Validations of research methods for urinary incontinence in women

Scored questionnaires in clinical practice and epidemiological research

Atle Klovning

Dissertation for the degree of philosophiae doctor (PhD)

at the University of Bergen, Norway.

2010

ISBN 978-82-308-1590-8

Validations of research methods for urinary incontinence

Scored questionnaires in clinical practice and epidemiological research

Atle Klovning

Department of Public Health and Primary Health Care

Faculty of Medicine and Dentistry

University of Bergen, Norway

2010

Contents

Acknowledgements

26.2 miles. Better known as 42,195 m. After completing the scientific marathon a PhD is, I have so many wonderful people to thank!

First of all, I owe an endless number of thanks to my principal supervisor, Steinar Hunskaar at the Department of Public Health and Primary Health Care at the University of Bergen for his continual inspiration and kind support in all phases of this PhD project; uphill and downhill, mile after mile! His thoroughness and profound knowledge of every aspect of this research field has facilitated my completion of this PhD-thesis, which at times felt like a "mission impossible,"- now hopefully converted into a "mission accomplished"- the end of a long run!

I was originally tracked into **Study I** while strolling along the cobbled streets of Bergen along with my bike with Steinar Hunskaar an afternoon after work as a young GP at Engen Legesenter in Bergen in August 1993. He needed "someone to just enter some data into SPSS 4.0, to analyze these data, and perhaps even work part-time as a research assistant at the University…" I accepted his challenge, and learned the early SPSS with its initial scripting procedures, working on my Mac PowerBook 145 (1993 model), which still functions!

My eminent co-supervisor since the second part of this PhD has been Hogne Sandvik, keeping me at pace and on track with his discrete and elegant support since the preparatory phases of **Study II**. Hogne deserves a "Guinness book of records"-prize for his fabulous ability to give the "swiftest and most thorough feedback in the World."

I am utterly grateful for the time and patience Steinar and Hogne have spent guiding a more "villedbar enn veiledbar" (distractible than possible to guide) person like me!

Many thanks to Bjarne Eriksen, who kindly let us research the high quality data he and colleagues collected for **Study I** at the Gynaecological department at The Norwegian University of Science and Technology in Trondheim, and to Anita Vanvik, who collected the urotherapist data.

I have spent many wonderful years with my colleagues at the University of Bergen since 1993, who provided an intellectually stimulating environment, letting me explore the fascinating tools of evidence-based medicine in my teaching and in all my side tracks on my way to this PhD-thesis. For me, EBM has since then been a way of clearing up medical clutter, and Steinar Hunskaar was the one who opened the doors for me. Thanks! And thanks to Trisha Greenhalgh for letting me tutor at her EBM workshops, Andy Oxman at the Nordic workshops, Toby Lipman at the Durham North-England workshops, and the late Udo Kastner at an Austrian one! I have enjoyed globetrotting with EBM!

Along the way, my fascination for EBM has resulted in publications and Web sites with great colleagues: on alendronate (with Ole Frithjof Norheim), several books and book chapters on EBM (Arild Bjørndal and Signe Flottorp), meta-analyses on osteoarthritis of the knee (Jan Magnus Bjordal, Lars Slørdal and the late Elisabeth Ljunggren), on national guidelines for antenatal care for the Directorate of Health and Social Welfare (Brit Roland), on evidence-based laboratory medicine (EBLM) and American guidelines for point-of-care testing for diabetes (Sverre Sandberg); and several Web sites like www.NettDoktor.no, www.cmwr.org and www.medicalwebresearch.net, www.forskningsenheten.no and www.frognerhelsesenter.no.

In 2001, after NettDoktor, a virtual Centre for Medical Web Research (CMWR) was launched with the invaluable help from Finn Steen and Jarle Petterson. In that phase, I once again knocked on the doors of Steinar Hunskaar and Hogne Sandvik, telling them I would like to take up again UI research, but this time to explore Web-based research of UI. This led

to the next phase of my UI research, **Study II** in this thesis, where I also had the pleasure of working with Kerry Avery, the key developer of the ICIQ-UI SF, in validating and scaling this ICI recommended instrument.

During nearly all these years, I have greatly appreciated being able to combine my research with clinical work as a GP- at Ulriksdal Legesenter, Olsvik Legesenter and now at Frogner Helsesenter thanks to Kjell-Olav Svendsen and Nils Johnsen. And I thank all my patient patients for reminding me every day of how the life of a real GP is like in the non-virtual World outside the University! I love being a GP for real people, too!

Special thanks to Jørund Straand, Head of General Practice Research Unit at the University of Oslo, for providing superb facilities and a stimulating research environment for my future life after moving to Oslo in 2006, and for opening the doors to forthcoming projects! We have been friends and colleagues working together for many years as GPs and researchers in Bergen, and we both enjoy singing Bellman songs.

I also thank Head of Section for General Practice/Family Medicine at the University of Oslo, Per Hjortdahl, for gentle and motivating support not only in my previous projects, but also for joining running projects in Cyberspace!

Although all my "diversions" have in some way represented obstacles to finalising this PhD, I would not have been without any one of them! I am deeply indebted to all my collaborators and friends who have stood cheering alongside this marathon track, and for giving me the privilege to pursue my research dreams!

Finn Steen has followed me through the past 30 years, all my ups and downs, and has at times been a co-worker at our virtual reality site cmwr.org. I thank him for all the great work he has accomplished on our Web projects, his frankness and ability to deromanticise my prolixity. I also am eternally grateful for the support Egil Lundal has given me over the

past 25 years- he has at times been the only person able to give me a notion of actually understanding what I really thought I meant and felt!

Thanks to all the great moments at the "Selskap for rasjonelle valg" ("Society for rational choices") with Ole Frithjof Norheim, Jon Ketil Johnsen and Geir Brekke- thanks for great discussions and wonderful cookery over the years – and for virtual running!

I owe my families and many friends unlimited gratitude for putting up with me through all these projects and all my misdemeanours, and who through the different phases of my research have reminded me that life is more than work and running. Although my life has taken on many directions, and many challenges have had to be solved on the way, I deeply honour and cherish all the wonderful moments I have encountered along the winding roads with families and friends!

I thank my dear children, Daniel, Theodor and Sofie for being just the wonderful persons they are! I was once upon a time a curious child, looking forward to yet new, exciting projects, just the way that they now are looking forward to finding their respective ways!

I am ever so grateful to my dear Bippi for her eternal inspirational cheerfulness, her ability to assist me in getting back on track when led astray, for the endurance that she has shown in the final phases of this process, and all the joyful moments and good company on the way; and for my bonus family- Simen, Sofie and Frida!

Finally, thanks to Apple and Nike+ for creating that wonderful virtual world of running with iPods and 187g LunarRacers! My next goal is running marathons in beautiful cities around the World after attending conferences. There will definitely be more time for family, friends, culture and singing. I'm so happy to have discovered the joys of long distance running! After passing the PhD finish line, life will never be the same!

Oslo, 2nd September 2010,

Atle Klovning

Abstract

Validations of research methods for urinary incontinence.

Scored questionnaires in clinical practice and epidemiological research.

Atle Klovning (Dissertation for the degree of philosophiae doctor, PhD).

Many clinicians and researchers have claimed that too many questionnaires for diagnosing and assessing urinary incontinence have been developed, and that the time has come for recommending a subset of validated scored questionnaires to be used in clinical practice and research. Our research group has over the past 25 years applied different strategies for validating scored questionnaires to be used in clinical practice and epidemiological research on urinary incontinence (UI).

This PhD-thesis is based on the findings in two studies: one clinical, diagnostic study, **Study I,** with findings published in **Paper I**, and one Web-based epidemiology study, **Study II**, with findings published in **Paper II** and **Paper III**.

In **Study I** we validated a scored questionnaire, the Detrusor Instability Score (DIS). The study patients were first assessed by a urotherapist (a specialised nurse) by structured history taking and preliminary tests, prior to the consultation with a gynaecologist. The urotherapist used a non-validated, standardized multiple-choice form for obtaining the urological history, specially devised for urogynaecological problems. The DIS was embedded in this questionnaire, and was administered by the urotherapist in a clinical, gynaecological outpatient setting. The DIS is a validated scored questionnaire developed to detect detrusor instability, and consists of 10 items, to be scored between 0 and a maximum of 20 points. The resulting score was not calculated before the study was over, thus blinding the urotherapist and the gynaecologist to the DIS. Altogether 250 patients were included.

Urine samples were examined, and cultured if infected. The patients filled in frequency/volume-charts and performed pad-weighing tests at home, so that the results could be presented to the gynaecologist at the next consultation. The need for incontinence aids was also determined.

The gynaecologist's consultation consisted of the medical history, urogynaecological examination including the assessment of prolapse/atrophy, perinealneurological examination, measurement of the residual urine volume, palpation of the pelvic floor, including assessment of the active contraction ability, stress-test, urodynamic investigations, and urethrocystoscopy on special indications. The gynaecologist recorded two sets of diagnoses for research purpose, A: Urodynamic diagnosis, and B: Clinical diagnosis after a comprehensive assessment of all available data except the DIS. This comprehensive clinical assessment was defined to be the gold standard for diagnosing genuine stress incontinence (GSI).

In **Study I** we found that the originally proposed cut-off level at 7 for the DIS resulted in patients with too many false positive findings for us to consider it to be useful as a preoperative tool. In 159 of the 250 women (64%) having GSI as defined diagnostically by a cut-off level at 7 for the DIS, we found that 41 of 250 women (16%) were actually given a false positive diagnosis. This could have been acceptable for conservative (non-surgical) treatments in primary health care settings, but not for surgical treatment. However, if the cut-off level was lowered to a cut-off level at 5 for the DIS, we found that 112 of 250 women (45%) would be diagnosed as having GSI, with only 20 of 250 women (8%) having a false positive diagnosis. The important issue here was whether these women, if otherwise feasible and indicated, might be able to undergo continence surgery without preoperative urodynamics. Consequently, we concluded that a lower cut-off point than originally

proposed was needed for the DIS to become a useful preoperative tool for continence surgery.

In **Study II**, we had two aims: To analyse how Web-based recruitment performed compared with postal surveys (**Paper II**), and to validate the International Consultation on Incontinence Questionnaire-Urinary Incontinence Short Form (ICIQ-UI SF) against the Incontinence Severity Index (ISI), in order to construct a severity grading for the ICIQ-UI SF (**Paper III**). We used the Web to invite a convenience sample of women to join a women's health study by self-selected participation (n=1,812) in 2002. The study was performed using the software Inquisite. Female users of major Norwegian Web sites were asked to join the study by three different routes: a general health Web site (NettDoktor.no), the health section of a general-purpose Web portal (StartSiden.no), and the newspaper Web site of Verdens Gang (VG.no). By answering "Yes" to a question defining the respondent as having UI, the respondents were branched into two validated questionnaires, the ISI as items number 2 and 3 of the EPINCONT questionnaire on web page 4 (10 items), and the ICIQ-UI SF on web page 5 (four items).

In **Study II** (**Paper II**), the results of 1,812 Web-recruited respondents were compared with 27,936 postally recruited study subjects, using the same epidemiological questionnaire to study UI as used in the EPINCONT study (Epidemiology of Incontinence in the County of Nord-Trøndelag).

Comparative analysis of results of the corresponding variables used in the WEB-EPI UI and the EPINCONT studies was done by calculating the 95% CIs with the CIA software,[1] using the Newcombe method for comparing independent proportions. Single asterisks (*) were placed in Table 1 of **Paper II** to mark where the point estimate of one variable was not an element of the 95% CI of the corresponding variable, thus indicating a statistically significant difference. Double asterisks (**) were placed to mark where the 95%

CI for the difference between the independent proportions did not contain zero, indicating a statistically significant difference.

