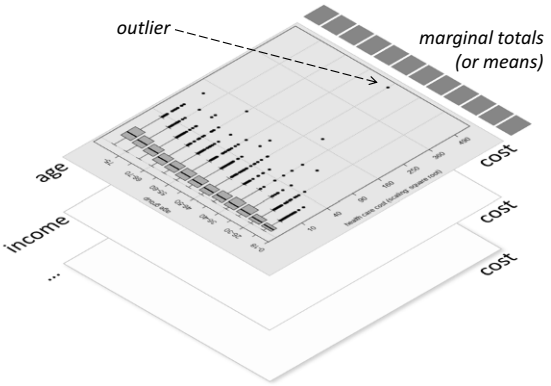

Fig. 7 Slicing the baseline survey data: The top plane or slice shows the relation between the variables (health care) *cost* and *age* (group). The population means (and totals) of health care costs are known for the marginal distribution by age group but not for all other breakdown variables (like income, etc.)

For the time being, an outlier (or extreme value) shall mean an atypical and/or influential observation in the sample (a more formal outlier definition will be given later). Also, for ease of discussion, we consider the specific situation shown in Fig. 7. The figure shows a schematic representation of the (health care) cost data in the baseline survey, cut into slices by breakdown variables (age, income, etc.). The top plane shows the slice `cost` $\sim$ `age` (group). This slice is special for two reasons. First, variable age (group) is used in the microsimulation as a breakdown variable to study the redistributive effects by age. Second, the population means of health care costs by age groups are known (from administrative CHI data). Hence, no estimation is required for the analysis of health care cost by age group (unless we are interested in a characteristic other than the arithmetic mean), yet this setting enables us to adjust the cost data (or, equivalently, the sampling weights) such that the weighted sample means (by age group) are aligned with the known population values. The adjusted data then allow us to obtain (presumably) more accurate estimates of average health care costs for *other* breakdown variables, whose population cost averages are not known (e.g., income, see Fig. 7), compared to not having adjusted the data in the first place (i.e., not having utilized the auxiliary information of the top slice). Such methods refer to the calibration principle of Deville and Särndal (1992, p. 376) that "weights that perform well for the auxiliary variables also should perform well for the study variable" (under the assumption that the study variable is correlated with the auxiliary variables).

It is crucial that the alignment methods applied at the top slice (to stick with our visual metaphor in Fig. 7) work properly, for otherwise alignment issues transmit to other slices causing distorted estimates there. Thus, we must avoid that an alignment problem in one place turns into an estimation problem at another place. For that matter it is of crucial importance *how* an alignment method achieves its goal *in the presence* of outliers. The naive alignment approach which scales the cost data (or weights) by $\bar{y}/\widehat{\bar{y}}$, where $\bar{y}$ and $\widehat{\bar{y}}$ denote, respectively, the population mean and the weighted sample mean of health care cost (for some age group), ensures that alignment is achieved for this breakdown variable. However, outliers in the data may exercise a huge impact on $\widehat{\bar{y}}$ and thus on the scaling factor. To see this, consider the age group of 41–45 years old women. For this group, the population

mean of cost is $\bar{y} = 3102$ CHF. The sample mean amounts to $\widehat{\bar{y}} = 5421$ CHF, which is an overestimate by approximately 75% because of (most notably) one heavy outlier; see Fig. 7. The resulting scaling factor is approx. 0.57, which implies that all observations—even the "good" ones—(or their sampling weights) are heavily shrunken. Such heavy shrinkage may cause disastrous underestimation for other breakdown variables. This problematic behavior is not limited to the naive method. Even more sophisticated alignment methods are not immune to this. In fact, any alignment method which is based on non-robustly estimated characteristics will be influenced by outliers. This is also the case for the (traditional) Deville–Särndal calibration method (Duchesne 1999).

Before we address alignment and estimation methods that can cope with outliers, it is helpful to formalize our definition of outliers.

5.1 Representative and nonrepresentative outliers

Compared with "classical" statistics, outliers are a different concept in design-based survey sampling. In the sampling context, outliers are extreme values selected from the population under study that deviate from the bulk of data. Following Chambers (1986), distinguishing *representative* from *nonrepresentative* outliers is helpful. Representative outliers are extreme but correct values and are thought to represent other population units similar in value. A nonrepresentative outlier is an atypical or extreme observation whose value is either deemed erroneous or unique in the sense that there is no other unit like it.[26]

Furthermore, and in contrast to classical statistics, we also must consider the sampling weights because design-based estimators are functions of the weights and the observed values. Depending on the type of estimator, observations *not* considered outliers (e.g., situated in the bulk of the data) can still heavily influence the estimate because of their large sampling weight. We call such observations *influential values* (Lee 1995). Conversely, outliers well separated from the majority of observations are not necessarily influential when they have small weights. The problem worsens if large values have large sampling weights.

Outliers and influential values are typically dealt with in two separate steps: detection followed by treatment. Another option is the application of robust estimation techniques (e.g., M-estimators; see below), which combine the steps of detection and treatment. All techniques aim to avoid untreated outliers and influential values because these can heavily compromise the variance–bias profile—or equivalently the mean square error (MSE)—of the estimator of interest. So, leaving erroneous outliers untreated implies biased estimates and inflated variance of the estimator. In case of representative and nonrepresentative outliers, the situation is more complicated because the outliers' influence on the MSE depends on the sample size (Hulliger 1995; Lee 1995). If the sampling fraction is large or the sample size is

[26] For the sake of illustration, consider that person in the population who incurred the largest amount of costs. If this particular person gets selected into the sample, it receives a weight of approximately 476 (under our simplified calculations). Now, it is evident that this person is a non-representative outlier because it does not represent approximately 476 individuals with the same amount of costs.

large, the problem is less troublesome. When the sample size is small, however, and whence—as a rule—the variance is the dominant factor in the MSE, small biases introduced through robustification (e.g., reducing values or shrinking weights) can be worthwhile if the variance can be significantly reduced. Thus, for small samples, there is a tradeoff between variance and bias. However, in some cases, the introduced bias can be substantial and may render a robust procedure grossly inefficient. This phenomenon occurs all the more as the sample size increases because the variance decreases, but the bias typically does not. As a result, the bias tends to dominate the MSE for large samples.[27]

5.2 Robust estimation and alignment methods

In contrast to our discussion on Pareto tail modeling for income, we have no comparable parametric model for health care cost because empirical and theoretical evidence on the distributional shape of health care costs is scarce (compared with the well-studied Pareto assumption in income research). Thus, we adopt robust nonparametric methods.

In what follows, y_i ($i \in s$) denotes the variable of interest (health care cost). The goal is to obtain y-totals (or means) as *weighted linear statistics* of the sample data, which are

(i) outlier resistant (robust),
(ii) and (if applicable) aligned with known population totals (or means).

