Harvesting collective agreement in community oriented surveys: the medical case.

Federico Cabitza

Abstract The chapter discusses the role of simple and lightweight Web-based systems in promoting a different approach to the externalization of practice-related knowledge within communities of professionals. This approach exploits common online questionnaire systems to collect the preferences of large numbers of domain experts to interesting paradigmatic work cases and proposes a statistically sound evaluation of these responses to evaluate the agreement reached within the community. We tested this approach in a case study that involved a large international medical association, that we chose as an example of a large and highly distributed community of expert professionals; in this study we challenged more than 1,000 surgeons about some border-line clinical cases where tacit notions based on life-long practice and situated experiences coexist (and sometimes clash) with scientific evidences drawn from the specialistic literature. We make the point that a sound evaluation of the collective agreement is a necessary precondition to use such lean Web-based tools in bottom-up knowledge elicitation initiatives. To this aim, existing measures of agreement and survey-related heuristics can be exploited to get a more precise picture of the "opinion of the many" in collective settings like communities of practice.

1 Background and motivations

Surveys and polls are important means to collect information since the beginning of the nineteenth century [20] and the rationalization of both the instruments and methods to process, analyze and interpret the collected data can be said coeval to the development of the machines that were to evolve in the modern programmable computers [21]. Surveys have been used extensively in different disciplines, like sociology, psychology, economics, education and also in medicine. In this domain,

Università degli Studi di Milano-Bicocca, Viale Sarca 336, I - 20126 Milano (Italy), e-mail: cabitza@disco.unimib.it

they have been frequently used for a range of different objectives, from the collection of data on the beliefs, attitudes and behaviors of both patients and doctors [28], to the collection and analysis of quantitative data that are essential in clinical epidemiology and health services research [33]. In the last ten years, due to the crescent diffusion of personal ICTs, either computer-assisted or online questionnaires have been increasingly employed to leverage the clear advantages that Web-based surveys provide in terms of both total costs, recruitment effort and analysis efficiency. A simple research on Pubmed on the expressions "Web-based surveys" and "online surveys" in either the title or abstract of indexed publications reveals less than five papers written before year 2000 against more than 1700 contributions written since 2000 to date.

Yet, despite their number, their cost and the level of interest in their findings, surveys conducted among physicians usually get inadequate response rates, thus raising concerns that non-response bias could affect the validity and generalizability of the findings [28]. Quite surprisingly (if considering, e.g., the very high level of digitization of general practice in Europe [9]) the phenomenon of low response rates is observed to be significantly *worse* for Web-based surveys in comparison to surveys that are based on traditional mail and telephone [5, 43, 32, 47, 29]. Many researchers have focused on various and complementary strategies for achieving higher rates of response [45] where assurance of confidentiality, monetary incentives and keeping questionnaires short are the most frequent recommendations within more general and complex framework for sound surveying (e.g. [8]). Since no magic formula probably exists, even recently researchers advocate further studies in this vein [14]. Our point is that one of the reasons why surveys do not attract a vast interest from doctors lies in the actual use that researchers make of the results coming from such initiatives. For instance, when a survey is employed to probe respondents on their preferences about clinical procedures and indications, the related analyses are usually kept at the simple level of comparing response percentages and the agreement among respondents is characterized in very qualitative ways, e.g., by merely confronting percentages of responses (e.g., [34]). This a problem that begins to emerge even in the specialist literature [10] and that, together with the correlated problem of low response rates, undermines the reliability of these initiatives, and therefore their potential to become reliable but lightweight methods to extract tacit knowledge from the grassroot level of medical communities.

This will be the main point of the chapter, as it will be articulated in some detail in Section 2 and in particular in Section 5, where we will briefly discuss how to possibly improve reliability in medical surveys. Section 3 will review some of the methods by which the agreement among the respondents of a survey can be measured with some objectivity, as a necessary precondition to ground on these estimates the externalization/production of new knowledge on the practices carried out within the surveyed community; Section 4 reports a case study where these ideas were first applied all together to extract consensus-based "best practices" from a medical association counting more than 1,000 members all around the world. Section 6 will conclude the chapter by providing some indications of how this work could be extended in next similar initiatives.

2 Surveys as tools for knowledge externalization

Our initial point is that the real impact that surveys may have on the process of knowledge externalization and sharing within a community does not lie on the instrument in itself, but rather on the method and rigour with which questionnaires are designed and applied, responses are collected and analysed and results are interpreted and returned to the respondents and policy makers: this is especially true in those ambits where the existence of reliable and socially valued knowledge has an important impact on professional practice and this, in its turn, has an impact on themes that rightly catch the attention of the commonalty, like healthcare quality, treatment appropriateness and patient safety [35, 6]. Our more specific point is that online questionnaire systems have a potential to contribute to medical knowledge (and hence medical practice) that has not fully been explored so far, especially in light of the increasing pervasiveness that characterizes Web-based technologies and, in particular, of the recent interaction modalities that are enabled by the so called “Web 2.0” platforms [46].