The data were also analyzed as 5-year age groups, and for any significant difference for variables between the three different websites; none were detected, and the sample was analyzed as a whole. Statistical significance was accepted at the 5% level ($P<0.05$).

We found that our Web sample of women with UI was younger than the EPINCONT sample, 37 versus 48 years, $P<0.05$. The proportion of women 60 years or older was only 3.3% in our study, while it was 29.0% in the EPINCONT study. We found that the unadjusted prevalence of UI was lower in our study (20%) than in the EPINCONT study (25%), but age-stratified prevalence rates were higher in all age groups. In the Web sample, we found fewer women with slight UI in all age groups, and more with moderate (30-39 and 50-59-year age groups) and severe UI (20-29, 30-39, and 40-49-year age groups).[1] We concluded that we recruited a younger population with more severe UI than the EPINCONT study. Web-based approaches seem to be less feasible than postal methods for studies on conditions with higher prevalence in the older population, and UI is such a condition.

In **Study II** (**Paper III)** we also performed a Web-based comparison of two scored questionnaires assessing the severity of UI, (the ICIQ-UI SF vs. the ISI), using the ISI as a gold standard.

The ICIQ-UI SF has been developed by the International Consultation on Incontinence (ICI), has so far been translated to 38 languages, and is now recommended by the ICI as a gold standard outcome measure for future research and clinical practice, according to the proceedings of the 4th ICI (pp. 1771-2).[2] The committee recommends using

[1] Erratum: In **Paper II**, the groups with severe UI were erroneously reported as 30-39, 40-49 and 50-59.

high quality questionnaires (Grade A) for the assessment of the patient's perspective of incontinence symptoms and their impact on quality of life, and recommend other Grade A questionnaires for more detailed assessment.

We used split-half sampling for developing and validating the severity grading of the ICIQ-UI SF, using SPSS to extract a random half of the 343 women with UI, yielding a development sample (n=171) and a validation sample (n=172). The respondents in the first sample were used to develop the scale for the ICIQ-UI SF, while the remaining respondent sample was used to validate the severity scaling of the ICIQ-UI SF. Four levels of the ISI were plotted against the ICIQ-UI SF sum-score with and without the HRQoL dimension. The association between the ISI and ICIQ-UI SF scores was investigated by Spearman's rank correlation coefficient (rho), as this correlation is used for ordinal variables. Kappa values were calculated using the SPSS on 4x4 contingency tables of the severity (slight, moderate, severe, very severe) of UI by arbitrarily changing the severity intervals until maximum Kappa was obtained. As SPSS was only able to produce unweighted Kappa statistics, the 4x4 contingency tables with maximum unweighted Kappa values produced by SPSS were manually entered into the Web pages provided by Professor emeritus Lowry, enabling us to calculate Kappa scores with linear and quadratic weighting. In order to create a scale for the ICIQ-UI SF based on the ISI as the assumed gold standard, we iteratively calculated the weighted Kappas for the unweighted Kappas that SPSS produced for the different intervals for the severity of the ICIQ-UI SF and the ISI. Accordingly, the weighted Kappas were calculated for the validation sample.

There were strong correlations between the four-level ISI and ICIQ-UI SF scores with versus without the HRQoL item; Spearman's rho was 0.62, $P < 0.01$ versus 0.71, $P < 0.01$. By adjusting the intervals for the ICIQ-UI SF total score for the study subjects in the first scale development file to obtain maximum agreement with the four levels of the ISI, we

could define the following intervals for the ICIQ-UI SF (n = 171): slight (1-5), moderate (6-12), severe (13-18), and very severe (19-21) (Kappa with quadratic weighting = 0.61). Similarly, for the ICIQ-UI SF without the HRQoL item, we could define the following levels: slight (1-3), moderate (4-5), severe (6-9), and very severe (10-11), (Kappa with quadratic weighting = 0.71). Applying these intervals to the second sample (n = 172) in order to validate our findings, Kappa with quadratic weighting for ICIQ-UI SF with and without the HRQoL item was 0.61 and 0.74, respectively.

Our findings suggest that the ICIQ-UI SF may be divided into the following four severity categories: slight (1-5), moderate (6-12), severe (13-18) and very severe (19-21) UI. Disregarding the HRQoL-item, the four severity grades would be slight (1-3), moderate (4-5), severe (6-9) and very severe (10-11).

Sammendrag (Abstract in Norwegian)

Valideringer av forskningsmetoder for urininkontinens.

Skårede spørreskjemaer i klinisk praksis og epidemiologisk forskning.

Atle Klovning (Avhandling for graden philosophiae doctor, ph.d.).

Mange klinikere og forskere har hevdet at det har vært utviklet for mange ulike spørreskjemaer for diagnostisering og vurdering av effekten av behandling av urininkontinens (UI), og at tiden er moden for å anbefale et avgrenset sett av validerte skårede spørreskjemaer. I løpet av de siste 25 årene har forskningsgruppen vår anvendt ulike strategier for å validere skårede spørreskjemaer til bruk i praksis og epidemiologisk forskning på UI.

Denne ph.d.-avhandlingen bygger på resultatene fra to studier; en klinisk diagnostikkstudie, **Studie I**, med funn publisert i **Artikkel I**, og en Web-basert analytisk epidemiologistudie, **Studie II**, med funn publisert i **Artikkel II** og **Artikkel III**.

I **Studie I** validerte vi et spørreskjema, Detrusor Instability Score (DIS). Pasientene ble først vurdert av en uroterapeut (en spesialisert sykepleier) ved hjelp av et ikke-validert, strukturert anamneseskjema og med innledende tester, forut for konsultasjonen med en gynekolog. Uroterapeuten brukte et standardisert flervalgsskjema for å ta opp den urologiske anamnesen, spesielt utviklet for urogynekologiske problemer. DIS var bygget inn i dette spørreskjemaet som ble administrert av en uroterapeut på en gynekologisk poliklinikk. DIS er et validert, skåret spørreskjema utviklet for å detektere detrusor instabilitet, og består av 10 elementer, og som skal skåres 0-20 poeng. Poengsummen ble ikke beregnet før studien var over, og dermed var både gynekologen og uroterapeuten blindet for DIS. Til sammen 250 pasienter ble inkludert.

Urinprøver ble undersøkt, og dyrket dersom infisert. Pasientene fylte ut miksjonsliste og utførte bleieveiingstester hjemme, slik at resultatene kunne forelegges gynekologen ved neste konsultasjon. Behovet for inkontinenshjelpemidler ble også vurdert.

Gynekologens konsultasjon omfattet anamneseopptak, urogynekologisk undersøkelse inkludert vurdering av prolaps/atrofi, perinealneurologisk undersøkelse, måling av resturin, palpasjon av bekkenbunnen, inkludert vurdering av bekkenbunnsmuskulatur, stresstest, urodynamiske undersøkelser, og urethracystoskopi på spesielle indikasjoner. Gynekologen registrerte to sett med diagnoser for forskningsformål, A: Urodynamisk diagnose og B: Klinisk diagnose etter en helhetlig vurdering av alle tilgjengelige data bortsett fra DIS. Denne helhetlige, kliniske vurderingen ble definert å være gullstandarden for diagnostisering av genuin stressinkontinens (GSI).

I **Studie I** fant vi at den opprinnelig foreslåtte avskjæringsverdien ved 7 for DIS førte til for mange falske positive funn til vi kunne anse DIS å være nyttig som et preoperativt verktøy. Hos 159 av 250 kvinner (64%) som hadde GSI definert ved en avskjæringsverdi på 7 for DIS, fant vi at 41 av 250 kvinner (16%) faktisk fikk en falsk positiv diagnose. Dette kunne tenkes å være akseptabelt for konservativ (ikke-kirurgisk) behandling i primærhelsetjenesten, men ikke for kirurgisk behandling. På den andre siden, dersom avskjæringsverdien for DIS ble senket til 5, ville det resultert i at 112 av 250 kvinner (45%) fikk definert diagnosen GSI, hvorav 20 av 250 kvinner (8%) fikk en falsk positiv diagnose. Det sentrale spørsmålet her var om disse kvinnene, dersom det ellers var indisert, kunne tilbys kontinenskirurgi uten preoperativ urodynamisk undersøkelse. Som en følge av våre beregninger, fant vi at et lavere avskjæringspunkt enn det opprinnelig foreslåtte var nødvendig om DIS skulle kunne være et nyttig verktøy for preoperativ vurdering før kontinenskirurgi.

I **Studie II**, hadde vi to mål: å analysere hvordan Web-basert rekruttering funksjonerte sammenlignet med brevbaserte undersøkelser (**Artikkel II**), samt å validere spørreskjemaet International Consultation on Incontinence Questionnaire-Urinary Incontinence Short Form (ICIQ-UI SF) mot Incontinence Severity Index (ISI), for deretter å utarbeide en alvorlighetsgradering av ICIQ-UI SF (**Artikkel III**). Vi brukte WWW til å invitere et selvselektert utvalg kvinner (n=1812) i 2002 til å delta i en kvinnehelseundersøkelse. Studien ble utført ved hjelp av programvaren Inquisite. Kvinnelige brukere av store norske Websteder ble rekruttert via tre forskjellige ruter: et generelt helsewebsted (NettDoktor.no), helseseksjonen av webportal (StartSiden.no), og nyhetsweben VG.no. Ved å svare bekreftende på et spørsmål som definerte at respondenten hadde UI, ble respondentene forgrenet til to validerte spørreskjemaer, ISI som spørsmål 2 og 3 i EPINCONT på web side 4 (10 spørsmål), og ICIQ-UI SF på web side 5 (fire spørsmål).

I **Studie II** (**Artikkel II**) ble resultatene fra 1 812 Webrekrutterte respondenter sammenlignet med 27 936 brevrekrutterte respondenter, med det samme epidemiologiske spørreskjemaet som ble brukt i EPINCONT studien (Epidemiology of Incontinence in the County of Nord-Trøndelag).

Alle konfidensintervallene ble beregnet etter Newcombes metode for å sammenligne uavhengige proporsjoner ved hjelp av et MS DOS-basert program, CIA.[1] Alle konfidensintervallene ble beregnet én og én, og tilordnet stjerner etter sammenligning av konfidensintervallene for EPINCONT og WEB-EPI UI studiene. En enkeltstjerne (*) markerte i Tabell 1 i **Artikkel II** de tilfellene hvor punktestimatet av en variabel ikke var delmengde av 95% konfidensintervallet til den korresponderende variabelen, og således indikerte en statistisk signifikant forskjell. Dobbeltstjerner (**) markerte de tilfellene hvor 95% konfidensintervallet for forskjellen mellom uavhengige proporsjoner ikke inneholdt null, og dermed indikerte en statistisk signifikant forskjell.

Dataene ble også analysert som 5-års aldersgrupper, og for statistisk signifikante forskjeller mellom deltagerne fra de tre forskjellige webstedene; ingen forskjeller ble avdekket, slik at utvalget ble analysert som et hele. Statistisk signifikans ble akseptert på 5% nivået ($P<0,05$).

Webrespondentene var yngre enn EPINCONT-respondentene, 37 versus 48 år, $P<0,05$. Andelen kvinner 60 år eller eldre var bare 3,3% i vår undersøkelse, mot 29,0% i EPINCONT studien. Vi fant at den ujusterte råprevalensen av kvinner med UI var lavere i vår studie (20%) enn i EPINCONT studien (25%), mens den aldersstratifiserte prevalensen var høyere i de enkelte aldersgruppene. I Web-gruppen fant vi færre kvinner med mild UI i alle aldersgrupper, og flere med moderat (30-39 og 50-59-års aldersgrupper) og alvorlig UI (20-29, 30-39 og 40-49-års aldersgrupper).[2] Vi konkluderte med at vi rekrutterte en yngre populasjon av kvinner med mer alvorlig UI enn EPINCONT studien. Web-baserte tilnærminger synes å være mindre hensiktsmessige enn postale metoder for studier av tilstander med høyere prevalens i den eldre delen av befolkningen, og UI er en slik tilstand.