We deliberately speak of weighted linear statistics, not estimators in order to cover estimation and alignment methods. That is, when a statistic is used to estimate an unknown population parameter or characteristic, it is called an estimator. Unlike estimation methods, the population parameter or characteristic of interest is a known quantity in the application of alignment methods. Therefore, the device to achieve alignment is not called an estimator. We use the term *aligned value* to denote the sample-based weighted linear statistic (e.g., weighted mean), which is based on the modified observations or weights to achieve alignment. Clearly, the aligned value is equal to the known population quantity (if alignment was successful).

For estimation methods, we demand only that requirement (i) is met (i.e., outlier robustness, see enumeration above), whereas for alignment methods, both requirements (i) and (ii) must be fulfilled. By way of illustration, consider the visual metaphor of the data slices in Fig. 7. Since the population means are known for the top slice (i.e., `cost` $\sim$ `age`), estimation is pointless and we focus only on outlier resistant alignment. For all other slices, the goal is robust estimation of the y-total

[27] Such troublesome situations have been of great concern to advocates of robust methods. For instance, Chambers (1986) proposes a robust estimator in which the incurred bias is estimated by a robust technique and then "added back" to some extent to the robust estimator; the resulting estimator is called bias-corrected. Hulliger (1995) suggests another solution to the problem, insofar as he considers a set of eligible estimators that include the nonrobust, but consistent, estimator (e.g., Horvitz–Thompson estimator). His method is called minimum estimated risk estimator and is an adaptive procedure because it searches for an optimal variance–bias configuration among the set of eligible estimators.

or -mean (taking the modified sampling weights or observations into account that have been obtained at the top slice).

To fix notation, let w_i denote the sampling weight ($i \in s$). We denote by w_i^* outlier resistant weights (that are possibly adjusted to meet alignment goals), and which are defined as $w_i^* = u_i w_i$, where the u_i's are factors to downweight outliers and achieve alignment. We will discuss the choice of the u_i's later. By the identity

$$\sum_{i \in s} w_i^* y_i = \sum_{i \in s} w_i y_i^* \tag{6}$$

we see that the estimated y-total can equivalently be represented with the help of modified observations $y_i^* = y_i u_i$.[28] Also, we may regard the y_i^*'s as imputed values which are free from outliers and ensure (together with the w_i's) that the alignment goals are achieved (granted that alignment goals were imposed). More importantly, we have the freedom to work, in the later course of the simulation, with the tuples (w_i^*, y_i), (w_i, y_i^*), or directly with the u_i's.

Next, we address three methods to compute the u_i's (and thus the y_i^*'s or w_i^*'s).

5.2.1 Robust estimation

For the further course of discussion, it is helpful to focus on robust M-estimators in the context of finite population estimation (although these estimators do not seek alignment with known population values). We restrict attention to the robust Horvitz–Thompson (HT) estimator of Hulliger (1995) because it is outlier resistant and it can be written as a weighted linear estimator.

Let ψ denote the Huber ψ-function defined as $\psi(x,k) = \min\{k, \max(-k, x)\}$ for $x \in \mathbb{R}$, where $k > 0$ is a robustness tuning constant; we let $\widehat{\sigma}$ be a preliminary robust estimate of scale, for example, the interquartile range of the cost data y_i. The robust estimator of the weighted mean is the solution $\widehat{\mu}_k$ of the estimating equation (Hulliger 1995)

$$\sum_{i \in s} w_i \psi\left(\frac{y_i - \mu}{\widehat{\sigma}}, k\right) = 0. \tag{7}$$

The tuning constant k determines the amount of robustness we want to achieve.[29] Estimator $\widehat{\mu}_k$ can be expressed as a weighted estimator,

$$\widehat{\mu}_k = \frac{\sum_{i \in s} w_i u_i y_i}{\sum_{i \in s} w_i u_i} \quad \text{where} \quad u_i = \frac{\psi(e_i, k)}{e_i} \quad \text{with} \quad e_i = \frac{y_i - \widehat{\mu}_k}{\widehat{\sigma}}, \tag{8}$$

and can thus be brought into the form of (6). The u_i take values in the interval [0,1].

[28] Likewise, we have a Hajek type estimator for the mean, $\sum_{i \in s} w_i^* y_i / \sum_{i \in s} w_i^*$.

[29] A small value of k reduces the influence of outliers and influential observations. By contrast, if $k \to \infty$, $\widehat{\mu}_k$ is equal to the Hajek mean. Computations are performed with the R package `robsurvey` of Hulliger et al. (2019).

5.2.2 Robust adaptive M-estimator with an alignment penalty

In this paragraph, it is assumed that the population y-mean, $\bar{y} = \sum_{i \in U} y_i / N$, is a known quantity ($U$ denotes the set of population indices). Observe that the estimator $\widehat{\mu}_k$, which is defined as the solution to the estimating equation in (7), does not impose alignment goals. As a result, $\widehat{\mu}_k$ may differ considerably from $\bar{y}$. In order to incorporate the auxiliary information that $\bar{y}$ is known, we propose to compute an adaptive M-estimator that minimizes an approximate estimate of the mean squared error of $\widehat{\mu}_k$,

$$\widehat{\mathrm{MSE}}(\widehat{\mu}_k) = \widehat{\mathrm{var}}(\widehat{\mu}_k) + (\widehat{\mu}_k - \bar{y})^2, \tag{9}$$

where $\widehat{\mathrm{var}}$ denotes the estimated variance. Observe that the squared bias term on the r.h.s. of (9) is evaluated with respect to the known population mean $\bar{y}$. The squared bias works like an alignment penalty that penalizes estimates that deviate too much from $\bar{y}$. Formally, we seek the M-estimator which minimizes (9) on the set of tuning constants $\{k : k \in \mathbb{R}_+\}$. The optimal estimator is $\widehat{\mu}_{k_{\mathrm{opt}}}$, where

$$k_{\mathrm{opt}} = \underset{k \in \mathbb{R}_+}{\arg\min}\ \widehat{\mathrm{MSE}}(\widehat{\mu}_k). \tag{10}$$

The proposed estimator is inspired by the minimum estimated risk estimator in Hulliger (1995); our method differs from Hulliger's insofar that he defines the squared bias as $(\widehat{\mu}_k - \widehat{\mu})^2$, where $\widehat{\mu}$ is the weighted sample mean. For ease of reference, we call the estimator $\widehat{\mu}_{k_{opt}}$ with k_{opt} defined in (10) the *minimum risk M-estimator* (MRM). Although the MRM estimator is not explicitly aligned or benchmarked with $\bar{y}$, it often coincides with $\bar{y}$ (or is at least close to the benchmark); see empirical illustration, below. Furthermore, deviations of $\widehat{\mu}_{k_{opt}}$ from $\bar{y}$ are unproblematic (or even intended) provided that the MSE of $\widehat{\mu}_{k_{opt}}$ is considerably smaller than the MSE of any competing estimator. That is, we deliberately relax the alignment requirement slightly whilst the gains in MSE outweigh the incurred bias.