The key point here is “scalability”. It has always been impractical for investigators to administer a questionnaire to all potential respondents in a target population, e.g. the members of a community of practice; for this reason, it has been a common (yet delicate) practice to address a so called “sampling frame” instead of the whole population, i.e., the target population from which to extract a sample by means of different sampling techniques according to the research objectives and resources (e.g., random sampling, cluster sampling). Nowadays in the healthcare sector, like many other sectors where public registries of professionals exist and are constantly maintained, it is much easier and cheaper to identify large populations of possible respondents and contact them at very low cost; indeed, almost every member of a speciality association and employee of an healthcare facility has (at least) an email address and most of the addresses of health professionals can be found in either private or public registries (e.g., MMS[1], NHSnet[2]). The large numbers that ICT can help achieve account for nothing less than *increased* precision and descriptive power, as it is known that the margin of error in survey-based researches does not depend on the size of the population of interest but rather on the sample size (at a desired confidence level). But the larger the numbers involved, not only the higher the statistical precision; we may wonder if reaching the grassroot levels of an arbitrarily large community of professionals could allow the investigators to end up by probing phenomena that are more traceable back to the concept of “collective intelligence” [22, 27] than to those of either census of practices or survey of attitudes/preferences. In what follows, we will concentrate on these themes in the specific domain of medicine, but we are confident that many of the results gained in such a delicate and knowledge-intensive domain can be also applied or better yet, can inform, research that is oriented to other professional domains.

[1] http://www.mmslists.com/

[2] http://www.nhs.net

In the healthcare domain, relatively small-sample surveys have been recently employed to assess knowledge of and compliance with evidence-based recommendations (e.g. [36, 48]) of practitioners in their actual practice. In this same vein, our current research question is whether community-wide online surveys can contribute in shedding light on those “treatments of choice” and rationales for decision (e.g., particular indications that make a specific treatment or procedure advisable) that the majority of practitioners prefer over possible alternatives in the context of borderline and exemplificatory cases; and whether such indications, which attract the interest and earn the preference of a multitude of (usually silent) experts, can rise to the status of *collective consensus-based recommendations*. As ICT researchers, we focus on the tools that could enable the explicitation and formalization of these recommendations and leave to the debate of the medical communities whether the preferences that emerge from the “rank and files” of hospitals and private practices could flank and, maybe, complement, what are now considered the “scientific evidences” of lowest level in medicine, the so called evidences of level III or D [3], which are built on the basis of the opinions of (a limited number of) respected authorities or of (allegedly) respected and influential expert committees.

To address this research question, our approach focuses on a rigorous approach towards two aspects of (online) surveys: *generalizability* of findings (tackled from the perspective of the interrater reliability); and “consensus quality” (or *agreement assessment*). Agreement assessment regards how much it is true that the practitioners agree on a specific treatment. To this respect, our contribution will be presented in Section 3 and regards the experimentation and consolidation of heuristics proposed elsewhere in the literature, as well as the introduction of a novel measure, that is simple to calculate and is more sensitive to large numbers of respondents. Generalizability regards how much the collected respondents are representative of a whole category of practitioners; to this respect, in Section 5 we will discuss a novel heuristics to determine the optimal timing for sending a reminder and therefore to help assess and minimize non-response bias.

2.1 The quest for objectivity in (online) surveys

For the success of initiatives of online surveying, scholars have underlined the importance of proper design and of compliance to recommendations that gave some evidence of efficacy (e.g., [1, 47, 6, 14]). In addition to good design (e.g., survey length) and effective strategies (e.g., incentives, reminders), it is also important to make the best use of the responses collected. To this aim, proper analytical procedures and techniques must be applied in a proper manner. It is the employment of these techniques (eventually embedded in the response processor of the computer-based system) that makes a research survey essentially different from the most trivial online survey systems like, for instance, the almost ubiquitous Web-based *polls*; these are usually very lean tools, more and more frequently made available in mashups form, that are becoming increasingly common in several web sites, ranging between institutional newspaper websites, corporate portals and social network platforms (e.g., blogs) to probe the readers' opinion on almost any subject.

The differences between knowledge-oriented surveys and discussion-oriented polls lie, as simple as it can be, on how questions are asked and how responses are analyzed; or, in more technical terms, on the "study design" and on the "hypothesis formulation". In fact, assuming that the right questions (i.e., which are able to raise the interest of a competent community) are formulated in the right manner (i.e., without revealing a specific preference or discriminating and biasing the respondents), the point is not in the underlying technology per se, but on the result analysis and how this can affect discussion and consensus-building practices that are triggered and supported by incremental uses of the system. In particular, to detect consensus and concordance patterns in a heterogeneous and distributed community of practitioners, such a system must be designed so that statistical processing of responses can be carried out as objectively as possible. Only in this way, the system can be seen as a technology at least potentially capable to contribute – in ways that were simply not possible before – in building wide-consensus-based evidences that take the real attitudes, habits and practices of the "rank and file" of practitioners into account.

The simplest output that a system collecting responses to closed questions can provide is the frequency with which each possible alternative has been selected by the respondents. Yet, knowing how these frequencies distribute over the full range of responses is not sufficient to draw the knowledge that can enable further informed discussion and comparison of alternatives, as it is expressed in the following two questions: "what is the degree of agreement that we observe among respondents (in our case, doctors)?"; and "to what extent these agreements can be generalized and reproduced?". While the first question seems to refer only to what scholars call *interrater reliability*, the latters seems to refer to generalizability and replicability; as a matter of fact, both aspects are tightly intertwined.