I **Studie II** (**Artikkel III)** gjennomførte vi også en Web-basert sammenligning av to skårede spørreskjemaer som vurderer alvorlighetsgraden av UI, ICIQ-UI SF og ISI, hvor vi anvendte ISI som gullstandarden.

ICIQ-UI SF er utviklet av The International Consultation on Incontinence (ICI), og har så langt blitt oversatt til 38 språk, og er nå anbefalt av ICI som gullstandarden for fremtidig forskning og klinisk praksis, ifølge rapporten fra den 4de ICI (side 1771-2).[2] Komiteen anbefaler bruken av høykvalitets spørreskjemaer (Grad A) for vurdering av

[2] Erratum: I Artikkel II er gruppen med ”severe UI” feilaktig beskrevet som 30-39, 40-49 og 50-59.

pasientens egne syn på urininkontinenssymptomene og deres betydning for livskvaliteten, og anbefaler andre Grad A spørreskjemaer for mer detaljerte vurderinger.

Vi anvendte splittmetodikk for å lage et utviklings- og et valideringsutvalg for alvorlighetsgradering av ICIQ-UI SF, ved å bruke SPSS til å splitte utvalget bestående av 343 kvinner med UI i to tilfeldige halvdeler, slik at vi fikk et utviklingsutvalg på 171 og et valideringsutvalg på 172 respondenter. Respondentene i det første utvalget ble brukt til å utvikle graderingsskalaen for ICIQ-UI SF, mens den gjenværende halvdelen ble brukt til å validere denne alvorlighetsskalaen. Fire nivåer av ISI ble plottet mot ICIQ-UI SF skåren med og uten livskvalitetsdimensjonen. Assosiasjonen mellom ISI og ICIQ-UI SF ble undersøkt ved hjelp av Spearman's rank correlation coefficient (rho), ettersom metoden brukes for ordinale variabler. Kappaverdier ble beregnet ved hjelp av SPSS for 4x4-tabeller av alvorlighetsgrader (mild, moderat, alvorlig og svært alvorlig) av UI ved å systematisk endre alvorlighetsintervallene inntil maksimal Kappaverdi ble funnet. Siden SPSS bare var i stand til å produsere uvektet Kappastatistikk, ble de 4x4-tabellene med maksimal uvektet Kappa manuelt lagt inn i et webbasert program utviklet av professor emeritus Lowry, slik at vi kunne regne ut Kappaverdier med bade lineær og kvadratisk vekting. For å skalere ICIQ-UI SF basert på ISI som den antatte gullstandarden, beregnet vi gjentatte ganger vektet Kappa for de uvektede 4x4-tabellene som SPSS produserte for de forskjellige alvorlighetsintervallene av ICIQ-UI SF og ISI. Til slutt ble vektede Kappaverdier for valideringsutvalget beregnet.

Vi fant høy korrelasjon mellom fire-nivå ISI og ICIQ-UI SF skårene med versus uten livskvalitetsdelen, Spearmans rho var 0,62 ($P < 0,01$) versus 0,71 ($P < 0,01$). Ved å justere intervallene for ICIQ-UI SF totalskår for forsøkspersonene i utviklingsfilen for å oppnå maksimum overensstemmelse med de fire gradene av ISI, fant vi følgende intervaller for ICIQ-UI SF (n = 171): mild (1-5), moderat (6-12), alvorlig (13-18), og svært alvorlig (19-21)

(Kappa med kvadratisk vekting = 0,61). Tilsvarende, for ICIQ-UI SF uten livskvalitetsdelen, fant vi følgende nivåer: mild (1-3), moderat (4-5), alvorlig (6-9), og svært alvorlig (10-11), (Kappa med kvadratisk vekting = 0,71). Ved å anvende disse intervallene på valideringsfilen vår (n = 172), fant vi at Kappa med kvadratisk vekting for ICIQ-UI SF med og uten livskvalitetsdelen var henholdsvis 0,61 og 0,74.

Våre funn viste at total ICIQ-UI SF kan deles inn i følgende fire alvorlighetsgrader: mild (1-5), moderat (6-12), alvorlig (13-18) og svært alvorlig (19-21) UI, og dersom vi ikke tar med livskvalitetsdelen, blir graderingen mild (1-3), moderat (4-5), alvorlig (6-9) og svært alvorlig (10-11).

List of publications

Paper I. Klovning A, Hunskaar S, Eriksen BC. Validity of a scored urological history in detecting detrusor instability in female urinary incontinence. *Acta Obstet Gynecol Scand* 1996;75:941-5.

Paper II. Klovning A, Sandvik H, Hunskaar S. Web-based survey attracted age-biased sample with more severe illness than paper-based survey. *J Clin Epidemiol* 2009;62:1068-74.

Paper III. Klovning A, Avery K, Sandvik H, Hunskaar S. Comparison of two questionnaires for assessing the severity of urinary incontinence: The ICIQ-UI SF versus the Incontinence Severity Index. *Neurourol Urodyn* 2009;28:411-5.

The papers are reproduced with permission from the respective publisher for academic use, and are referred to by their Roman numerals, as **Paper I**, **Paper II** and **Paper III** in the thesis.

Abbreviations and definitions

Accuracy	Rate of true positives and true negatives
AUC	Area under the curve
BMI	Body mass index
CI	Confidence interval
DARE	Database of Abstracts of Reviews of Effects
DI	Detrusor instability (outdated). Idiopathic detrusor overactivity
DIS	Detrusor instability score
DO	Detrusor overactivity
DOR	Diagnostic odds ratio
EAEMP	European agency for the evaluation of medicinal products
EBLM	Evidence-based laboratory medicine
EBM	Evidence-based medicine
EPINCONT	Epidemiology of Incontinence in the County of Nord-Trøndelag
FDA	Food and Drug Administration
GSI	Genuine stress incontinence (outdated)
HRQoL	Health Related Quality of life
HTA	Health Technology Assessment
HTML	HyperText Markup Language
ICC	Intra-class correlation
ICI	International Consultation on Incontinence
ICIQ-UI SF	International Consultation on Incontinence Questionnaire-Urinary Incontinence Short Form
ICS	International Continence Society
ICUD	International Consultation on Urologic Diseases
IIQ	Incontinence Impact Questionnaire
Internet	The physical network that interconnects computers in order to deliver services like the WWW
Intranet	A physical network, often password protected, that delivers services to local users at offices
IPSS	International prostate symptom score
ISI	Incontinence Severity Index
ISP	Internet Service Provider
IUGA	International Urogynecological Association
JAMA	Journal of the American Medical Association
Kappa	Statistical method for assessing inter-observer agreement
KHQ	King's Health Questionnaire
LR-	Negative likelihood ratio
LR+	Positive likelihood ratio
LUTS	Lower urinary tract symptoms
MCU	Multichannel urodynamics
MUI	Mixed urinary incontinence
MySQL	A relational database management system
N&U	Neurourology and Urodynamics

NHS	National Health Service
NICE	National Institute for Health and Clinical Excellence
NPV	Negative Predictive Value, the proportion of patients with a negative test not having the disease
OAB	Overactive bladder
OCR	Optical character recognition
OR	Odds ratio
MS-DOS	Microsoft Disk Operating System
PASW Statistics 18.0	Now IBM SPSS Statistics: www.spss.com/ibm-announce/
PFD	Pelvic floor dysfunction
POP	Pelvic organ prolapse
PPV	Positive Predictive Value, the proportion of patients with a positive test having the disease
Prevalence	The proportion of patients having a disease
PRO	Patient reported outcome, see 1.1.3 for details
PRO-questionnaires	Patient reported outcome questionnaires
QoL	Quality of life, shorter form for HRQoL
R&D	Research and development
RAND	A contraction of the term research and development. 30 Nobel Laureates have been affiliated with this US corporation. http://www.rand.org/about/history/
ROC	Receiver operating characteristics
RR	Relative risk
SD	Standard deviation
s.e.m	Standard error of the mean
Sensitivity	The proportion of patients with a disease having a positive test (=True positive rate)
SnNout	For a high sensitivity (≥80%) a negative test rules out the condition
Specificity	The proportion of patients with no disease having a negative test (=True negative rate)
SpPin	For a high specificity (≥80%), a positive test rules in the condition
SPSS	Statistical package for the social sciences, now PASW Statistics 18.0
sROC	Summary receiving operating characteristic
SUI	Stress urinary incontinence
UDI	Urodynamic investigation
UI	Urinary incontinence
Urodynamics	Functional study of the lower urinary tract
USI	Urodynamic stress incontinence (formerly GSI)
UTI	Urinary tract infection
UUI	Urgency urinary incontinence
WEB-EPI UI	Web-based epidemiology of urinary incontinence
WHO	World Health Organization
WWW	World Wide Web, service delivered for use on the Internet

1 Introduction

1.1 The ICS definition of UI

1.1.1 From "Social and hygienic problem, objectively demonstrable" to "Any leakage"

The original ICS definition of UI

In 1988, the International Continence Society committee on standardisation of terminology of lower urinary tract function defined UI as "involuntary loss of urine, which is objectively demonstrable and a social or hygienic problem."[3] In accordance with this, the consensus statement at the 1st International conference for the prevention of incontinence proposed the following statement in 1997[3]:

> *"Since there are numerous definitions of Urinary Incontinence, it was agreed that the International Continence Society definition be adopted by the Conference: Urinary Incontinence is a condition in which involuntary urine loss is a social or hygienic problem and is objectively demonstrable. However, the 'objectively demonstrable' criteria may require modification in large-scale epidemiological work."*

A simple Google-search (March 2010) for the definition of "urinary incontinence" showed that many Web sites still continue to use the original definition, thus misleading the Web community.

---

[3] http://www.continence-foundation.org.uk/in-depth/prevention-of-incontinence.php

The ICS Committee on Standardisation of Terminology

The ICS has gone through many processes in the different phases of standardising the terminology related to LUTS. The reports have been published in parallel (dual publication) in several journals to reach the relevant specialist milieus: urologists, gynaecologists and urogynaecologists. The First,[4, 5] Second,[6-8] Third,[9-11] Fourth,[12, 13] Sixth,[3, 14, 15] and Seventh report[16] have been published in urological and gynaecological journals. The Fifth seems to be missing; it has not been published as far as I can see from my literature searches. The latest version of the terminology was published in January 2010, and is from now on a joint effort between the IUGA and the ICS (see section 1.1.2).[17, 18]

Criticism of the original ICS definition

Foldspang and Mommsen argued strongly against the original ICS definition in a study published in the Journal of Clinical Epidemiology in 1997.[19] They showed that it was problematic that UI had to be "objectively demonstrable," and incorrect to define UI as "being a social and hygienic problem" instead of just "any leakage."

They based their conclusion on results from conducted a postal, cross-sectional, age-stratified study of 3,114 women in Aarhus, Denmark, and found that only 63% of 388 women considered UI to be a "social or hygienic problem," and only 22% conformed that they had ever abstained socially because of UI. The authors concluded that the ICS definition of UI at that time presented intrinsic logical problems that invalidated its use in research.

In 1998, Holtedahl and Hunskaar demonstrated the effect different definitions of UI had on prevalence rates, when applied to the same population of women.[20] In this study, 47% reported "any incontinence," 31% reported leakage twice or more often per month, while only 19% admitted incontinence as defined by the ICS.