In the presence of outliers and influential values, the MRM estimator tends to be superior in terms of MSE compared with competing methods (see below). However, it can be heavily biased when the population mean $\bar{y}$ is much larger than $\widehat{\mu}_{k_{opt}}$; whence, the squared bias dominates the MSE and the MRM estimator does not achieve any gains in MSE over the weighted sample mean (yet, the MRM estimator is never inferior to the weighted sample mean).

5.2.3 Robust self-calibration

In this paragraph, we introduce a robust calibration method that *explicitly* ensures alignment (under the assumption that the (sub-) population quantities $\bar{y}$ and N are known).[30] To this end, we follow Duchesne (1999), who proposed a robustification of the (traditional) calibration method of Deville and Särndal (1992). In practice, the

[30] If the (sub-) population size N is unknown, it can be replaced by the estimate $\widehat{N} = \sum_{i \in s} w_i$.

traditional calibration (see Appendix B) is used to re-weight a vector of auxiliary variables, say, $\boldsymbol{x}_i \in \mathbb{R}^p$ $(i \in s)$—not the variable of interest, y_i—such that the sample x-totals are aligned with their population values. Our approach, however, seeks calibration or alignment directly for the study variable y_i. Therefore, we call the method (robust) *self-calibration*.

We follow Duchesne (1999) and fix a set of tuples of constants $\{(q_i, r_i) : i \in s\}$. The choice of the constants will be discussed later.[31] Next, we define—still following Duchesne (1999)—a set of weights $\{v_i : i \in s\}$ and consider minimizing the distance function $\sum_{i \in s}(v_i - r_i)^2/q_i$ subject to alignment or calibration constraints (s.t.c.). This choice of distance function is problematic because the resulting weights v_i can be negative. In order to restrict the calibrated weights v_i to the interval $[L, U]$, where L and U are pre-determined boundaries $(0 \leq L < U < \infty)$, we consider instead the following minimization problem

$$\min \frac{1}{2}\sum_{i \in s} h(v_i, q_i, r_i) \qquad \text{s.t.c.} \qquad \begin{bmatrix} \sum_{i \in s} v_i \\ \sum_{i \in s} v_i y_i \end{bmatrix} = \begin{bmatrix} N \\ \sum_{i \in U} y_i \end{bmatrix}, \tag{11}$$

where minimization is with respect to the v_i's, and

$$h(v_i, q_i, r_i) = \begin{cases} \dfrac{(v_i - r_i)^2}{q_i} & \text{if } v_i \in [L, U], \\ \infty & \text{otherwise.} \end{cases} \tag{12}$$

The distance function in (12) is due to Duchesne (1999), and it is a slight modification of Case 7 in Deville and Särndal (1992). We impose two calibration constraints; see r.h.s. of (11). Observe that our second constraint is specified with respect to the study variable, y_i, not an auxiliary variable (this marks the major difference to the proposal of Duchesne 1999). Together, the two constraints ensure that the Hajek estimator of the y-mean, $\sum_{i \in s} v_i y_i / \sum_{i \in s} v_i$, is aligned with the population y-mean.[32]

The choice of the constants (q_i, r_i) is of great importance in order to achieve robustness. We take $(q_i, r_i) = (u_i w_i, u_i w_i)$ for all $i \in s$, where $u_i = \psi(e_i, k_{opt})/e_i$ with $e_i = (y_i - \widehat{\mu}_{k_{opt}})/\widehat{\sigma}$ and $\widehat{\mu}_{k_{opt}}$ is the M-estimator with k_{opt} defined in (10). Observe that this choice implies that $q_i = r_i$ $(i \in s)$, which is sensible and easy to compute but may not be the best specification possible. That is to say, it can sometimes be advantageous to take the constants to be $(w_i u_i, w_i u_i')$, where $u_i' = \psi(e_i, k')/e_i$ with k' other than k_{opt}. However, this approach poses the difficulty of choosing the tuning constant k'. We stick with the choice $q_i = r_i$

[31] The constants define the class of QR estimators in the sense of Wright (1983), and QR estimators can be regarded as calibration estimators (Duchesne 1999).

[32] For reasons of efficiency (see e.g., Särndal et al. 1992, 182), we prefer the Hajek mean over the Horvitz–Thompson estimator of the mean.

because of its simplicity, and then we solve (11) to get the calibrated weights v_i ($i \in s$).

In the later course of the simulation, we are free to work with the tuples (v_i, y_i) or (w_i, y_i^*), where $y_i^* = y_i u_i^*$ with $u_i^* = v_i / w_i$, or we may store the u_i^*'s in the baseline survey for future usage ($i \in s$).

5.3 Empirical illustration

We study the empirical performance of the three methods for estimation and alignment of health care cost by age group (cf. top slice in Fig. 7). Clearly, estimation is actually not needed because the population means (by age group) are known quantities. Therefore, we are mainly concerned whether the methods achieve alignment. Fig. 8 shows the aligned values or estimates of average cost by age group for several methods. The known population means are shown as a thick grey line. From the visual display, we observe that the weighted sample mean overestimates the population mean for the age group of 41–45 years old women by approx. 75% (i.e., CHF 5421 vs. 3102) because of a few outliers. A similar behavior—albeit less pronounced—is apparent for the age groups 26–30, 46–50 and 51–55 years. The estimates of the minimum risk M-estimator (MRM) are robust against outliers and influential values, and the estimates coincide with (or are at least close to) the population means in the age groups below 54 years. For the age groups above 54 years, however, the MRM estimator underestimates the population means quite noticeably. The reason for this behavior lies in the nature of the method. As an M-estimator, the method works by downweighting outlying observations; yet, for the age groups above 54 years, the method should actually react by up-weighting (which it is incapable of doing by design).[33] The method robust self-calibration produces values which are perfectly aligned with the known population means (as expected). If alignment is the only method selection criterion, we prefer robust self-calibration over the other methods.

For a comprehensive assessment of the estimation/ alignment methods, we shall also study the methods' MSE. To fix notation, let $\widehat{\mu}$ denote a generic estimator or alignment method. We estimate the MSE of $\widehat{\mu}$ by $\widehat{\mathrm{var}}(\widehat{\mu}) + (\widehat{\mu} - \bar{y})^2$. For the weighted sample mean (Hajek estimator) and the MRM method, we use standard (approximate) variance calculation procedures to compute $\widehat{\mathrm{var}}(\widehat{\mu})$; see e.g. Särndal et al. (1992, 182). Since robust self-calibration is not an estimating method, the aligned means have zero variance.[34] However, we shall nevertheless compute an approximate variance estimate for the robust self-calibration method. The variance estimator mimics the variance of the Hajek estimator, though it neglects

[33] In principle, an M-estimator may result in a behavior that corresponds to up-weighting. To this end, the M-estimator must downweight small values more than large values; hence, the estimate increases compared to a weighting scheme that downweights only large outliers. However, for our data, we were unable to tune the method in order to achieve such behavior.

[34] The zero variance property is a direct implication of the calibration constraints. To see this, consider (11) and note that the second constraint imposes estimator $\sum_{i \in s} v_i y_i$ to be aligned with $\sum_{i \in U} y_i$, which is a population quantity; hence, it has zero variance.