In the domain of treatment evaluation, generalizability (and therefore the external validity of the user study) refers to the extent to which the assessment of appropriateness for a specific treatment is influenced by the specific doctors involved; or, alternatively, to the extent a different group of doctors would have yield the same responses within a tolerable margin of error. This is a thorny matter since much of what pertains to doctors' evaluation (e.g., diagnosis, prognosis) is bound to their specific experience, sensitivity and interpretative capabilities; to this respect, subjectivity and interpretive differences are unavoidable factors of medical profession and could be considered a source of bias only from the merely statistical point of view. Yet, provided that researchers have selected respondents randomly from a population of potential practitioners, if they detect that the respondents' answers are affected by an excessive degree of subjectivity this could indicate the need for further refinement for either the case descriptions, the expression of the alternatives or the choice of the evaluation categories. For this reason, reproducibility can be seen as a kind of reliability, arguably the strongest to achieve and demonstrate. Reliability accounts for the degree of agreement that is observed among independent observers; therefore, the more observers agree on the responses they provide, the more comfortable we can be that their responses are exchangeable with those provided by other observers [18], reproducible, and trustworthy.

3 How to "gauge" collective agreement?

In this section we focus on interrater reliability. In Section 5 we will see a contribution to make a reliable survey more soundly generalizable. Interrater reliability is defined as the extent to which different evaluating doctors (more than two and independent from each other), each assessing the same treatment for the same case, come to the same decision, i.e., either select the same treatment/option or assign the same appropriateness category for the option in hand.

The simplest and most common method of reporting interrater reliability is the percent agreement statistic, also called 'proportion of agreement' (Po). Po is an estimation of the probability that two (randomly selected) raters assign the same appropriateness grade to a given treatment. Unfortunately, this measure tends to overestimate the degree of clinically important agreement since it does not take into account the agreement that would have been expected due solely to chance[3]. To overcome this shortcoming, scholars often employ the Fleiss's (multirater) Kappa score[4]. This score is interpretable as a measure of agreement beyond that due solely to chance, where values between 1 and 0 indicate agreement better than chance, a value of 0 indicates a level of agreement that could have been expected by chance, values between 0 and -1 indicate levels of agreement that are worse than chance.

Although Kappa is a measure of interrater reliability that is often found in literature, some authors have argued that is not suitable for the majority of agreement analysis. In fact, the Fleiss' Kappa is overly conservative [39] especially when evaluation includes several categories. In these cases the possibility of chance agreement appears negligible, thus leading to several cases in which the Kappa score is very low even when proportions of agreement are very high [11]. Moreover, others argue that the Fleiss Kappa should be applied only when the number of ratings on each subject (treatment) is constant [41] and when assessments are limited to nominal data with no clear order between categories [19].

Since ordinal variables are a convenient way to assess both the appropriateness of treatments and the level of personal accordance with some option, in our analyses we opted for two complementary measures that can be used with any number of observers, namely the free-marginal multirater Kappa [41], which is not influenced by prevalence bias and do not require an a priori knowledge of marginal distributions; and the Krippendorff's Alpha [19], which generalizes across different scales of measurement and can be computed with or without missing data. Both are scores that define a reliability scale from 1 for perfect agreement and 0 for absence of agreement. Yet, a test of the null hypothesis that all agreement is due to chance and

[3] Obviously a doctor does not choose a treatment by chance; but it is the doctors that are (should be) recruited by chance.

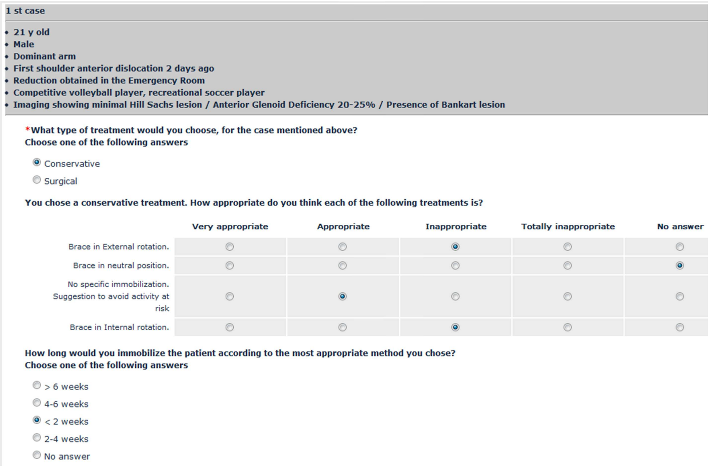
[4] Not to be confused with the Cohen Kappa, suitable for multi-case two-rater assessments.

agreements are not reliable (K=0, A=0) often relates to a statistically significant but mediocre level of real agreement. For this reason, to assess the strength of agreement conventional benchmarks are proposed in the literature: in regard to the Kappa score, Landis and Koch [31] characterized values of Kappa less than 0 as indicating no agreement, values between 0 and .20 as slight, .21 and .40 as fair, .41 and .60 as moderate, .61 and .80 as substantial, and between .81 and 1 as almost perfect agreement; Fleiss proposed a two threshold benchmark where values less than .4 are associated with poor agreement, values between .4 and .75 with intermediate to good agreement and values above .75 with excellent agreement [15]. In regard to the Krippendorff's Alpha, social scientists commonly consider generalizable agreements with reliability greater than .8, while they draw only tentative conclusions for data whose agreement measures are less than that threshold [30].