Criticism of the original definition was also raised in the epidemiology sections in the proceedings from the meetings of the 1st ICI[21] (pp. 216-7) held in Monaco in 1998 and the 2nd ICI[22] (pp. 194-5) held in Paris in 2001, and onwards. In the latter publication, it was the Committee on Epidemiology, chaired by Professor Steinar Hunskaar that very clearly stated that there was a need to change the ICS definition of UI to "any leakage." The committee further suggested that epidemiological studies should contain a minimum data set consisting of a screening question for UI ("any leakage"), frequency and quantity measures, duration, type and severity. In addition, using bother- or HRQoL-measures was recommended.

Acknowledging these criticisms, the ICS definition was changed in 2001[4], and has since then been phrased as "the complaint of any involuntary leakage of urine."[23] In the 3rd ICI on page 298, the process of accepting UI as "any leakage" was discussed.[24]

1.1.2 The present (2010) IUGA/ICS joint terminology

The joint report on the terminology for female PFD (pelvic floor dysfunction) was presented in the January 2010 issue of N&U.[17, 18] The authors clearly stated that there was a need for a core terminology for clinical use and research, and that this terminology had to be "female-specific". The authors argued that such a report should also be as user-friendly as possible, clinically based, able to indicate origin of the term, and to provide explanations.

The joint report further divides the terminology into "**Symptoms,**" "**Signs,**" and "**Urodynamic investigations and associated pelvic imaging**," and left out the

[4] Erratum: The paper referred to says 2002; 2001 is the correct year (e-mail from Professor Steinar Hunskaar)

terms “Urodynamic observation” and “Conditions.” The joint IUGA/ICS terminology report[17] defines the following UI **symptoms** (The abbreviations below are my suggestions):

UI symptoms

(i) Urinary incontinence (symptom): Complaint of involuntary loss of urine. (UI)

(ii) Stress (urinary) incontinence: Complaint of involuntary loss of urine on effort or physical exertion (e.g., sporting activities), or on sneezing or coughing. N.B.: “activity-related incontinence” might be preferred in some languages to avoid confusion with psychological stress. (SUI)

(i) Urgency (urinary) incontinence: Complaint of involuntary loss of urine associated with urgency. (UUI)

(ii) Postural (urinary) incontinence: (NEW) Complaint of involuntary loss of urine associated with change of body position, for example, rising from a seated or lying position. (PUI)

(iii) Nocturnal enuresis: Complaint of involuntary urinary loss of urine which occurs during sleep. (NE)

(iv) Mixed (urinary) incontinence: Complaint of involuntary loss of urine associated with urgency and also with effort or physical exertion or on sneezing or coughing. (MUI)

(v) Continuous (urinary) incontinence: Complaint of continuous involuntary loss of urine. (CUI)

(vi) Insensible (urinary) incontinence: (NEW) Complaint of urinary incontinence where the woman has been unaware of how it occurred. (IUI)

(vii) Coital incontinence: (NEW) Complaint of involuntary loss of urine with coitus. This symptom might be further divided into that occurring with penetration or intromission and that occurring at orgasm. (CoI)

Bladder storage symptoms

(i) Increased daytime urinary frequency: Complaint that micturition occurs more frequently during waking hours than previously deemed normal by the woman.

(ii) Nocturia: Complaint of interruption of sleep one or more times because of the need to micturate (void).

(iii) Urgency: Complaint of sudden compelling desire to pass urine, which is difficult to defer.

(iv) Overactive bladder: (OAB, urgency) syndrome: Urinary urgency, usually accompanied by frequency and nocturia, with or without urgency urinary incontinence, in the absence of UTI or other obvious pathology.

UI signs

Signs are defined as any abnormality indicative of disease or a health problem, discoverable on examination of the patient by health care workers; an objective indication of disease or a health problem. The urinary incontinence **signs** are defined as:

(v) Urinary incontinence: Observation of involuntary loss of urine on examination: this may be urethral or extraurethral.

(vi) Stress (urinary) incontinence (clinical stress leakage): Observation of involuntary leakage from the urethra synchronous with effort or physical exertion, or on sneezing or coughing.

(vii) Urgency (urinary) incontinence: Observation of involuntary leakage from the urethra synchronous with the sensation of a sudden, compelling desire to void that is difficult to defer.

(viii) Extraurethral incontinence: Observation of urine leakage through channels other than the urethral meatus, for example, fistula.

(ix) Stress incontinence on prolapse reduction (occult or latent stress incontinence): (NEW) Stress incontinence only observed after the reduction of co-existent prolapse.

UI diagnoses

As a consequence of these new definitions, the report highlights six **diagnoses** that are common in the sense that there is evidence for a prevalence of 10% or more in women presenting with symptoms of PFD (pelvic floor dysfunction). These six are:

(i) Urodynamic Stress Incontinence. (USI)

(ii) Detrusor Overactivity. (DO), replaces detrusor instability.

(iii) Bladder Oversensitivity. (BO), increased bladder sensitivity, replaces sensory urgency.

(iv) Voiding Dysfunction (VD),

(v) Pelvic Organ Prolapse (POP) and

(vi) Recurrent UTIs.

The joint report also includes definitions for other related PFDs, but these are outside the scope of this thesis, and therefore not mentioned here.

The group suggested that future publications should acknowledge these standards in the "Methods and Materials" section of any publication with this exact text:

> *"Methods, definitions and units conform to the standards jointly recommended by the International Incontinence Society (ICS) and the International Urogynecological Association, except where specifically noted."*

As a curiosity, I remark that the text is erroneous, and should have been phrased "International **Continence** Society" and not "International **Incontinence** Society."

1.1.3 Future assessment: From urodynamics to Patient Reported Outcome questionnaires?

A patient reported outcome (PRO) is defined as any report coming directly from patients, without prior interpretation by physicians or anyone else, about how they function or feel in relation to a disease or treatment.[25] It seems that the technology optimism that had been associated with urodynamic investigations, probably due to the former definition requiring the documentation of objectively demonstrable UI, has now seemingly turned in favour of using PRO-questionnaires to a larger extent for clinical practice and research.

Patient reported outcome (PRO)-questionnaires have now been proposed as a preferred choice for use in clinical practice and research by the 4th ICI, as validated instruments applying non-invasive methods for assessing UI.[2] This committee strongly advocates the use of PRO-questionnaires in the chapter "Initial assessment of urinary and faecal incontinence in adult male and female patients" as their contribution to the proceedings,[26] and this document fully replaces the earlier versions produced by the first three ICI scientific committees led by Jenny Donovan,[27-29] whose reports had pioneered the evidence-rating of all questionnaires that had been developed, grading them from A to C.

In the UK, the National Institute for Health and Clinical Excellence (NICE) have produced many guidelines, also for UI, in full text versions, shorter versions and "pullouts."[30-34] The NICE also recommends the use of questionnaires, and emphasizes that the initial assessment and treatments should be done in primary care settings. The 2006 NICE guideline[30] on UI provides strategies for the evaluation and management of stress incontinence. In this guideline, the NICE argues that that there is no need for invasive diagnostic procedures like UDIs prior to instigating conservative treatments in women with pure SUI.

Dmochowski criticized this view in an editorial,[35] based on the findings of a large, but retrospective audit study by Digesu et al.:[36] "Do women with pure SUI need urodynamics?" Digesu et al. studied 3,428 women aged 24-81, and of these, 52% complained of urinary incontinence, whereas 48% self-reported to be continent. Only 308 women (9%) could be classified as having pure SUI. Of the 308 women who complained of having only SUI, 78% had USI, 8% had DO, 3% had combined USI and DO, and 11% had inconclusive urodynamics (no urodynamic abnormality). The women with inconclusive urodynamics were investigated further with urethrocystoscopy and/or ambulatory urodynamics. Ambulatory urodynamic evaluations revealed that all of these women had DO. The authors therefore conclude that since nearly 20% of the women with pure SUI according to the KHQ in fact have DO, this group needs different treatment options or management. I do not consider this to be problematic. The challenge of finding the 20% with DO still needs to be catered for, before operative procedures are considered. But for conservative measures, this is not so. In **Study I** we found only 8% with DI if the cut-off level was set to 5 for the DIS, while it was 20% for the KHQ, based on the symptoms domain of the KHQ.

In a review conducted by Avery et al.[37] based on a literature search in relevant databases until 2004, applying the same standardized recommendation grades as used in the 4th ICI book,[26] they found 150 randomized trials, investigating 130 treatments of UI. Interestingly, only 50 (38%) of the trials included a grade A questionnaire as an outcome measure, and only 25 (19%) of these trials included grade A questionnaires that were considered to attain the highest level of rigour. Researchers should keep in mind that considerable advances have been made in the non-invasive assessment of urinary incontinence, with 18 questionnaires now achieving the highest level (grade A) of scientific rigour.[37] The assessment of UI symptoms and its impact on patient lives is now characterized by high quality, validated questionnaires and more consistent use of these instruments in RCTs, facilitating future cross-comparison of results between studies.

1.2 Scored questionnaires

In 1994, Professor Abrams postulated the basic requirements for useful questionnaires. The questionnaires had to be facile, each item should demonstrate a causal relationship with the condition to be measured, the score should be able to assist in determining the appropriate treatment, and finally, use of the questionnaire should improve patient care.[38]

In addition, there was a need for evidence-grading and quality assessment of already developed instruments. As part of this process, the ICS adopted the ICUD-process that led to the development of the International prostate symptom score (IPSS). Accordingly, this led to a series of International Consultations on Incontinence (ICIs).

1.2.1 The process of EBM-grading scored questionnaires

The 1st International Consultation on Incontinence (Monaco, 1998)

The first detailed expert-based review of recommended questionnaires for use in assessing UI was provided in 1999 in the proceedings (book) of the 1st International Consultation on Incontinence held in Monaco in 1998,[39] sponsored by the WHO and organized by the ICS and ICUD, WHO-recognized, non-governmental organisations.

The 1st ICI followed an organisational template that has been used in all ICUD-consultations from 1991 and onwards in order to produce a book combining consensus methods with methods of systematic reviews.[39] First, the ICS appoints an ICI Executive Committee which, in consultation with the major relevant scientific societies worldwide, appoints an appropriate range of committees to cover the topic of

the individual consultation. Next, the ICI Executive Committee appoints chairs and scientific committee members with broad academic qualities, representative of the global scientific community and relevant specialties. Each committee is responsible for defining its subject matter. In the 12-month period prior to the next consultation, a systematic review of the relevant medical literature is performed, as a basis for the content of the committee's chapter. Usually, at least three drafts are written and reviewed by the committee members at preliminary meetings, typically held at the American Urological Association or the European Association of Urology, in advance of the final meeting at the consultation itself. At this consultation, the committee chair presents the final draft of this chapter, which is then edited and published as a book chapter, together with the work of the other committees. Each ICI scientific committee, consisting of the Executive Committee and the chairs of the individual committees, provide a series of recommendations for the investigation and treatment of patients, based on the findings of the respective committees at the end of the consultation, and these are formulated in the final book chapters.

The 1st ICI appointed 24 individual committees, including a committee on "Symptom and QoL-assessment" relevant for this thesis. The committee members applied grades of recommendations to already developed questionnaires based on their type of validation, and encouraged the use of questionnaires with the highest possible level of recommendation, both in clinical practice and for research on UI.

The 2nd International Consultation on Incontinence (Paris, 2001)

At the Second International Consultation on Incontinence in Paris in 2001, the scientific committee developed standardised grades of recommendation for all questionnaires,[28] attempting to reflect the Levels of Evidence devised by the Oxford Centre for Evidence Based Medicine. These were applied to evaluate questionnaires

concerning urinary incontinence. At the 2nd ICI, the scientific committee had devised three grades of recommendations:[28]

A: Questionnaires were "Highly recommended" and given a Grade **A** if the committee found that published data indicating that the questionnaire was valid, reliable and responsive to change following standard psychometric testing. Evidence had to be published on all three aspects, and questionnaires had to be relevant for use in persons with incontinence.