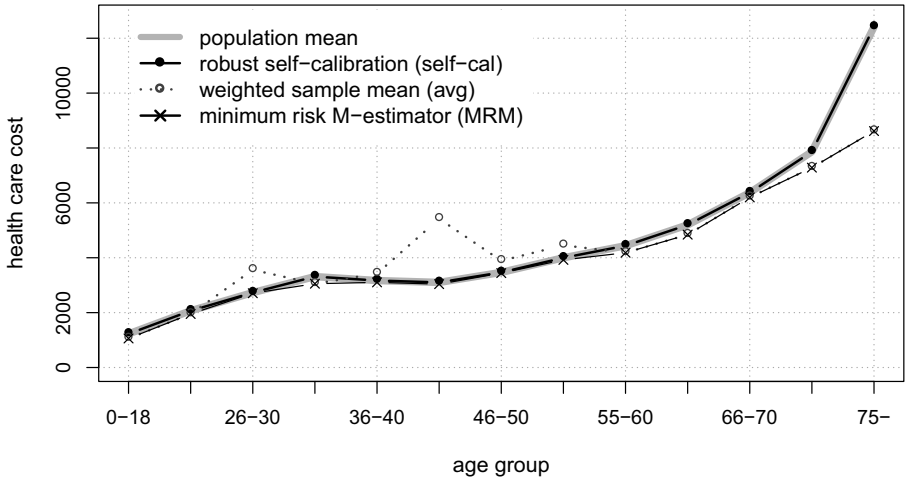

Fig. 8 Estimates (or aligned values) of average health care cost for women (by age group) for several estimation and alignment methods; the known population means are shown as a thick grey line; source: BAG (2017)

the fact that the calibrated sampling weights are dependent on the y_i's. As a result, the approximate variance estimator tends to underestimate the true variance.[35]

Table 4 shows the relative MSE (relMSE) for the methods in Fig. 8. The relMSE is the ratio of an estimator's MSE to the MSE of robust self-calibration. Values smaller (larger) than 1.0 indicate that the method under study is more (less) efficient than robust self-calibration. First, we note that the weighted sample mean (avg, Hajek estimator) is extremely inefficient compared with all other methods. Second, when $avg < \bar{y}$ (see last column of Table 4), the MRM estimator is as inefficient as method avg. In these cases, the estimate $\widehat{\mu}_{k_{opt}}$ is equal to avg because the penalty term (squared bias in the MSE, see Eq. 9) dominates the MSE and pulls the estimate onto avg. The MRM estimator can—in principle—escape from this trap if it would downweight small observations more than large outliers. However, it is not capable of doing so in our application. By contrast, MRM is more efficient than the robust self-calibration method in all cases where $avg > \bar{y}$. For some age groups, the gains in efficiency over self-cal are considerable (partly because we have tuned self-cal rather conservatively)[36]. Third, although method self-cal does not achieve the most efficient estimate/aligned value for one particular age group, it clearly shows the best *ensemble* efficiency (i.e., mean or total efficiency over all age groups). That is, self-cal achieves a fairly good compromise. Moreover, and when alignment is of key importance to the microsimulation modeler, self-cal is the preferred method because it ensures alignment at reasonable efficiency. For simulations with small

[35] The (standard) approximate residual variance estimator for the MRM method also tends to underestimate the variance. Therefore, the variance estimators for MRM and the robust self-calibration method are on an equal footing.

[36] See our discussion on the choice of the constants $q_i = r_i$ in Sect. 5.2.3.

Table 4 Estimates (by age group) of the relative mean square error (relMSE) for the methods robust self-calibration (self-cal), minimum risk $\boldsymbol{M}$-estimator (MRM), and the weighted sample mean (avg); relMSE is computed with respect to method self-cal; see text for further explanations

Age	Relative mean square error			avg $< \bar{y}$
group	self-cal	MRM	avg	
0–18	1.00	4.66	4.66	yes
19–25	1.00	1.21	1.21	yes
26–30	1.00	0.64	13.05	no
31–35	1.00	1.26	1.26	yes
36–40	1.00	0.51	1.91	no
41–45	1.00	0.33	49.11	no
46–50	1.00	0.78	4.15	no
51–55	1.00	0.96	4.32	no
55–60	1.00	1.32	1.32	yes
61–65	1.00	1.20	1.20	yes
66–70	1.00	0.97	0.97	yes
71–75	1.00	1.65	1.65	yes
75–	1.00	16.30	16.30	yes

sample sizes (not the case in our application), efficiency consideration become more important than perfect alignment; hence, MRM is a good choice.

Next, we address the robust estimation problem when the population means are unknown (see slices other than the top slice in Fig. 7). This time we consider estimation of average health care cost by household type. Clearly, we cannot use alignment methods. In principle, we could estimate average cost by a robust estimator of the Hajek mean for each category of variable household type. There is nothing wrong with this approach, except that it does not incorporate the auxiliary information from the alignment exercise at the top slice (to stick with the visual metaphor). In other words, this approach does not utilize the calibration principle of Deville and Särndal (1992, p. 376) that "weights that perform well for the auxiliary variables also should perform well for the study variable". Thus, we estimate the average costs by household type with the Hajek type estimator $\sum_{i\in s} w_i u_i y_i / \sum_{i\in s} w_i u_i$, where the u_i's depend on the method under consideration. For method avg, we have $u_i \equiv 1$; for MRM and self-cal, we take the u_i's that have been generated in the previous alignment exercise. Now, we cannot examine how close an estimate is to the population value since the latter is unknown. Therefore, we focus our discussion on the efficiency of the methods, measured by the variance of the estimators. We computed the relative variances (relVAR) with respect to method self-cal. Thus, values smaller (larger) than 1.0 indicate superior (inferior) efficiency compared with method self-cal; see Table 5.

The extreme outlier that we have already encountered in the alignment exercise (cost $\sim$ age group; see also Fig. 7) shows up in the household type "families with two or more children", and it inflates the estimated variance for the weighted sample mean (avg). The two other methods are robust against the outlier(s). Also, see from Table 5 that the MRM estimator has a smaller variance than method self-cal in households with children (and vice versa). This pattern is caused by the alignment methods and then "imported" to the current situation. That is, the MRM estimator was superior (with a few exceptions) in terms of efficiency for age groups below

Table 5 Estimated relative variance (relVAR) of the weighted sample mean (Hajek estimator) of health care costs by household type, where the weights are taken from robust self-calibration (self-cal), minimum risk $\boldsymbol{M}$-estimator (MRM), or just the sampling weights (method avg); relVAR is computed with respect to method self-cal

Household type	self-cal	MRM	avg
Single-parent families	1.00	0.98	1.00
Families with 1 child	1.00	0.99	1.28
Families with 2 or more children	1.00	0.91	3.14
Singles	1.00	1.11	1.09
Couples (no chilren)	1.00	1.13	1.01
Pensioner households	1.00	1.09	1.02

54 years (see Table 4). This effect then carries over to the current estimation problem because individuals in households with children (i.e., parents) range typically in age brackets below 54 years; as a result relVAR is lower. Since self-cal and MRM are so close in terms of relative variance, it is hard to prefer one method over the other. However, if take up the discussion of the previous paragraph, we may favor method self-cal if we value alignment (at the top slice) more than efficiency (and vice versa). Notably, in very small samples, efficiency considerations become more important and thus MRM is preferred over method self-cal.