In addition to Alpha and Kappa scores, we also propose a heuristic-based measure of agreement that is simply based on a Chi Square test (performed on the difference in distributions between alternative categories), from the obvious observation that the higher the Chi square score, the more polarized responses are (that is, the farer the respondents from perfect balance between the options). This latter score is highly sensitive to large samples of respondents, assuming that the fact that several people agree on a specific option is more significant than the mere proportions between options (i.e., 8 vs. 2 is associated to a smaller agreement than 800 vs. 200). Our simulations carried out on samples of more than thirty respondents for both nominal and ordinal data[5] led to proposing the following indicative thresholds: no agreement below 10, poor agreement up to 20, good agreement from 30 up (obviously the higher the score, the better the agreement).

All that said, although interrater reliability is a generally accepted indicator of how much consensus lies in the ratings/opinions given by a group of people/experts, there is no general... consensus on how this indicator should be precisely measured and related scores interpreted. This uncertainty calls for further research and for validation of formal approaches through tests run at the field of work. In those cases, we should also address the fact that an excellent agreement on a specific treatment choice (e.g., like that reported in Section 4 for the case no. 8) should not induce researchers to discarding the opinion of small minorities, but rather bring them to considering why, say, one (out of ten) expert doctor expresses an opinion that contrasts that of the other nine. In fact the point of initiatives where agreement is assessed should not in erasing differences or in making opinions more extreme [2]; but, on the contrary, these systems should be seen as tools to let hidden common preferences emerge, share rationales for such preferences, make practitioners aware of differences and help them understand the reasons for divergence, as well as probe them whether a candidate solution can be the best compromise.

[5] In this latter case we compared the values of Chi with the Kendall's coefficient of concordance, which is a normalized score between 0 and 1 as the above mentioned Alpha and Kappa.



1 st case

- 21 y old
- Male
- Dominant arm
- First shoulder anterior dislocation 2 days ago
- Reduction obtained in the Emergency Room
- Competitive volleyball player, recreational soccer player
- Imaging showing minimal Hill Sachs lesion / Anterior Glenoid Deficiency 20-25% / Presence of Bankart lesion

*What type of treatment would you choose, for the case mentioned above?
Choose one of the following answers

● Conservative

○ Surgical

You chose a conservative treatment. How appropriate do you think each of the following treatments is?

	Very appropriate	Appropriate	Inappropriate	Totally inappropriate	No answer
Brace in External rotation.	○	○	●	○	○
Brace in neutral position.	○	○	○	○	●
No specific immobilization. Suggestion to avoid activity at risk	○	●	○	○	○
Brace in Internal rotation.	○	○	●	○	○

How long would you immobilize the patient according to the most appropriate method you chose?
Choose one of the following answers

○ > 6 weeks

○ 4-6 weeks

● < 2 weeks

○ 2-4 weeks

○ No answer

Fig. 1 A screenshot from a page of the ESSKA survey.

4 The ESSKA Case Study

In order to validate our approach, we conceived a multi-page survey in which the members of a large community of specialists were invited to either choose their treatment of choice or rank alternative treatments in the context of eight fictitious clinical cases. These cases were conceived to be somehow “borderline” with respect to which treatment would be the best one (i.e., the most appropriate) to apply and they were described in terms of short summaries at the top of each survey page (see Figure 1 to see how a typical case was shown in each page of the online questionnaire). The initiative was co-designed and then patronaged by the scientific board of an international association of surgeons specialized in sports traumatology and knee surgery, ESSKA and this made it noteworthy for two main reasons. First, this association counts more than 1,000 members from 66 countries from all five continents; in that, it represents a heterogeneous gathering of professionals that, on one hand, share all the same interests and competencies in a specific medical specialty (i.e., sport traumatology) but that, on the other hand, have been trained and practice in quite different settings and environments. Second, surgery specialities are generally known to be the most conservative ones with respect to the adoption and inclusion in daily practice of evidences drawn from the academic literature [12]; for this reason, the scientific board of the association shared with us the research objective to see if lightweight Web-based tools, like online surveys are, could be leveraged to find potential agreement and achieve consensus on best practices in their specialty, even in those cases where no scientific evidence exist or is effectively applied by the grassroot level of their members.

At the end of the survey, we collected 374 completed questionnaires (36% of the target population), 38 partial questionnaires and 25 contacts, i.e., respondents that opened the first page of the survey but did not proceed. Almost three fourth of