B: Questionnaires were "Recommended" and given a Grade **B** if the committee found published data indicating that the questionnaire was valid and reliable following standard psychometric testing. Evidence had to be published on two of the three main aspects, usually validity and reliability.

C: Questionnaires were considered to have "Potential" and given Grade **C** if the committee found published data (including abstracts) indicating that the questionnaire was valid or reliable or responsive to change following standard psychometric testing.

The 3rd International Consultation on Incontinence (Paris, 2004)

At the 3rd ICI,[29] these grades were yet revised and updated to take into account the increasing numbers of published questionnaires concerned with LUTS and incontinence, and also broadening of the field to include pelvic organ prolapse (POP) and faecal incontinence.

The 4th International Consultation on Incontinence (Paris, 2008)

The book of the 4th ICI represented a very marked shift in the recommendations,[26] as the ICI scientific committee recommended that no other questionnaires other than the relevant ICIQ modules were to be used, apart from at very special occasions. The ICI

scientific committee argued that the number of high quality questionnaires already having been developed was too high, and meant that there were now sufficient questionnaires for most purposes, so it was not necessary to encourage the development of new questionnaires, except for particular patient groups. The ICI scientific committee expected that at the time of the next ICI, "Grade A new" questionnaires would either be promoted to "Grade A" because of further high quality publications or relegated to Grade B if further development did not occur.

Although the 4th ICI represented a change of direction from the previous recommendation schemes, questionnaires would still be graded A, B, or C as outlined earlier. Within the description of the ICIQ modular structure (see 1.2.2), the grade assigned to each module would be indicated. In case none of the modular questionnaires were found appropriate for specific research or clinical purposes, the ICI's recommendation was to use a Grade A questionnaire as previously recommended; and where no suitable instrument existed- a Grade B or C questionnaire. The new grading criteria are shown in the documents Table 1 from the document:[2]

Table 1. Criteria for recommendation of questionnaires for UI and UI/LUTS at the Fourth Consultation 2008

Grade of recommendation	Evidence required (published)
Recommended (**Grade A**)	Validity, reliability and responsiveness established with rigour in several data sets
Recommended (**Grade A^{new}**)	Validity, reliability and responsiveness indicated with rigour in one data set
(**Grade B**)	Validity, reliability and responsiveness indicated but not with rigour. Validity and reliability established with rigour in several data sets. To be used if suitable questionnaires not available in ICIQ modular format or Grade A or Grade A new

This version also introduces a new concept: For UI and UI/LUTS, the scientific committees examined the quality of the psychometric evidence. Only where published data were scientifically sound was the label 'with rigour' allowed.

In **Study II** we have used the ISI and ICIQ-UI SF questionnaires (described in greater detail in chapters 1.2.3 and 1.2.4); both are Grade A questionnaires.

1.2.2 The ICIQ Modular structure

Acknowledging the fact that there has been a need for universally applicable PFD-questionnaires that could be widely applied across the population for clinical practice and research, the 1st ICI scientific committee initiated in 1999 the development of a set of questionnaires that would facilitate the cross-comparisons of findings from different settings and studies in a manner similar to that of the International Prostate Symptom Score (IPSS).[39] Hence, the 1st ICI scientific committee decided to develop the ICIQ-modules according to the standard methods of psychometric testing outlined by the "Symptom and QoL assessment" committee.[27]

The ICIQ Advisory Board was formed in 1999, and the project was discussed with the board, and early in the process a decision was made to extend the concept further and develop the ICIQ Modular Questionnaire. The first questionnaire to be developed in this module was the ICIQ-UI SF for urinary incontinence, which has now been fully psychometrically validated.[40]

Given the initial intent to produce an internationally applicable questionnaire, the Advisory Board developed a protocol for the production of translations of its modules. This protocol prescribes the production of the new language version by 2 native speakers of the target language (step 1), back translation into the source language (English) by a native English speaker (step 2), resolution of any differences between the original and the language version (step 3), and revalidation of the new language version. If backward translation is not successful, it is suggested that the questionnaire may need validation in that language.[25, 39] As of May 2010, the ICIQ-UI SF has been translated into 38 languages,[25] among them Italian,[41] Japanese,[42] Spanish[43] and Arabic.[44]

The production of new questionnaires is extremely time-consuming and costly. Based on the recommendations of the "Symptom and QoL assessment" committee at the 1st and 2nd ICI consultations in 1998 (Monaco) and 2001 (Paris), a number of already existing validated questionnaires have been adopted as ICIQ modules and renamed with the permission of their authors. For the first 5 years after inclusion as ICIQ modules, it has been recommended that the original questionnaire should be cited in any ICIQ publications.

In the proceedings of the 4th ICI (2008),[26] Table 3 in that text (see below) shows the ICIQ modular structure, with existing modules, modules that are being developed for urinary tract, vaginal and lower bowel symptoms, and additional modules that are condition specific, dealing e.g. with sexual matters and HRQoL. Eventually, patient satisfaction modules will also be developed, as an important part of assessing treatment effectiveness. The ICIQ modules will eventually evolve into grade A with rigour of all modules as the validation process continues.

Table 3. The ICIQ Modular Structure

CONDITION	RECOMMENDED MODULES	OPTIONAL MODULES	RECOMMENDED ADD-ON MODULES			
	Core Modules (symptom assessment)		HRQL	Generic HRQL	Sexual Matters	Post-Treatment
Urinary symptoms	Males: ICIQ-MLUTS Females: ICIQ-FLUTS	Males: ICIQ-MLUTS LF Females: ICIQ-FLUTS LF	ICIQ-LUTSqol	SF-12	Males: ICIQ-MLUTSsex Females: ICIQ-FLUTSsex	ICIQ-Satisfaction*
Vaginal symptoms and sexual matters	ICIQ-VS		ICIQ-VSqol*	SF-12		
Bowel symptoms and quality of life	ICIQ-B			SF-12	Males: ICIQ-Bsex* Females: ICIQ-Bsex*	
Urinary Incontinence	ICIQ-UI Short Form	ICIQ-UI LF*	ICIQ-LUTSqol	SF-12	Males: ICIQ-MLUTSsex Females: ICIQ-FLUTSsex	
CONDITION	B) Specific patient groups		HRQL	Generic HRQL	Sexual Matters	
Nocturia	ICIQ-N		ICIQ-Nqol	SF-12	Males: ICIQ-MLUTSsex Females: ICIQ-FLUTSsex	
Overactive Bladder	ICIQ-OAB		ICIQ-OABqol	SF-12	Males: ICIQ-MLUTSsex Females: ICIQ-FLUTSsex	
Neurogenic	ICIQ-Spinal Cord Disease*			SF-12		
Long-term catheter users	ICIQ-LTC*			SF-12		
Children	ICIQ-CLUTS*		ICIQ-CLUTSqol*			

Gray: In development; black: Grade A or A (New)

370

1.2.3 The ICIQ-UI SF (Appendix 3 and 4)

The ICIQ-UI SF is a sum-score developed by the International Consultation on Incontinence Modular Questionnaire study group (www.iciq.net).[39] This questionnaire is the UI element in a modular package of questionnaires for related PFD problems. Avery et al. have psychometrically validated the ICIQ-UI SF, in a paper that in a transparent manner describes the validation process of the ICIQ-UI SF, the thoroughness and completeness all of the modules of the ICIQ have to undergo.[40] The ICIQ-UI SF is developed for assessing the prevalence, severity, impact on quality of life, and type of UI. Two studies have compared the ICIQ-UI SF with urodynamics since its introduction in 1999.[45, 46] The ICIQ-UI SF has been translated to 38 languages,[25] among them Norwegian, Swedish, Danish and Finnish. It has undergone many validation studies, and has been used in many different types of studies. The ICIQ-UI SF has received the highest grade of recommendation by the committees of the 2nd and 3rd International Consultations on Incontinence,[28, 29] and is now recommended by the ICI as a gold standard outcome measure for future research and clinical practice, according to the proceedings of the 4th ICI (pp. 1771-2).[2] The committee recommends using high quality questionnaires (Grade A) for the assessment of the patient's perspective of incontinence symptoms and their impact on quality of life, and recommend other Grade A questionnaires for more detailed assessment.

The ICIQ-UI SF consists of four items. Only the first three items are part of the sum score. The fourth item included was meant to be a self-assessment of the aetiology, and was included by the expert committee because it was thought to be useful in clinical practice, to understand patients' perception of the cause and type of leakage. This part of the questionnaire has not been subjected to validation processes.

The third item was constructed as an HRQoL-scale, in the form of a VAS ranging from 0 (“not at all”) to 10 (“a great deal”).

The complete form we used in Norwegian is shown in Appendix 3, while the corresponding web version is shown in Appendix 4. The four items of the ICIQ-UI SF (the three first are sum-scored items) of the ICIQ-UI SF are:

Item [1] “How often do you leak urine?” (Tick one box) [Scores 0-5]

0 “Never”

1 “About once a week or less often”

2 “Two or three times a week”

3 “About once a day”

4 “Several times a day”

5 “All the time”

Item [2] “How much urine do you usually leak (whether you wear protection or not)?” (Tick one box) [Scores 0, 2, 4, or 6]

0 “None”

2 “A small amount”

4 “A moderate amount”

6 “A large amount”

Item [3] “Overall, how much does leaking urine interfere with your everyday life?” (Please ring a number between 0 (not at all) and 10 (a great deal) [Scores 0-10]. As it is presented in the original questionnaire, this is a eleven-point ordinal scale more than a visual analogue scale, ranging from 0 “Not at all” to 10 “A great deal,” as there is no continuous line.

Item [4] “When do you leak urine?” (Please tick all that apply to you). [Unscored]. This item covers different aspects of UI: no UI (#1), UUI (#2), SUI (#3, #5 and #6), NE (#4), IUI (#7) and CUI (#8).

The answers to the first three items result in a sum score, ranging from a minimum score of 0 (“no UI”), to a maximum score of 21. Preliminary cut-off scores were set to 0= “no UI” and ≥1= “UI.” The first two items are “objective” measures, summing up to a range of 0 to 11, while the third item is a “subjective” measure ranging from 0 to 10.

We used an official Norwegian language version (bokmål) of the ICIQ-UI SF that was translated from English (Appendix 3) by the ICI modular questionnaire study group. This form was incorporated into Web survey, as shown in the screen dumps in Appendix 4. In the web-form, items 1-3 are presented in the opposite direction compared to the authorised version.