6 Conclusion

The credibility of microsimulation modeling with the research community and policymakers depends on the availability of high-quality baseline surveys and the application of sound statistical methods. In this paper, we addressed two potential quality issues that both relate to skewed heavy-tailed distributions.

First, we reviewed how the presence of unit nonresponse can lead to biased simulation and estimates. In our application, we found that the top income bracket (and to a lesser extent also households in the lowest income bracket) are significantly under-represented in the baseline survey, compared with tax register data. Notably, we discovered that too few high- and ultra-high-income households were included in the sample because—as the literature shows—survey compliance decreases with increasing income. Other survey-related errors may have contributed to the under-representation of the top income bracket. Altogether, the estimate of average income underestimates the known population average. Based on progressive taxation, underestimation of the average implies downward-biased results for the simulation of taxes (and possibly other simulated variables). Although the Deville–Särndal calibration eliminated under-representation of the top income group, it could not achieve alignment of estimated average income in the right tail of the distribution with known population values *without* distorting the empirical distribution. The problem is rooted in the inability of the calibration method to cope with skewed heavy-tailed distributions. To overcome the problem, we introduced a parametric Pareto model to describe the right tail of the income distribution. With the help of the tail model, we adjusted the sample income distribution in the tail such that average income in the

top income bracket was aligned with known values. Henceforth, income data from the adjusted sample is more representative for the population distribution in terms of the first moment and with respect to tail probabilities. Our method of imputing expected order statistics from the Pareto distribution in place of the empirical order statistics has two major advantages over random imputation: the ranks of the observed household incomes are preserved, and the differences between observed and imputed values are small (except for the highest order statistics).

Under-representation of the top income bracket is a *common* issue of household surveys and is not limited to the Swiss SILC survey. This claim is substantiated by, for instance, the analysis of Törmälehto (2017) who presents empirical evidence of under-representation for 31 countries in the 2012 EU-SILC data, and the theoretical arguments in Korinek et al. (2006). Since sample surveys in general have difficulties in capturing top incomes, our method can be a useful tool for microsimulation modelers working with survey income data.

The second contribution of the paper also refers to the treatment of skewed heavy-tailed distributions. Here, we are concerned with variables from an outlier-prone, skewed population distribution that have been added to the baseline survey by record linkage. In our empirical application, individual health care costs from register data have been linked to the baseline survey. Because the baseline survey is a random sample with a small sampling fraction, the sampling weights (i.e., the inverse of the sample inclusion probabilities) are relatively large. An outlying observation in the cost data together with a large sampling weight can thus heavily influence or even ruin a sample estimate of the mean, total, or any similar characteristic. In contrast to our discussion on Pareto tail modeling for income, we have no comparable parametric model for health care cost; therefore, we adopt robust non-parametric methods.

In terms of methods, we distinguish between estimation and alignment methods for health care costs (by breakdown variables like age, income or household type). Alignment methods seek modifications of the data or the sampling weights such that the sample characteristics (e.g., mean or total) are aligned with *known* population values; hence, no estimation is required (unless we are interested in characteristics other than the ones that were benchmarked). However, the population characteristics of health care costs are only known for *some* breakdown variables. In our application, health care costs are known by age group, but not for other breakdown variables like `cost` $\sim$ `household type`. Therefore, we cannot impose alignment goals for average health care costs by household type. However, and by referring to the calibration principle of Deville and Särndal (1992, p. 376), that "weights that perform well for the auxiliary variables also should perform well for the study variable", we seek alignment for `cost` $\sim$ `age group` and then use the modified observations (or weights) for the analysis of `cost` $\sim$ `household type`.

Alignment and estimation methods are required to be outlier resistant. When non-robust alignment methods are applied to achieve alignment for one breakdown variable (e.g., `cost` $\sim$ `age group`), the cost data or weights are at risk of being distorted in the presence of outliers, which in turn may cause biased estimates for other breakdown variables (e.g., `cost` $\sim$ `household type`). Thus, we must avoid that an alignment problem in one place turns into an estimation problem at

another place. Any method that is based on non-robustly estimated sample-based characteristics (namely, the naive alignment method and the Deville–Särndal calibration method) is not protected against the presence of outliers. Therefore, we have proposed two alignment methods which are outlier resistant: robust self-calibration (self-cal) and the minimum risk M-estimator of the mean (MRM). The latter method is inspired by Hulliger (1995).

Our empirical analysis shows that the method self-cal achieves alignment with known population characteristics for reasonable levels of efficiency (mean square error, MSE) in the presence of outliers. In contrast, the weighted sample average is heavily influenced by outliers and is very inefficient. The MRM estimator does not impose explicit alignment goals and still produces estimates that are very close to the known population values with one exception: the MRM estimate is not even close to the benchmark when the sample mean is considerably smaller than the known population mean (formally, $\widehat{\bar{y}} < \bar{y}$). Apart from this case, the MRM is superior in terms of MSE. That being said, we prefer method self-cal over MRM when the sample size is relatively large for the following reasons: Self-cal achieves the alignment goals and its ensemble efficiency (i.e., total or mean efficiency over all categories of a breakdown variable, e.g., household type) is superior; in other words, self-cal achieves a good efficiency compromise. If, however, the sample size is small, efficiency considerations become more important. Hence, we favor the MRM estimator when $\widehat{\bar{y}} > \bar{y}$ because it exhibits gains in MSE over self-cal, and we suggest self-cal for the cases where $\widehat{\bar{y}} < \bar{y}$. Our methods are universally