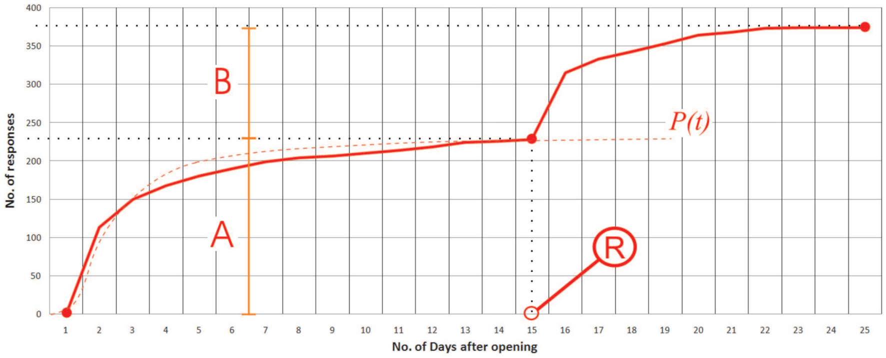
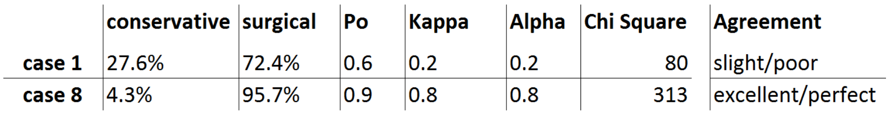


Fig. 2 The response rate of the survey

the respondents were from a European country. All together, we gathered 412 questionnaires where at least one case had been considered (i.e., a question addressed), accounting for a final response rate of 40% of the target population. This was in line with, or even beyond, the initial expectations, not only considering the really tight schedule of orthopaedic surgeons (57% of respondents declared to perform more than 200 operations yearly) and the nature of their work (that does not imply long stays at a computer), but also with respect to other survey initiatives involving orthopaedic surgeons [32] and populations of similarly great size (i.e., >1000 respondents) that are characterized by loose social control, high distributedness and lack of incentive structures [16]. Figure 2 depicts the particular response rate that we obtained: the dispatch of a reminder message after approximately two weeks since the opening of the collection session caused the increase of the 63% of responses as we will see in more details in Section 5.

The design of the study and the statistical results gathered through the online questionnaire have been described in another contribution [40]. In what follows, we will focus on the responses collected about two specific clinical cases that were considered of particular interest for the community due to their potential overlapping (or clash) with existing medical evidences of level I and II (i.e., the highest available for the community at hand). We present these cases in a row by presenting the concise description that was conveyed to the respondents and then discussing how these expressed their opinions and how their preferences can be matched with the existing literature in their field.

case 1 **21 y old Male; Dominant arm; First shoulder anterior dislocation 2 days ago; Reduction obtained in the Emergency Room; Competitive volleyball player; recreational soccer player; Imaging showing minimal Hill Sachs lesion / Anterior Glenoid Deficiency 20-25% / Presence of Bankart lesion**
In regard to whether a patient presenting a case like that should be operated or not (conservative vs. surgical treatment), the proportions of responses indicated a clear majority in favor of the surgical approach (72% vs. 28%, in Figure 3); that notwithstanding, the measures of agreement proposed in Section 2.1, gave



	conservative	surgical	Po	Kappa	Alpha	Chi Square	Agreement
case 1	27.6%	72.4%	0.6	0.2	0.2	80	slight/poor
case 8	4.3%	95.7%	0.9	0.8	0.8	313	excellent/perfect

Fig. 3 Response proportions and agreements for the selected cases.

results that are heuristically interpreted in terms of either slight or poor agreement (see the assessments of agreement in the rightmost column in Figure 3). This is an interesting point in favor of a more objective approach to the assessment of response proportions, as these can bring to intuitive results that highly overstate the actual agreement existing in communities (thus polarizing opinions and indications towards an unfounded consensus). If we focus on conservative options (the choice of the large majority of the respondents), we observe results by which we can assert that 'Brace in Internal rotation' and 'Brace in neutral position' are appropriate management techniques of similar cases with statistical significance. These two treatments are preferred with respect to the others (p=.003, 63% vs. 37%) but between these two, we can not say which one is the preferred one (p=.118); yet, "Brace in Internal rotation" was considered the most appropriate treatment on the 37% of cases (vs. 26%). This finding clashes with the lack of evidence from randomised controlled trials that a conservative management is more appropriate than others [17]. Moreover, differently from what asserted in a randomized controlled trial performed on 2007 [23], which asserts that external rotation is better (in terms of risk of recurrence) than the conventional method of immobilization in internal rotation, the community considers this treatment not appropriate (p<.001). To this regard, yet, respondents did not exhibit a significant agreement (P0, percentage of overall agreement 50%, Kappa score = .004). In regard to the optimal length of immobilization period for those who chose external rotation, we detected that the community expresses a clear preference (p=.025, 50 vs. 30, associated with the 2nd treatment) toward a length between 2 and 4 weeks. This finding complements the lack of evidence coming from a study of level II [42]

case 8 **69 y old Diffuse knee arthritis (all the compartments); Relative improvement from conservative treatments done elsewhere (VAS 8 to 5); Shows up again complaining for increase of daily pain. 75 Kg, 181 cm**

This is one of the cases in which the respondents reached the most clear consensus; obviously, they were unawares of each other's opinion and this marks the most notable difference between traditional ways to reach consensus on medical treatment, i.e., through focus groups where experts discuss cases and evidences at length till some agreement is reached. The survey indicated that most of the surgeons would go for a surgical treatment (96% vs. 4%) and this high polarization of preference is reflected by scores of almost perfect agreement. Also in regard to the kind of implant, respondents exhibited a clear preference (p=<.001) for the Cemented Total Knee implant. This confirms the inclusion in practice of corresponding evidences from literature [13]. Likewise, where there is a lack of evidence between Posterior Cruciate Retaining Implant and Posterior Stabilized Total Knee Prosthesis [24], respondents split quite evenly for either techniques (163 vs 171 ,p=.662).