1.2.4 The ISI (Appendixes 2 and 4)

The Incontinence Severity Index (ISI) was developed in Professor Steinar Hunskaar’s research group by Hogne Sandvik for use in epidemiological surveys to identify the severity of urinary leakage in women with UI. The ISI is a semi-objective and quantitative measure, which purposely does not include a HRQoL dimension or other subjective perceptions of leakage as being a problem or not, and thus reflects the current UI definition of ”any leakage.” Due to limited power, the first study published in 1993 was only able to validate a simplified 3-level version of the ISI.[47] In a second study published in 2000, more women (n=265 with 315 pad-weighings) were included, and Sandvik et al. were able to demonstrate that a four-level index was just as valid.[48] Sandvik et al. recommended using the four-level index as it also gives a more balanced distribution in clinical studies. Since its introduction in 1993, the ISI

has been used in many different studies of UI, both epidemiological[49-63] and clinical studies.[64-79] The ISI has received the highest grade of recommendation by the committees of the 2nd and 3rd International Consultations on Incontinence.[28, 29]

In **Paper II,** the 3-level Incontinence Severity Index (ISI) developed by Sandvik et al.[47] was used to characterise the severity of incontinence. This index is calculated by multiplying the reported frequency (four levels) by the amount of leakage dichotomised to two levels:

Item [1] "How often do you experience urinary leakage?" (Four levels):

1 "Less than once a month"

2 "One or several times a month"

3 "One or several times a week"

4 "Every day and/or night"

Item [2] "How much urine do you lose each time?" (Two levels):

1 "Drops or little"

2 "More"

By multiplying the scores of question [1] and [2], the resulting score is a multiplicative index score with values from 1 to 8. The resulting index scores 1 to 8 points, and is further categorised into three levels:

"Slight" 1 to 2 points

"Moderate" 3 to 4 points

"Severe" 6 to 8 points

Typically, slight incontinence denotes leakage of drops a few times a month, moderate incontinence denotes daily leakage of drops, and severe incontinence denotes larger amounts at least once a week. In this development study, Sandvik et al.

found in 1993 that slight incontinence meant a leakage of 4 g/24 hours (95% [CI]: not calculated); moderate meant 17 g/24 h, and severe meant 63 g/24 h.[47]

The 3-level ISI has later been validated against pad-weighing tests in two studies, one by Sandvik et al. in 2000[48] and one by Hanley et al. in 2001.[80] In their validation study, Sandvik et al. found that slight incontinence meant a leakage of 6 g/24 hours (95% [CI]: 2 to 9); moderate meant 17 g/24 h (13 to 22), and severe meant 56 g/24 h (44 to 67).[48]

In the validation study by Hanley et al., they found that reliability and responsiveness of the 3-level ISI was satisfactory. They found that slight urinary incontinence represented a median leakage of 32 g/48 hours; moderate 29 g/48 h, and severe 143 g/48 h ($\chi^2 = 14.9$, $P < 0.001$; mean ranks 41.8, 50.2, and 80.7 respectively).[80]

In **Paper III,** the 4-level ISI was used in its original form in Norwegian translation. It consists of two items, defining frequency (four levels) and volume (three levels) of leakage.[48] The ISI is a multiplicative score based on these two items:
Item [1] "How often do you experience urinary leakage?" (Four levels):

1 "Less than once a month"

2 "A few times a month"

3 "A few times a week"

4 "Every day and/or night"

Item [2] "How much urine do you lose each time?" (Three levels):

1 "Drops"

2 "Small splashes"

3 "More"

By multiplying the scores of question [1] and [2], the resulting score is a multiplicative index score with values from 1 to 12. This index score is then further categorised into four levels of incontinence severity:[48]

"Slight"	1 and 2 points
"Moderate"	3, 4 and 6 points
"Severe"	8 and 9 points
"Very severe"	12 points

The ISI has later been scored "0" for "no incontinence" in studies where e.g. treatments result in patients turning from being incontinent to continent.

The 4-level ISI has been validated against pad-weighing in two studies, the previously described study by Sandvik in 2000[48] (also 3-level ISI), and in a Spanish study in 2006.[81]

For the 4-level ISI used in a Norwegian population, Sandvik et al. found slight incontinence to indicate a leakage of 6 g/24 hr (95% [CI]: 2 to 9), moderate incontinence 23 g/24 hr (15 to 30), severe incontinence 52 g/24 hr (38 to 65), and very severe incontinence 122 g/24 hr (84 to 159).[48]

In the Spanish study, Sandvik et al. found slight incontinence in primary care vs. hospital care to indicate a leakage of 10 g/24 hr (95% [CI]: 2 to 17) vs. 6 g/24 hr (3 to 9), moderate incontinence 32 g/24 hr (17 to 47) vs. 44 g/24 hr (24 to 63), severe incontinence 100 g/24 hr (49 to 151) vs. 102 g/24 hr (70 to 134), and very severe incontinence 223 g/24 hr (-8 to 453) vs. 193 g/24 hr (124 to 261), respectively.

1.3 From paper to Web-based epidemiological research

1.3.1 Postal methods

Many researchers have experienced that conducting postal surveys for epidemiological research is costly and time-consuming, with many demanding manual phases of work. The manual processes have been eased since the days of the classic paperweight, after e.g. the introduction of a Pitney Bowes machine for folding forms and feeding them into envelopes at our office. The three photos depict a letter weight (Photo 1), a paper folding and enveloping machine (Photo 2), and questionnaires folded, enveloped and stamped, ready to be delivered by the post office (Photo 3).


(Photo 1: Atle Klovning)


(Photo 2: Atle Klovning)


(Photo 3: Atle Klovning)

Since the manual punching and coding of data may often lead to erroneous data entry, and in order to compensate for most of these manual efforts, companies enabling an automatisation of these processes from defining a layout of the questionnaire, producing the final form, merging it with an address list, printing, enveloping, stamping and posting these forms have solved some of the researchers' requirements. When the forms are returned, they may be OCR-read and entered directly into a database.

1.3.2 The Internet and the World Wide Web

The English scientist Tim Berners-Lee invented the World Wide Web in 1989, and was knighted in the UK for this achievement in 2004. It is important to distinguish between the Internet and the World Wide Web (WWW). While the Internet is the hardware and software technologies that connect the physical computers, the World Wide Web is one of the many services offered on the Internet; other services are e.g. e-mail, chat, online games and video. Today, the community-enabling software Facebook, with its over 500 million users might represent a new research arena to be explored. As an example of this, I was recently asked to review a paper that uses Facebook as an arena for qualitative research, and I expect that many studies using such community Webs will be performed in the future, and to enable new arenas for recruitment to Web-based studies, for both quantitative and qualitative research methods.

1.3.3 Web-based research

Searching Medline February 2010 with the term ”World Wide Web”, I retrieved 40,717 articles, the earliest from 1994. Using the search term “Internet”, I retrieved 39,765 articles. Using the Boolean operator “OR” between the two terms in order to define any complementary set of papers, yields 40,717 papers. Although, the defined MeSH-term is “Internet,” this simple search shows that the phrase search “World Wide Web” is a more complete search term. By browsing through the list of papers retrieved, the first years of these papers mainly covered IT-communication and teaching/learning, thereafter imaging and telemedicine.

The first paper I found relevant for this thesis is one on ”Health status assessment via the World Wide Web,” published in 1996.[82] This study lasted one year, and collected data from 4,876 individuals on the Web using the RAND 36-item

multiple-choice questionnaire. The authors were optimistic and concluded that "the use of Web technology to administer patient surveys could dramatically lower the cost of performing both randomised clinical investigations and routine outcomes monitoring. As a result, the WWW may play an important role in advancing health services research and outcomes-based patient care."

Eysenbach cited this paper in a Letter to the BMJ, stating: "Obviously, the Web community is not a representative sample of the whole population, and results obtained with questionnaires on the WWW are biased towards self-selection; thus they must be interpreted with care and verified in an unbiased population."[83]

The largest Web-based epidemiological study I found in my literature search is a Swedish follow-up study of 96,000 women born between 1943 and 1962 (then aged 30-49 years) residing in the Uppsala region, and who were invited to fill in a posted paper form in 1991/92.[84] The overall response rate was low, 51% (49,248 women). Of the original 96,000 women, 47,859 (50%) were recruited to answer questions about smoking, body size and shape, use of oral contraceptives and their reproductive history, altogether approximately 70 items to answer. In section 6.2 of her thesis,[85] Ekman points out that the questionnaire was large, with 90- not 70 items, and took 1½ hours to fill in.

The Web-based follow-up study was launched in February/March 2003, and invited 47,859 women to answer a web-based questionnaire only. The response rate to this study was 33% (15,922 women). The 31,937 non-responders were randomised to 5 different response modes depending on whether they had provided their e-mail address or not.

Among the 30,880 women without an e-mail address, 4,974 received a postal reminder with a paper questionnaire or optional Web-form (Group 1.1), 25,906

received a postal reminder only with the Web option (Group 1.2). The 1,057 who had provided their e-mail addresses were randomised into 3 groups. Group 1.3 received a postal reminder with either a paper or a Web questionnaire, Group 1.4 received an email with a login to the Web questionnaire, while Group 1.5 received an email with a direct link to the Web questionnaire, not requiring login.

The overall response rate after this first reminder increased from 33% to 45%. Of these additional 12%, 3,476 (61%) used the Web option. But, when given the option of paper versus Web, the women preferred paper. In Group 1.1, only 139 of 2,149 (6%) chose Web, and in Group 1.3 20 of 191 (11%) did so.

After the second reminder, the overall response rate rose from 45% to 72%. The women were either randomised to a postal reminder with a paper questionnaire (25, 145 women) or e-mail reminder with login (1,135 women). Web responses accounted for only 198 of these 26,280 (0.8%) second reminder responders.

In total, after 2 reminders, 41% responded to the Web questionnaire, while 31% responded to the paper questionnaire. Analysing response rates to a more profound extent, they found that the Web-, paper- and non-responders respectively did not differ significantly in age, physical activity levels, and BMI. The responders answering either the Web or paper questionnaires had a higher level of education and income and a lower level of smoking than non-responders. The RRs for the association between different sociodemographic variables showed that using the Web did not introduce any important issues compared to using traditional, postal methods. The authors found no mode effect. The bias associated with collecting information using Web questionnaires were not greater than that caused by paper questionnaires. This paper is one of the papers in Alexandra Ekman's thesis.[85] She argued that

Sweden is well situated for Web-based surveys and epidemiological studies because of the massive outreach of the Internet.

Summing up, this study showed that a Web-only solution would only give a response rate of 41%, and only by having the option to fill in a paper version would the overall response rate be 72%. However, it should be noted that the response rate in 1992 was only 51%. The follow-up study showed that a combination of different response modes gave the highest response rate.

Representation issues and biases

Couper published three useful papers on the different issues of importance for researchers using the Web for research- a review in 2000,[86] a study on designing Web-based surveys,[87] and a paper on representativity issues.[88]

In the paper on "Web surveys: A review of issues and approaches,"[86] Couper discusses the pros and cons of the Web-survey methods. Three biases are important: coverage and sampling bias, nonresponse bias and measurement bias. Nonresponse bias occurs when not all people included in a sample are willing to complete a survey, and measurement bias arises when responses deviate from their true values. This could be because of lack of motivation, comprehension problems, poor wording or design. Using telephone or an interviewer gives a possibility to explain and clarify the questions. Coverage and sampling bias occurs when one does not reach the appropriate target population.

The paper also discusses different modes of recruitment that might be used to increase participation rates in Web-based studies. The table below is from this paper, and illustrates the main types of Web surveys. **Study II** in this thesis is an unrestricted self-selected survey.

Table 2. Types of Web Surveys

Nonprobability Methods	Probability-Based Methods
1. Polls as entertainment	4. Intercept surveys
2. Unrestricted self-selected surveys	5. List-based samples
3. Volunteer opt-in panels	6. Web option in mixed-mode surveys
	7. Pre-recruited panels of Internet users
	8. Pre-recruited panels of full population

In another paper, the same Couper discusses these issues of representation in eHealth research; with a focus on Web-surveys.[88] In short, Couper urges caution, particularly in replacing existing research methods with Web-based methods only.

Producing Web-based forms

Couper also investigated different modes of designing and presenting a survey on the Web.[87] This Web survey was designed to study the use of a progress indicator or not, multiple-item screens versus single-item screens, and radio buttons versus entry boxes. Couper found that entry boxes were easier to avoid answering, but rather than arguing for one certain approach, he suggested that a more tailored response should be applied. Couper found only marginal evidence for the hypothesis that a progress indicator reduces respondent abandonments. He found faster completion times and less missing data for multiple-item screens.

In a randomised testing of alternative survey formats amongst 4,208 anonymous volunteers over three months on the WWW, Bell et al. found that the matrix format speeded up the completion time of the SF-36, compared to a list format.[89]

Another useful guide, is the one published by Birnbaum on how to perform Web-based research.[90] In this overview, he describes the major issues concerning designing, programming, and executing of Web-based research, and discusses the different related pitfalls.