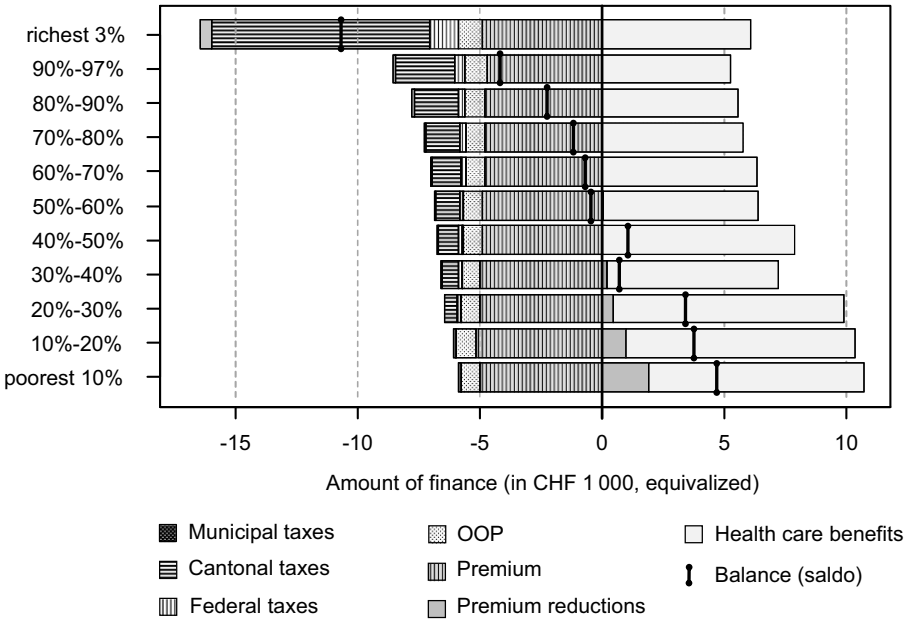

Fig. 9 Redistributive effects: Balance between payments made to and benefits received from CHI by household income bracket (see text for further explanations).

applicable to outlier-prone and skewed data when achieving alignment goals is demanded.

To illustrate the impact of the discussed methods, we study the redistributive effects in CHI by household income. Fig. 9 shows a comparison of average payments made to the system (taxes, premium, OOP) versus average financial aids (e.g. premium reductions) and average health care benefits received from CHI by income bracket. Payments and benefits are equivalized by the EUROSTAT equivalization scale[37] for reasons of comparison. Moreover, Fig. 9 does not contain confidence intervals for ease of simplicity. We observe from the display that households above the 40%-50% income bracket are net payers (see balance/ saldo). It is also noteworthy that households in the top income bracket make (mainly through taxes) a major financial contribution to the system. If the Pareto tail adjustment for the income distribution is omitted, we would observe significantly underestimated tax payments in the top income bracket. Fig. 9 shows other interesting facts—to be discussed elsewhere. We refer the reader to Schoch et al. (2013), where we study other breakdowns (e.g., gender, household composition) and more sophisticated measures of the redistribution effects (e.g., Gini coefficient).

Acknowledgements We are grateful to Beat Hulliger and three anonymous reviewers whose comments helped to improve the paper. Special thanks go to the authors of the R software packages `data.table` (Dowle and Srinivasan 2018) and `survey` (Lumley 2019).

Funding Open access funding provided by FHNW University of Applied Sciences and Arts Northwestern Switzerland

Appendix

A Model overview

In this paragraph, we give a brief overview of the microsimulation model. All simulations refer to the baseline survey; we do not simulate at the level of the population.

Simulation model (SILC 2016 data)

Let s_H and s_I denote, respectively, the sample of households and individuals of the baseline survey. We denote by f a generic function. We write $\texttt{variable}_i$ to mean

[37] This scale is also called OECD-modified scale and assigns a value of 1.0 to the household head, a value of 0.5 to each additional adult member and 0.3 to each child.

the value of a (generic) variable for individual or household i. For ease of display, we work with *umbrella terms*, e.g. income shall stand for—depending on the context—gross, net personal or household income (incl. employee income, benefits or losses from self-employment, pensions, old-age benefits, housing allowances, inter-household cash transfers, etc.), or disposable household income, etc.

Taxes: For each individual i living in household $h \in s_H$, tax payments are simulated (separately at the federal, cantonal and municipality level) by

$$\begin{aligned}\texttt{tax}_i \leftarrow f(&\texttt{income}_i, \texttt{place of residence}_i, \texttt{household type}_i,\\ &\texttt{marital status}_i, \texttt{age}_i, \texttt{canton}_i)\end{aligned}$$

The totals of the simulated tax revenues (at the federal, cantonal and municipality level) are aligned with the known population totals.

Premium reductions: For each couple or family $h \in s_H$ (or individual $i \in s_I$), it is determined whether it is entitled to premium reductions (and if so to what extent), using

$$\begin{aligned}\texttt{premium reductions}_i \leftarrow f(&\texttt{income}_i, \texttt{canton}_i, \texttt{place of residence}_i,\\ &\texttt{household type}_i, \texttt{marital status}_i,\\ &\texttt{tax deduction}_i, \texttt{family allowances}_i,\\ &\texttt{number of children}_i).\end{aligned}$$

In addition, for each child or young adult (18–25 years old) j in family or household h, it is checked whether child or young adult j is entitled to personalized premium reductions if the family is not eligible,

$$\begin{aligned}\texttt{premium reductions}_j \leftarrow f(&\texttt{income}_j, \texttt{age}_j, \texttt{place of residence}_j,\\ &\texttt{undergoing education}_j, \texttt{canton}_j).\end{aligned}$$

The averages of the simulated $\texttt{premium reductions}_i$ and $\texttt{premium reductions}_j$ are aligned with the known population averages by canton × age group, where age group is one of the categories: child, young adult or adult.

Deductible and premium of CHI: For each individual $i \in s$, the deductible (choices: 300, 500, 1000, 1500, 2000 or 2500 CHF) and the premium are taken from register data and matched to the baseline survey by record linkage,

$$(\texttt{deductible}_i, \texttt{premium}_i) \leftarrow \texttt{register data}[(\texttt{deductible}_i, \texttt{premium}_i)],$$

The frequency distribution of $\texttt{deductible}_i$ is aligned with the known frequency distribution (by canton × age group, where age group is one of the categories: child, young adult or adult).

Health care costs: For each individual $i \in s$, the amount of incurred health care costs are taken from register data and matched to the baseline survey by record linkage,

$$\texttt{health care cost}_i \leftarrow \texttt{register data}[\texttt{health care cost}_i],$$

and the averages of $\texttt{health care cost}_i$ are aligned with the known population averages (by $\texttt{sex} \times \texttt{canton} \times \texttt{age group}$) using the methods discussed in the paper (variable $\texttt{age group}$ is a categorical variable with age binned into brackets of size 5 years, incl. the boundary intervals [0,18] and [76, max]).

Out-of-pocket payments (OOP): For each individual $i \in s$, the out-of-pocket payments are simulated by

$$\begin{aligned}\texttt{out-of-pocket payments}_i \leftarrow f(&\texttt{deductible}_i, \texttt{health care cost}_i\ \texttt{income}_i,\\ &\texttt{household type}_i \texttt{means-tested benefits}\\ &\texttt{from OASI}_i),\end{aligned}$$

where OASI denotes old-age and survivor's insurance. Total OOP is aligned with the known population total (by canton).

Breakdown variables for the analysis

The redistributive effects are studied by contrasting 1) payments to the system (taxes at the federal, cantonal and municipality level, out-of-pocket-payments, premiums) with 2) financial benefits received from the system (premium reductions, health care benefits and means-tested benefits to finance out-of-pocket payments) by breakdown variable; see Fig. 9. Our basic model implements the following breakdown variables:

(i) individual level
  (a) age (categorical variable with 12 categories)
  (b) gender (men and women)
  (c) nationality (Swiss and foreigner)
  (d) health status (categorical variable with 5 categories)

(ii) household/family level
  (a) equivalized disposable household income (categorical variable with 11 categories)
  (b) household type (categorical variable with 6 categories; e.g., couples without children)
  (c) cross (or Cartesian) product of household type and equivalized disposable household income

The effects for these breakdown variables have been published in Schoch et al. (2013) and BAG (2018). Our model is technically not limited to this choice of breakdown variables.