From the cases outlined above, we make the point that where statistical significance is achieved, and agreement assessed, the system can provide practitioners with "evidences" that either confirm or oppose those drawn by scientific methods. On the basis of how much sound the findings are, the system could then pass these results to the discussion of the experts. These, in turn, could try to understand the status of the findings and address whether it is reasonable to rank their level of evidence somehow in between evidences drawn from case-series studies and those based on the opinions of few experts of clear and respected authority[6]. We leave it to the reader's opinion whether to deem the agreed opinions on very specific matters of a wide majority of skilled professionals more authoritative than the opinion of "the chosen few", where the agreement level is backed by an analytical survey system that is properly tailored to specific research questions.

5 A proposal to improve generalizability of findings

Threats to the (external) validity of a study like the reported one regard those elements that can negatively affect the confidence that the findings can be generalized from the sampling frame to the entire population of interest, and from the context of the study to other contexts, i.e., other people, communities, places and times. At the end of the survey, when we presented the preliminary results to the ESSKA board during their biannual congress (that was hold just one week after the conclusion of the study), we have been asked whether the opinions gathered from the respondents could be representative of the whole community of the ESSKA members. As we said in Section 2.1, generalizability is a major point to address for any study and consensus-oriented studies enabled by online surveys should make no exceptions.

In our case, a little less than half of the target population (i.e., all the ESSKA members) returned their questionnaire, thus making participation a partial success. We could not pick every actual respondent in a purely random manner, and this irreversibly thwarted the fact that the size of the sample would be sufficient for reliable results at a standard confidence level. Therefore, we had to consider the probability that non-response bias could affect the generalizability of the findings. In the present study, non-response bias refers to the condition in which the surgeons who did not fill in their questionnaires have opinions on the treatments of choice that are systematically different from the opinions of those who completed their surveys.

In order to test for non-response bias, then, we decided to consider who responded after the reminder had been sent as *proxies* for non-respondents; this is

[6] These are, respectively, evidences of level II-3 and III or level C and D according to the evidence ranking developed by either the U.S. Preventive Services Task Force or the Oxford Centre for Evidence-based Medicine.

a technique that has been proved adequate in regard to homogeneous groups of potential respondents, such as physicians in one specialty [44], in the assumption that the variations that do exist within such groups may not be as associated with willingness to respond. To this aim, we compared the responses of those who completed the questionnaire *before* the reminder (group A in Figure 2), to those who completed *after* it (group B in Figure 2), as the latter ones can be considered a sample of non-respondents (to the first mailing). Now, this comparison can be considered arbitrary in the general case, but it is not if one considers the cumulate response rate curve, we reported in Figure 2. As Barclay et al. noted almost incidentally, looking at when the curve seems to “flatten out” enables to decide on the right timing when to deploy a reminder to the completion of the questionnaire [4]. Our point is that the study of the best interpolating curve can not only be used to determine when it is a good time to send an effective reminder but, more significantly, to begin sampling the “catchment area” of the population of non-respondents. An assessment of the slope of the cumulative response rate curve allows to estimate whether the questionnaire will be completed by many other respondents or, on the contrary, very few others will remember to complete the questionnaire (and hence consider an invitation they received weeks or even months earlier). In our case, the pattern that we detected by minimizing the summation of the squared deviations between our data and a generic S-shaped logistic curve [37][7] suggests that very few respondents of the “second” turn would be just late comers of the first turn. The corresponding logistic function P(t) saturates after a certain value of x (i.e., number of hours/days) over a value that is not noticeably incremented not even for very high values of x, thus backing our approach to non-response bias[8]. Also in this case, we propose a heuristic to determine whether the response rate has reached a value of y (i.e., number of responses) that is very close to the hypothetical maximum; this value, indicated with R in Figure 2 corresponds when the first derivative of the logistic function[9] becomes less than 0.01. Once that point has been detected, investigators can send a reminder and begin considering responses collected from then on as representative of the rest of the population and consequently partition responses in two groups, namely early-respondents and late-respondents, respectively A and B in Figure 2, with group B taken as proxies of potential non respondents. Coming back to our case study, we performed a Pearson Chi test and a Wilcoxon-Mann-Whitney test on the variables of interest in these two groups, and detected no significant difference (Asymp. Sig.>.4); the reweighting of the results according to the response / non-response proportions produced a negligible impact on the observed values.

[7] The logistic regression model we obtained is represented by the function $y(x) = \frac{a}{b+ce^{-ax}}$ with $a \simeq 1.04$, $b \simeq 0.004$, $c \simeq 0.06$ and it is indicated with P(t) in Figure 2

[8] This holds in the assumption that responses from the second turn are representative of the non-responses, and that people that can get convinced by a single reminder end up by exhibiting similar opinions to those that, conversely, are refractory to any reminder at all

[9] $y' = \frac{abce^{cx}}{(b+e^{cx})^2}$

6 Conclusions and future work

Porter pointed out that the basis for authority in science must be *quantitative objectivity* as “reliance on nothing more than seasoned judgment seems undemocratic” [38, p.7]. We agree with this stance but we also came to wonder whether a fully “democratic judgment” can also be taken as reliable. To this aim, in this chapter we have addressed the general problem of how to reach “quantitative objectivity” in the analysis of survey-related responses and, at the same time, how lightweight Web-based technologies can support the “democratic” creation of (medical) knowledge on the basis of this objectivity. The methods we proposed and tested in our case study can be taken as a first contribution toward the consolidation of enabing technologies and methods to “give voice” to the usually silent moltitude of experts, the so called “grassroot level” of a community of practice. To validate our approach, we proposed and discussed the use an online survey system as a flexible and convenient means to collect the preferences and attitudes of doctors towards appropriate treatments in medical practice.