One way of reducing nonresponse bias might be to use registry-based emails, as in alternative 7 in the previous table. In a systematic review of 17 Internet-based surveys of health professionals, Braithwaite et al. finds that response rates varied from 9 to 94% in 12 studies,[91] and discussed the issue of problematic external validity of findings from Web-based studies.

Demands on Web survey tools

Bälters et al. addressed the requirements of tools for Web-based epidemiological research.[92, 93] In the first paper published in European Journal of Epidemiology in 2005,[92] they argued in favour of Web-based epidemiology:

> *"Data collection in epidemiological studies is to a large extent made by printed questionnaires, telephone interviews, face-to-face interviews, or a combination of these methods. These methods are costly and time consuming. The main cost in using printed questionnaires comes from completing missing or unrealistic answers by phone interviews, and transferring the answers to computer readable format. Furthermore, the time period between the first distribution of a questionnaire and first statistical analyses may be long, maybe months or even years."*

They pointed out that Web-based surveys have the potential of reducing these problems significantly, and two main advantages of Web questionnaires compared to traditional printed questionnaires are the immediate control of answers, and instant electronic storage.

In a subsequent brief report in Epidemiology, they present the following supportive arguments for Web-based epidemiology:[93]

> *"Web questionnaires can be used for research purposes in population-based settings in which Internet access is high, although we found that the initial response rate was lower than for the traditional printed questionnaire. In comparison, the willingness to answer a second questionnaire was higher when using a Web questionnaire instead of a printed questionnaire. Personalised feedback in the Web questionnaire further increased the compliance rate for a second questionnaire. Total response rates for the second part of the questionnaire were similar for the printed and the Web questionnaires."*

Security issues

On the other side, Bälter et al. argue that the fear of entering sensitive data could reduce the number of respondents. [92] Already, the Web has been used to collecting sensitive data: information about drug dealing, drug and alcohol use, and sex habits. It seems that on the Web, people are even willing to expose themselves to poker and pornography, and as a consequence of this are at a great risk of exposing themselves and their computers to devastating security attacks. There have been performed several Web studies on perceived stigmatising conditions - like vaginal pain,[94] depression,[95] vestibular pain,[96] and illicit substance abuse,[97, 98] assuming that the "anonymous" study setting might ease people in exposing taboos. UI too, has been regarded as a stigmatizing disorder, and faecal incontinence an even more burdensome condition. It is a major concern that people are willing to share any kind of information on the Internet, thinking they are in the safe realms of their private homes, while it also may provide new and useful research arenas, as long as we can be sure that researchers, Web survey producers and ethical committees share a responsibility in taking care of the integrity and privacy of the WWW responders.

2 Aims of the two studies in this thesis

The aims of this thesis were to validate scored questionnaires to be used in clinical practice and epidemiological research on urinary incontinence (UI).

2.1 Study I

- To validate a scored questionnaire, the Detrusor Instability Score (DIS) (**Paper I**)

2.2 Study II

- To analyse how Web-based recruitment performs compared to postal surveys (**Paper II**)
- To validate the International Consultation on Incontinence Questionnaire - Urinary Incontinence Short Form (ICIQ-UI SF) against the Incontinence Severity Index (ISI) (**Paper III**)
- To construct a severity scale for the ICIQ-UI SF (**Paper III**)

3 Materials and Methods

3.1 Participants

3.1.1 Study I (Paper I)

The outpatient clinic

The findings in this study were based on 250 consecutively included patients at an outpatient clinic at the University Hospital in Trondheim, Norway. More detailed information about this clinic has been published in a Norwegian paper with English abstract.[99] In 1988 this clinic was awarded "Det nytter prisen" by the Norwegian government for its outstanding service, a prize for the most beneficial health service in 1988.

A urotherapist used a structured questionnaire to record the history (Appendix 1) and gathered other relevant information prior to the examination by a specialist in urogynaecology. The mean age of the women that were included (± s.e.m.) was 49 years (± l). Of the women, 96 of them (42%) had been incontinent for 10 years or more. Urodynamic investigations of these women revealed stress incontinence in 58%, sensory urgency in 19%, motor urgency in 21% and mixed incontinence in 32%. Using a 3-level severity index, we found that 7% had slight, 25% moderate, and 68% severe urinary incontinence.

3.1.2 Study II (Paper II and III)

Web-based recruitment

We used the Web (see Appendix 4) to invite a convenience sample of women to join a women's health study by self-selected participation, focusing on women's general health.

Female users of major Norwegian Internet sites were asked to join the study by three different routes: a general health Web site (NettDoktor.no), the health section of a general-purpose Web portal (StartSiden.no), and the newspaper Web site of Verdens Gang (VG.no). NettDoktor was, at that time, the Norwegian part of Europe's largest health Web site, StartSiden was Norway's largest Web portal, and VG.no, Norway's largest Web-based newspaper. At the first two of these Web sites we used fixed placed banners containing the logo of the University of Bergen, whereas at VG.no the study was linked to NettDoktor by a link in an interview in VG.no (Appendix 4). The three investigators were named on the introductory page of the Web questionnaire. The study was anonymous, and informed consent was not considered necessary, as the study collected no personal information and participation was voluntary.

Between February 23, 2002 and April 22, 2002, women accessing the NettDoktor Web site were recruited by a banner with the text: "Join the large women's health study at the University of Bergen" on the front page of www.NettDoktor.no (NettDoktor). Between April 25, 2002 and August 20, 2002, women were able to access our study by means of StartSiden (www.startsiden.no), where we used the text "UiB/Join the women's health study" on the front page of the health section (http:// www.startsiden.no/helse/). The VG på Nett (www.vg.no)

interviewed me about urinary tract disorders, incorporating a direct link to the study in the Web text, easily accessible in the period from March 4, 2002 to March 6, 2002.

Altogether 1,812 female Web users were recruited, and 343 of them were sub-branched into two incontinence questionnaires by answering "Yes" to a single question on whether they had "Any leakage of urine" (Appendix 4). Those who answered "No" were not entered into this part of the questionnaire; one of the features Web-based forms enable.

Split-half sampling for scaling the ICIQ-UI SF

We used split-half sampling for developing and validating the severity grading of the ICIQ-UI SF.[100] The random functions in SPSS were used to extract a random half of the 343 women with UI, yielding a development sample (n=171) and a validation sample (n=172). The respondents in the first sample were used to develop the scale for the ICIQ-UI SF, while the remaining respondent sample was used to validate the severity scaling of the ICIQ-UI SF.

3.2 Methods

3.2.1 Study I

The structured study questionnaire

The complete study questionnaire is shown in Appendix 1. It is a multiple-choice questionnaire, incorporating the 10-item DIS.[101] It consists of six sections with a total of 50 items, covering the gynaecological history, voiding history divided into the storage phase (sensation, detrusor activity, SUI), the emptying phase, and the severity of the UI. Based on this structured questionnaire, the urotherapist recorded the diagnosis. The DIS was not calculated. After the urodynamic investigations, the urodynamic diagnosis and the gynaecologist's final diagnoses were separately recorded. The urotherapist and the gynaecologist were blinded to each other's diagnosis and to the DIS. The DIS (Kauppila score) was independently calculated by the authors AK and SH of **Paper I** after the study was over, and the forms had been sent to the University of Bergen, Norway.

The DIS

Kauppila et al. developed this sum-score and published it in 1982. They had observed that the major cause of failure in the surgical treatment of stress urinary incontinence (SUI) in women was occult detrusor instability. In order to detect the degree of detrusor instability, urological histories were standardised by scoring the replies to ten specific questions with 0 (indicative of SUI), 1 or 2 (slightly and markedly indicative of detrusor instability, respectively). The sum of the scores was termed the "detrusor

instability score" (DIS). In the original paper they proposed a cut-off level at 7,[101] and chose a cut-off level at 5 when the DIS was validated this in a subsequent paper.[102] The aim of our study was to validate this cut-off level in an outpatient setting. Table 1 on page 138 of the development study by Kauppila et al. shows the different items of the DIS.[101] Note the incomplete wording of each question.

138 *A. Kauppila et al.*

Table I. *Questionnaire for recording the occurrence and degree of detrusor instability.*

Questions	Stress incontinence		Detrusor instability	
1. Feeling of urgency to void before involuntary loss of urine	No	(0)	Mild Strong	(1) (2)
2. Involuntary loss of urine during sudden physical stress	Yes	(0)	Also in other circumstances	(2)
3. Involuntary loss of urine after physical stress	Immediately	(0)	After a few seconds	(2)
4. Amount of urine escaped	Small	(0)	Moderate Large	(1) (2)
5. Ability to stop voiding	Yes	(0)	No	(2)
6. Painful sensation during voiding	No	(0)	Yes	(2)
7. Urgent need to void in haste, psychic strain or nervousness	No	(0)	Mild Strong	(1) (2)
8. Frequency of diurnal voiding	5 times or less	(0)	6–7 times 8 or more	(1) (2)
9. Frequency of noctural voiding	0–1	(0)	2–3 times 4 or more	(1) (2)
10. Previous urinary infections necessitating chemotherapy	0–1	(0)	2 or more Chronic infection	(1) (2)

Sum of the scores = detrusor instability score

In the development study,[101] the DIS was calculated for 134 patients both preoperatively and 2 years after operation, and 112 of these women were also evaluated by bead-chain urethrocystography (UCG) before operation. They found a 10% failure rate among the 72 patients with a DIS of 0-7, which they defined to be caused exclusively or nearly exclusively by detrusor instability, significantly less than the 32% failure rate in the 62 women with a DIS of 8-16, which they defined to be SUI complicated by marked detrusor instability. The 38% failure rate in 47 women of peri- or post-menopausal age and having a DIS of 8-16 was higher than the 10% in the remaining 87 women.

Further, they found that there was an increased risk of failure in patients who had a DIS of 8-16 combined with either a urethral inclination angle $\leq$ 80 degrees

(46% failure rate in 28 women) or a posterior urethrovesical angle ≤ 160 degree (43% failure rate in 30 women) in lateral bead-chain UCG during straining. As a measure of responsiveness, they observed that the mean DIS decreased after successful surgery, but remained constant in cases of failure. The scored urological questionnaire seemed to facilitate the detection of patients with detrusor instability. This was potentiated by lateral bead-chain UCG findings which indicated a low motility of the proximal urethra or bladder neck.[101] The authors concluded that in patients with a DIS of 0-5 and a positive Marshall test operation was indicated, whereas these patients should be treated conservatively if the Marshall test was negative. Patients with DIS 6-20 should undergo UDI prior to surgery for SUI.

The DIS has been used in several settings, like books,[103] studies on the agreement of anamnestic data by Voigt,[104] Kujansuu,[102] and in the preparation of Swedish guidelines.[105]

Urodynamics

According to Rosier et al. the conventional view of urodynamics was a series of more or less agreed-upon clinical tests consisting of e.g. flow- and pressure-flow studies, filling cystometry and/or assessment of the urethral closure function. [106] Also, Dmochowski argued that although UDI is a demanding procedure, it remains the only functional evaluation of bladder and urethral activity that can segregate detrusor and urethral contributions to incontinence and voiding function.[107]

In our study, pressure measurements were performed using a fibre-tip sensor connected to the Laborie system 2 000 (Camtech Ltd.). Simultaneously, urethrocystometry was done in a semi-prone position with temperated water at a filling speed of 25 ml/min. The volume at first sensation of micturition was measured. Urgency, bladder capacity, and uninhibited detrusor contractions in the filling phase

were registered. In patients with GSI anamnestically, a normal micturition list, and normal sensibility in the filling phase, bladder filling was restricted to 300ml. Urethral pressure-profile measurements in the semi-prone position were done at the above-mentioned capacity with the same sensor. With an automatic withdrawal unit the catheter was retrieved from the urethra at a speed of 2 mm/sec. We registered three resting profiles and three cough profiles. Functional urethral length, maximum urethral pressure and maximum closure pressure were calculated as a mean of these three. The differential pressure was registered simultaneously with the cough profiles. Finally, flowmetry was done with the patient sitting.