Earlier simulation model of Schoch et al. (2013)

In our earlier model, register data were not available. Therefore, the variables $\texttt{deductible}_i$, $\texttt{premium}_i$, and $\texttt{health care cost}_i$ were simulated. We adhere to the notation introduced above.

Deductible and premium of CHI: For each adult $i \in s$, the CHI deductible (choices: 300, 500, 1000, 1500, 2000 or 2500 CHF) is simulated by

$$\texttt{deductible}_i \leftarrow f(\texttt{income}_i, \texttt{age}_i, \texttt{health status}_i, \texttt{health care costs}_i, \texttt{education}_i, \texttt{sex}_i, \texttt{nationality}_i, \texttt{canton}_i).$$

The frequency distribution of $\texttt{deductible}_i$ is aligned with the known frequency distribution. The same method was used to simulated deductibles of children. The premium (which depends on $\texttt{deductible}_i$) is simulated by

$$\texttt{premium}_i \leftarrow f(\texttt{income}_i, \texttt{age}_i, \texttt{place of residence}_i, \texttt{canton}_i, \texttt{deductible}_i),$$

and average simulated $\texttt{premium}_i$ is aligned with the known population averages by $\texttt{deductible} \times \texttt{canton} \times \texttt{age group}$, where $\texttt{age group}$ is one of the categories: child, young adults or adult.

Health care costs: For each individual $i \in s$, the incurred health care costs were simulated by

$$\texttt{health care cost}_i \leftarrow f(\texttt{number of doctor visits}_i, \texttt{health status}_i, \texttt{chronic diseases}_i, \texttt{number of days with in-patient treatment}_i, \texttt{age}_i, \texttt{sex}_i, \texttt{restriction in everyday life due to illness}_i,),$$

and average of simulated $\texttt{health care cost}_i$ is aligned with the known population averages by $\texttt{sex} \times \texttt{canton} \times \texttt{age group}$; variable $\texttt{age group}$ is age grouped into brackets of size 5 years (incl. the boundary intervals [0,18] and [76, max]).

Means-tested benefits from old-age and survivor's insurance (OASI) to finance OOP:

$$\texttt{OASI-benefits}_i \leftarrow f(\texttt{income}_i, \texttt{household type}_i, \texttt{age}_i, \texttt{allowances from OASI}_i),$$

where OASI denotes old-age and survivor's insurance. The total is aligned with the known population total (by canton).

B Calibration estimation

In this paragraph, we give a brief overview of the traditional calibration method of Deville and Särndal (1992, from here on DS92). Suppose that a sample s of size n has been drawn from population U (of size N). In the sample, we observe the

variable of interest $y_i \in \mathbb{R}$ and a vector of auxiliary variables $\boldsymbol{x}_i \in \mathbb{R}^p$ $(i \in s)$. The population x-totals, $\boldsymbol{T}_x = \sum_{i \in U} \boldsymbol{x}_i$, are assumed to be known; by contrast, the population y-total T_y is unknown.

In practice, the calibration method is mainly used for two reasons: (i) Alignment: The data are re-weighted such that the estimated x-totals are aligned with the known $\boldsymbol{T}_x$ (i.e., the sampling weights are modified to incorporate auxiliary information). (ii) Efficiency: The calibrated weights are used to compute linearly weighted estimates, e.g., to estimate the y-total, under the hypothesis that "weights that perform well for the auxiliary variables also should perform well for the study variable" (DS92, p. 376).

The calibration method of DS92 seeks to compute modified weights v_i $(i \in s)$ that differ from the initial sampling weights w_i as little as possible and which maintain the calibration constraints on the r.h.s. in (13). Formally, DS92 suggest solving the optimization problem,

$$\min \frac{1}{2} \sum_{i \in s} \frac{(v_i - w_i)^2}{w_i} \qquad \text{subject to the constraints} \quad \sum_{i \in s} v_i \boldsymbol{x}_i = \boldsymbol{T}_x, \tag{13}$$

for the weights v_i $(i \in s)$. The new weights can then be used in place of the original sampling weights. Moreover, DS92 introduce the class of calibration estimators,

$$\widehat{T}_{y,cal} = \sum_{i \in s} v_i y_i, \tag{14}$$

for estimating the population y-total. For the distance function in (13), i.e. the objective function of the minimization problem, the estimator in (14) coincides with the generalized regression estimator (GREG). Hence, $\widehat{T}_{y,cal}$ inherits the nice properties of the GREG (i.e., efficient incorporation of auxiliary variables, asymptotic-design unbiasedness, etc.) and is (usually) more efficient than the Horvitz–Thompson estimator of T_y when y_i and $\boldsymbol{x}_i$ are correlated. The distance function in (13) has the disadvantage that some of the weights v_i can be negative. As a remedy, DS92 propose alternative distance functions which ensure positivity of the weights. The calibration method of DS92 is not robust against outliers or influential values in y_i or $\boldsymbol{x}_i$ (or both); see Duchesne (1999).

References

Alfons A, Templ M, Filzmoser P (2013) Robust estimation of economic indicators from survey samples based on Pareto tail modeling. J R Stat Soc Ser C Appl Stat 62:271–286

BAG (2017) Statistik der obligatorischen Krankenversicherung. Schweizer Bundesamt für Gesundheit, Bern

BAG (2018) Monitoring 2017: Wirksamkeit der Prämienverbilligung. Schweizer Bundesamt für Gesundheit, Bern

BFS (2016) Synthesebericht zur Revision der SILC 2014. Schweizer Bundesamt für Statistik, Neuchâtel

BFS (2017) Quality report: 2014 EU-SILC cross-sectional data, Switzerland. Schweizer Bundesamt für Statistik, Neuchâtel

Biemer PP, Lyberg L (2003) Introduction to survey quality. John Wiley & Sons, Hoboken