The doctors involved in the initiative responded favourably to the survey and to the specific tool we designed: some of them even contributed with suggestions on how to improve the tool, recognizing its potential especially for the younger and less expert surgeons. For this reason, we plan to further develop this idea by deploying an online survey that could adapt even more flexibly to the respondents’ specialization, experience and demands for the apt interpretation of the clinical cases proposed. The challenge posed by such a system is to find a way to offer to the users (i.e., to the doctors in the considered domain) pieces of information that are meaningful and “structured” enough to convey and evoke the experiential knowledge they need to make sense of a given context [7].

As a last remark, we believe that the potential for systems that are designed with the precise aim to collect collective preferences and soundly assess the agreement over these preferences is overly underestimated. We believe that further research should be undertaken to fully take advantages of their use in community-based domains. In fact, applications may range from the generic domains of distributed consensus conferences and distributed expert panels, to more specific ambits like, in medicine, the support of Delphi processes [25] and diagnosis team meetings [26]; in the academic domain, the support of the review process of novel research contributions for conference program committees and journal editorial boards; in the software engineering domain, the support of the phases of requirement elicitation, prioritization and validation. In particular to this latter domain, such systems would help analysts and designers probe the preferences and attitudes of a large number of potential users and stakeholders of a prospective system, thus contributing in the improvement of one of the most delicate phases in software development within the usually pressing time and cost constraints of ICT projects in cooperative settings.

References

1. Aday, L.: Designing and Conducting Health Surveys. Jossey-Bass (1996)
2. Alstyne, M.V., Brynjolfsson, E.: Global village or cyber-balkans? modeling and measuring the integration of electronic communities. Manage. Sci. **51**(6), 851–868 (2005)
3. Atkins, D., et al., M.E.: Systems for grading the quality of evidence and the strength of recommendations: Critical appraisal of existing approaches. BMC Health Services Research **4**(38) (2004)
4. Barclay, S., Todda, C., Finlayb, I., Grande, G., Wyattc, P.: Not another questionnaire! maximizing the response rate, predicting non-response and assessing non-response bias in postal questionnaire studies of gps. Family Practice **19**(1), 105–111 (2002)
5. Braithwaite, D., Emery, J., de Lusignana, S., Sutton, S.: Using the internet to conduct surveys of health professionals: a valid alternative? Family Practice **20**(5), 545–551 (2003)
6. Burns, K., Duffett, M., Kho, M., Meade, M.: A guide for the design and conduct of self-administered surveys of clinicians. Canadian Medical Association journal **179**(3), 245–52 (2008)
7. Cabitza, F., Simone, C., Sarini, M.: Knowledge artifacts as bridges between theory and practice: the clinical pathway case. In: KMIA'08: Proceedings of the International Conference on Knowledge Management In Action. Held in conjunction with the 20th IFIP World Computer Congress, 7 Jan 08 Milan, Italy (2008)
8. Dillman, D.: Mail and Internet Surveys: The Tailored Design Method. Wiley; 2 edition (2000)
9. Dobrev, A., Haesner, M., Hsing, T., Korte, W., Meyer, I.: Benchmarking ict use among general practitioners in europe. Tech. rep., European Commission - Information Society And Media Directorate General (2008)
10. Feinstein, A.R., Cicchetti, D.V.: High agreement but low kappa: I. the problems of two paradoxes. Journal of Clinical Epidemiology **43**(6), 543–549 (1990).
11. Feinstein, A.R., Cicchetti, D.V.: High agreement but low kappa: I. the problems of two paradoxes. Journal of Clinical Epidemiology **43**(6), 543 – 549 (1990)
12. Ferlie, E., Wood, M., Fitzgerald, L.: Some limits to evidence-based medicine: a case study from elective orthopaedics. Quality in Health Care **8**, 99107 (1999)
13. Gandhi, R., Tsvetkov, D., Davey, J., Mahomed, N.: Survival and clinical function of cemented and uncemented prostheses in total knee replacement. Journal of Bone and Joint Surgery **91-B**(7), 889–895 (2009)
14. Grava-Gubins, I., Scott, S.: Effects of various methodologic strategies. Can Fam Physician. **54**(10), 1424–1430 (2008)
15. Gwet, K.: Handbook of Inter-Rater Reliability. STATAXIS Publishing Company (2001)
16. Hamilton, M.B.: Online survey response rates and times. background and guidance for industry. Tech. rep. (2009)
17. Handoll, H., Hanchard, N., Goodchild, L., Feary, J.: Conservative management following closed reduction of traumatic anterior dislocation of the shoulder. Cochrane Database Syst Rev. **25**(1) (2006)
18. Hayes, A.: Statistical methods for communication science. Mahwah, NJ: Lawrence Erlbaum Associates, Inc. (2005)
19. Hayes, A.F., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. Communication Methods and Measures **1**(1), 77–89 (2007)
20. Headrick, D.R.: When information came of age : technologies of knowledge in the age of reason and revolution, 1700-1850. Oxford University Press (2000)
21. Heide, L.: Punched-Card Systems and the Early Information Explosion, 1880-1945. Johns Hopkins University Press (2009)
22. Hiltz, S., Turoff, M.: Network Nation - Revised Edition: Human Communication via Computer. The MIT Press (1993)
23. Itoi, E., et al., Y.H.: Immobilization in external rotation after shoulder dislocation reduces the risk of recurrence. a randomized controlled trial. J Bone Joint Surg Am. **89**(10), 2124–31 (2007)

24. Jacobs, W., Clement, D., Wymenga, A.: Retention versus sacrifice of the posterior cruciate ligament in total knee replacement for treatment of osteoarthritis and rheumatoid arthritis. Cochrane Database Syst Rev. **19**(4) (2005)
25. Jones, J., Hunter, D.: Consensus methods for medical and health services research. BMJ **311**, 376–380 (1995)
26. Kane, B., Luz, S.: Achieving diagnosis by consensus. CSCW **18**(4), 357–392 (2009)
27. Kapetanios, E.: Quo vadis computer science: From turing to personal computer, personal content and collective intelligence. Data Knowl. Eng. **67**(2), 286–292 (2008)
28. Kellerman, S., Herold, J.: Physician response to surveys. a review of the literature. Am J Prev Med. **20**(1), 61–7 (2001)
29. Kongsved, S., Basnov, M., Holm-Christensen, K., Hjollund, N.: Response rate and completeness of questionnaires: a randomized study of internet versus paper-and-pencil versions. J Med Internet Res. **9**(3) (2007)
30. Krippendorff, K.: Content analysis: An introduction to its methodology. Sage, Thousand Oaks, CA, USA (2004)
31. Landis, J., Koch, G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977)
32. Leece, P., Bhandari, M., Sprague, S., Swiontkowski, M.: Internet versus mailed questionnaires: A controlled comparison. J Med Internet Res. **6**(4) (2004)
33. Lipton, R., Liberman, J., Cutrer, F., Goadsby, P.: Treatment preferences and the selection of acute migraine medications: results from a population-based survey. The Journal of Headache and Pain **5**(2), 123–130 (2004)
34. Marx, R.G., et al.: Beliefs and attitudes of members of the american academy of orthopaedic surgeons regarding the treatment of anterior cruciate ligament injury. Arthroscopy: The Journal of Arthroscopic & Related Surgery **19**(7), 762–770 (2003).
35. McColl, E., Jacoby, A., Thomas, L., Soutter, J.: Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. Health Technol Assess. **5**(31), 1–256 (2001)
36. Mosca, L., et al.: National study of physician awareness and adherence to cardiovascular disease prevention guidelines. Circulation **111**, 499–510 (2005)
37. Parasuraman, A.: More on the prediction of mail survey response rates. Journal of Marketing Research **19**(2), 261–268 (1982)
38. Porter, T.M.: Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton University Press (1996)
39. Potter, W., Levine-Donnerstein: Rethinking validity and reliability in content analysis. Journal of Applied Communication Research **27**, 258–284 (1999)
40. Randelli, P., Cabitza, F., Arrigoni, P., Cabitza, P., Ragone, V.: Current practice in shoulder pathology: Results of a Web-Based survey among a community of 1,084 orthopedic surgeons. Knee Surgery, Sports Traumatology, Arthroscopy **forthcoming** (2011).
41. Randolph, J.J.: Free-marginal multirater kappa: An alternative to fleiss' fixed-marginal multirater kappa. In: Proceedings of the Joensuu University Learning and Instruction Symposium 2005, Joensuu, Finland, October 14-15th (2005)
42. Scheibel, M., Kuke, A., et al., C.N.: How long should acute anterior dislocations of the shoulder be immobilized in external rotation? Am J Sports Med. **37**(7), 1309–16 (2009)
43. Seguin, R., Godwin, M., MacDonald, S., McCall, M.: E-mail or snail mail? randomized controlled trial on which works better for surveys. Can Fam Physician. **50**(5), 414–9 (2004)
44. Sobal, J., Ferentz, K.: Assessing sample representativeness in surveys of physicians. Eval Health Professions. **13**, 367–372 (1990)
45. Thorpe, C., Ryan, B., Burt, A., Stewart, M., Brown, J.: How to obtain excellent response rates when surveying physicians. Fam Pract. **26**(1), 65–8 (2009)
46. Van De Belt, T.H., Engelen, L.J., Berben, S.A., Schoonhoven, L.: Definition of health 2.0 and medicine 2.0: A systematic review. Journal of Medical Internet Research **12**(2) (2010).
47. VanGeest, J., Johnson, T., Welch, V.: Methodologies for improving response rates in surveys of physicians: a systematic review. Eval Health Prof. **30**(4), 303–21 (2007)
48. Webster, B., et al., T.C.: Physicians' initial management of acute low back pain versus evidence-based guidelines. influence of sciatica. Journal of General Internal Medicine **20**, 1132–1135 (2005)