Biewen M (2001) Item non-response and inequality measurement: evidence from the German earnings distribution. AStA Adv Stat Anal 85:409–425
Bourguignon F, Spadaro A (2006) Microsimulation as a tool for evaluating redistribution policies. J Econ Inequal 4:77–106
Chambers R (1986) Outlier robust finite population estimation. J Am Stat Assoc 81:1063–1069
Cowell FA, Victoria-Feser MP (2007) Robust stochastic dominance: a semi-parametric approach. J Econ Inequal 5:21–37
Creedy J, Tuckwell I (2004) Reweighting household surveys for tax microsimulation modelling: an application to the New Zealand Household Economic Survey. Aust J Labour Econ 7:71–88
Crivelli L, Filippini M, Mosca I (2006) Federalism and regional health care expenditures: an empirical analysis for the Swiss cantons. Health Econ 15:535–541
David HA, Nagaraja HN (2003) Order statistics, 3rd edn. John Wiley & Sons, Hoboken
Dell F, Piketty T, Saez E (2007) Income and wealth concentration in Switzerland over the twentieth century. In: Atkinson AB, Piketty T (eds) Top incomes over the 20th century: a contrast between continental European and English-speaking countries. Oxford University Press, Oxford, pp 472–500 (chap 11)
Deville JC, Särndal CE (1992) Calibration estimators in survey sampling. J Am Stat Assoc 87:376–382
Doorslaer EV, Wagstaff A, der Burg HV, Christiansen T, Citoni G, Biase RD, Gerdtham UG, Gerfin M, Gross L, Häkinnen U, John J, Johnson P, Klavus J, Lachaud C, Lauritsen J, Leu R, Nolan B, Pereira J, Propper C, Puffer F, Rochaix L, Schellhorn M, Sundberg G, Winkelhake O (1999) The redistributive effect of health care finance in twelve OECD countries. J Health Econ 18:291–313
Dowle M, Srinivasan A (2018) data.table: Extension of 'data.frame'. https://cran.r-project.org/package=data.table (R package version 1.11.4). Accessed 28 Dec 2019
Drabinski T (2004) Umverteilungseffekte des deutschen Gesundheitssystems: Eine Mikrosimulationsstudie. Schriftenreihe des Instituts für Mikrodaten-Analyse, vol 2. Schmidt & Klaunig, Kiel (zugl. Diss. Univ. Kiel)
Duchesne P (1999) Robust calibration estimators. Surv Methodol 25:43–56
ESTV (2017) Direkte Bundessteuer: Steuerperiode 2016: Natürliche Personen. Eidgenössische Steuerverwaltung, Bern
Figari F, Paulus A, Sutherland H (2014) Microsimulation and policy analysis. In: Atkinson AB, Bourguignon F (eds) Handbook of income distribution. Handbooks in economics, vol 16. Elsevier, Amsterdam, pp 2141–2221 (chap 24)
Foellmi R, Martínez IZ (2017) Volatile top income shares in Switzerland? Reassessing the evolution between 1981 and 2010. Rev Econ Stat 99:793–809
Frick JR, Grabka MM (2005) Item nonresponse on income questions in panel surveys: Incidence, imputation and the impact on inequality and mobility. All Stat Arch 89(1):49–61
Godambe VP, Thompson ME (1986) Parameters of superpopulation and survey population: their relationships and estimation. Int Stat Rev 54:127–138
Grabka MM (2004) Alternative Finanzierungsmodelle einer sozialen Krankenversicherung in Deutschland: Methodische Grundlagen und exemplarische Durchführung einer Mikrosimulationsstudie. Technische Universität Berlin, Berlin (Diss. Fakultät Wirtschaft und Management)
Groves RM, Couper MP (1998) Nonresponse in household interview surveys. John Wiley & Sons, New York
Groves RM, Dillman DA, Eltinge JL, Little RJA (eds) (2002) Survey nonresponse. John Wiley & Sons, New York
Hannappel M, Troitzsch KG (2015) Mikrosimulationsmodelle. In: Braun N, Saam NJ (eds) Handbuch Modellbildung und Simulation in den Sozialwissenschaften. Springer, Wiesbaden, pp 455–489 (chap 16)
Hulliger B (1995) Outlier robust Horvitz–Thompson estimators. Surv Methodol 21:79–87
Hulliger B, Schoch T, Sterchi M (2019) robsurvey: robust survey statistics estimation. https://cran.r-project.org/package=robsurvey (R package version 0.1-1). Accessed 28 Dec 2019
Kleiber C, Kotz S (2003) Statistical size distributions in economics and actuarial sciences. John Wiley & Sons, Hoboken, NJ
Klevmarken NA (2002) Statistical inference in micro-simulation models: Incorporating external information. Math Comput Simul 59:255–265
Korinek A, Mistiaen JA, Ravallion M (2006) Survey nonresponse and the distribution of income. J Econ Inequal 4:33–55
Lee H (1995) Outliers in business surveys. In: Cox BG, Binder DA, Chinnappa BN, Christianson A, Colledge MJ, Kott PS (eds) Business survey methods. John Wiley & Sons, New York, pp 503–526 (chap 26)

Li J, O'Donoghue C, Loughrey J, Harding A (2014) Static models. In: O'Donoghue C (ed) Handbook of microsimulation modelling. Contributions to economic analysis, vol 293. Emerald, Bingley, pp 47–75 (chap 3)
Lohr SL, Raghunathan TE (2017) Combining survey data with other data sources. Stat Sci 32(2):293–312
Lumley T (2019) survey: analysis of complex survey samples. https://cran.r-project.org/package=survey (R package version 3.35-1). Accessed 28 Dec 2019
Müller A, Schoch T (2014a) Umverteilung in der Krankenversicherung: Eine Mikrosimulationsstudie. Soz Sicherh CHSS 6:180–183
Müller A, Schoch T (2014b) Vermögenslage der privaten Haushalte: Vermögensdefinitionen, Datenlage und Datenqualität. Wirtschaftliche und soziale Situation der Bevölkerung, vol 20. Bundesamt für Statistik, Neuchâtel
Myck M, Najsztub M (2015) Data and model cross-validation to improve accuracy of microsimulation results: Estimates for the Polish Household Budget Survey. Int J Microsimul 8:33–66
OECD (2011) OECD reviews of health systems: Switzerland 2011. OECD, Paris
R Core Team (2019) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. https://www.R-project.org/. Accessed 28 Dec 2019
Särndal CE (2007) The calibration approach in survey sampling. Surv Methodol 33:99–119
Särndal CE, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer, New York
Schoch T, Müller A, Bachmann T, Kraft E, Mattmann M, Walker P (2013) Umverteilungseffekte in der obligatorischen Krankenversicherung: Mikrosimulation für die Schweizer Bevölkerung auf Basis der SILC-Erhebung unter Berücksichtigung der kantonalen Strukturen: Studie zuhanden des Bundesamts für Gesundheit. Ecoplan, Bern
Schofield DJ, Zeppel MJB, Tan O, Lymer S, Cunich MM, Shrestha RN (2018) A brief, global history of microsimulation models in health: past applications, lessons learned and future directions. Int J Microsimul 11:97–142
Spielauer M (2011) What is social science microsimulation? Soc Sci Comput Rev 29:9–20
Thompson ME (2019) Combining data from new and traditional sources in population surveys. Int Stat Rev 87(S1):79–89
Törmälehto VT (2017) High incomes and affluence: evidence from EU-SILC. In: Atkinson AB, Guio AC, Marlier E (eds) Monitoring social inclusion in Europe. EUROSTAT, Luxembourg, pp 123–140
Wright RL (1983) Finite population sampling with multivariate auxiliary information. J Am Stat Assoc 78:879–884