For further explanation of terms, consult the most recent IUGA/ICS join report.[17]

Establishing the gold standard

Based on the descriptions of gold standards, the one that was used in **Study I**, can be described as en expert opinion diagnosis, with or without prior UDI.

In Appendix 1, a guide for filling in the forms is provided. In short, the urotherapist filled in the structured questionnaire, i.e. the three pages of the form that were stapled, filling in only one type of incontinence, but had the option to fill in other diagnoses in the "other" four fields. The urotherapist did not calculate the DIS (Kauppila score).

Page 4 of the questionnaire was loose-leafed and followed the patient's medical record. The gynaecologist had no knowledge neither of the structured study questionnaire, nor the diagnosis the urotherapist had set, nor the DIS. The gynaecologist recorded a urodynamic diagnosis (several options permitted) and the final clinical diagnosis based on the UDI and an extensive gynaecological assessment.

The gold standard chosen in this study was the gynaecologist's expert opinion based on UDI.

Sensitivity, specificity, PPV and NPV and accuracy

The calculations for the test characteristics sensitivity, specificity, PPV and NPV and accuracy were performed in accordance with the definitions used in the book "Clinical Epidemiology."[108]

ROC

ROC-curves are actually a plot of the true positive rate (sensitivity) on the Y-axis against the false positive rate on the X-axis (1-specificity), according to Sackett's book on p. 117.[108] It is also possible to convert the data to odds ratios to enable pooling of data for a meta-analysis.[109] A very useful Web site explaining the ROC is the one by Steve Simon.[110] We chose to use ROC-curves because they were thought to be the best way of illustrating test characteristics.

3.2.2 Study II

The WEB-EPI UI

All screen dumps of the Web-based questionnaire are shown in Appendix 4. The banners or links led to a short introductory page presenting the logo of the University of Bergen and our department. The questionnaire was titled "Women's Health Study 2002." The introductory text read:

> *"At Section for General Practice at the University of Bergen we have investigated several female health disorders for many years. Now we wish to use the Internet to conduct a new study. All entries will be anonymous, and the data collected will be used for research purposes. We hope you would like to participate in this study. This will be done by answering a few questions, taking only a couple of minutes. If you do not want to enrol, simply click your way out.*

Click 'Next Page' to continue, and 'Clear' to delete all. Best wishes, AK (Researcher), HS (Researcher), SH (Professor)."

Page 2 of the study presented six items (age, gender, menarche, menopause, pregnant, number of children). Page 3 presented six more items (number of voidings per 24 hours, night-time voidings, urinary tract infections, completeness of bladder emptying, urgency, urinary leakage). This last item was the only branching item in the entry form. It was formulated "Do you have urinary leakage?" ("Yes" or "No"), and clicking here was the only mandatory item to be entered. "No" directed the respondent to the exit page, where the respondent could choose between "Finish" and "Clear" in English language. "Yes" to this question defined the respondent as having UI, consequently branching the respondent into two validated questionnaires, the ISI as items number 2 and 3 of the EPINCONT questionnaire on page 4 (10 items), and the ICIQ-UI SF on page 5 (four items). Clicking "Next Page" or "Clear" then led to the exit page, with the choice between "Finish" or "Clear" in English language. After the exit page, the users were forwarded to a Web page on UI at NettDoktor. The respondents were not promised any score or feedback, and no kinds of incentives were offered. We had no initial contact with potential participants. Respondents stating they had no UI had to go through four Web pages, whereas those stating they had UI had to go through a total of six Web pages. Some of these pages were larger than normal screen resolution size, and had to be scrolled. All navigation buttons were non-modifiable and in English language. There were no "Back" buttons, but only "Next Page" or "Clear" on each page. Users were not provided a summary of their responses before the results were submitted.

Terminology

The terminology used in the published papers follow the ICS definitions as of 2002.[23]

The survey software

The survey was performed using the client-side software Inquisite (Inquisite Inc., Austin, TX, USA), and data were deployed to a database at a Web hotel located at UNI·C, the Danish IT centre for education and research. We used no passwords or login procedures. Colleagues piloted the usability and technical functionality of the survey before it was fielded. Log files were checked; they contained no person identification items, e-mail addresses, IP-addresses, or cookies. Neither the participation rate, nor the view rate, nor the completion rates were determined. No check was possible to prevent users accessing the survey several times, as we did not use cookies, collect IP addresses, or use login forms. The time stamping of data entry was manually checked in the final database. Although all data were time stamped, there was no track of the length of time used to fill in the form.

Confidence interval analysis (CIA)

All confidence intervals in Table 1 in **Paper II** were calculated by the Newcombe method for comparing independent proportions using the DOS-based software, CIA.[111] The CIs were calculated one by one, and the asterixes were assigned by comparing the confidence intervals between the EPINCONT and the WEB-EPI UI studies. The single asterisk (*) indicate the instances where the point estimate of one variable is not an element of 95% CIs of the corresponding variable, thereby indicating a statistically significant difference. The double asterisks (**) indicate the

instances where the confidence interval for the difference between the independent proportions does not contain zero, implying a statistically significant difference.

Correlation strategy

In **Paper III** the four levels of the ISI were plotted against the ICIQ-UI SF total sum-score with and without the HRQoL dimension. The association between the ISI and ICIQ-UI SF scores was investigated by Spearman's rank correlation coefficient (rho),[112] as this correlation is used for ordinal variables.

Determination of unweighted Kappa values

Kappa values were calculated using the SPSS on 4x4 contingency tables of the severity (slight, moderate, severe, very severe) of UI by arbitrarily changing the severity intervals until maximum Kappa was obtained.

Kappa with weighting (Lowry)

To my knowledge, SPSS is only able to produce unweighted Kappa statistics, so in order to achieve the weighted Kappa statistics, the contingency tables with maximum unweighted Kappa values produced by SPSS were manually entered into the dynamic Web pages provided by Professor emeritus Lowry.[113] This could probably have been programmed in SPSS, but functioned well on the Web site. By entering these tables into the Web site, we were able to calculate Kappa scores with linear and quadratic weighting.

Scaling the ICIQ-UI SF

In order to create a scale for the ICIQ-UI SF based on the ISI as the assumed gold standard, we iteratively calculated the weighted Kappas for the unweighted Kappas that SPSS produced for the different intervals for the severity of the ICIQ-UI SF and

the ISI. This process might also have been programmed, but functioned well the way it was done manually for the development sample (n=171) described at the end of section 3.2.1. Accordingly, the weighted Kappas were calculated for the validation sample.

4 Summary of results

4.1 Paper I

Klovning A, Hunskaar S, Eriksen BC. **Validity of a scored urological history in detecting detrusor instability in female urinary incontinence** *Acta Obstet Gynecol Scand* 1996;75:941-5.

Aim:

- To validate a scored questionnaire, the Detrusor Instability Score (DIS).

The mean age (s.e.m.) of the 250 women included was 49.2 years (0.9) (range 15-83). Mean DIS (s.e.m.) for all patients was 6.0 (0.2). Mean DIS (s.e.m.) for patients whom the gynaecologist classified as having GSI, mixed incontinence and pure urge incontinence was 5.2 (0.3), 8.0 (0.4) and 7.4 (0.5) respectively, when diagnoses were based on urodynamic findings alone, and 5.6 (0.3), 7.9 (0.6), and 7.6 (0.5) when diagnoses were based on both urodynamic and clinical assessment (the clinical diagnosis). We continued further evaluation with the clinical diagnosis alone, dichotomising the patients into those having genuine stress incontinence or not.

We found that the proposed cut-off level for the DIS at 7 resulted in too many false positive findings to be useful as a preoperative tool. In 159 women (64%) having GSI as defined by a cut-off value for the DIS set to 7, we found that 41 of these women (16%) were actually given a false positive diagnosis. This could have been

acceptable for conservative (non-surgical) treatments in primary health care settings, but not for surgical treatment. On the other hand, if the cut-off level was lowered to 5 for the DIS, 112 women (45%) would be diagnosed as having GSI, with only 20 women (8% of 250 women) having a false positive diagnosis. The important issue here is whether these women, if otherwise feasible and indicated, could undergo continence surgery without preoperative urodynamics. Also, this cut-off level was the level most optimal as defined by the ROC-curve, as it was the point nearest to the upper left corner.

Consequently, we concluded that a lower cut-off point than originally proposed was needed for the DIS to become a useful preoperative tool for continence surgery (DIS of 5).

4.2 Paper II

Klovning A, Sandvik H, Hunskaar S. **Web-based survey attracted age-biased sample with more severe illness than paper-based survey.** *J Clin Epidemiol* 2009;62:1068-74.

Aim:

- To analyse how Web-based recruitment performed compared to postal surveys

We recruited 988 respondents (2 months) from www.NettDoktor.no, 708 from www.VG.no (3 days), and 116 from www.startsiden.no (4 months), adding up to a total of 1,812 respondents, mean age 37 vs. 48 years, $P<0.05$. We excluded 36 men, 19 respondents with missing gender information, 38 with missing age information, 99 women below 20 years of age in order to have a sample comparable to the study population in the EPINCONT, and one with apparently nonsensical responses, leaving 1,619 cases for further analysis.

We found that the WEB-EPI UI sample was younger than the EPINCONT sample. The mean age (SD) for the 1,619 women included was 32 (10) years, and the median age was 30 years (range: 20 to 69 years) in the Web-based study. Corresponding figures for the EPINCONT study were: mean age 49 (17) years, and the participants' age ranged from 19 to 98 years. Only 11 women (3.3%) were older than 60 years in the Web-based study compared with 2,396 (29%) in the EPINCONT study. The age group 20 to 29 and 30 to 39 years were highly overrepresented in the

Web study. The crude unadjusted prevalence rate (95% CI) of UI in our study was 20% (18 to 22) (n= 325). Similarly, the crude unadjusted prevalence rate of UI in the EPINCONT study was 25% (24 to 25) (n= 6,876). The mean age (SD) for the 1,294 continent women in our study was 31 (9) years, and 37 (11) years for the 325 incontinent women, compared with 48 (17) and 53 (16) years in the EPINCONT study, respectively. The age-adjusted prevalence of UI in the WEB-EPI UI population higher or similar to, the EPINCONT study for all ages we have reliable data on.

We studied age-adjusted characteristics of the condition among incontinent women in our study compared with data from the EPINCONT study. We found the following statistically significant differences: in the WEB-EPI UI sample, we found fewer women with slight UI in all age groups, and more women with moderate (30 to 39 and 50 to 59-year age groups) and severe UI (20 to 29, 30 to 39 and 40 to 49-year age groups).[5] We found fewer women with stress UI (20 to 29 and 30 to 39-year age groups), more women with urge UI in the two youngest age groups, and more with mixed UI in the 30-39-year age group.

We concluded that we recruited a younger population with more severe UI than the EPINCONT study. Web-based approaches seem to be less appropriate than postal methods for studies of conditions with higher prevalence in the elderly population; and UI is such a condition.

[5] NB! The ”severe” group was erroneously described as 30 to 39, 40 to 49 and 50 to 59 in **Paper II**.

4.3 Paper III

Klovning A, Avery K, Sandvik H, Hunskaar S. **Comparison of two questionnaires for assessing the severity of urinary incontinence: The ICIQ-UI SF versus the Incontinence Severity Index.** *Neurourol Urodyn* 2009;28:411-15.

Aims:

- To validate the International Consultation on Incontinence Questionnaire - Urinary Incontinence Short Form (ICIQ-UI SF) versus the Incontinence Severity Index (ISI),
- To construct a severity scale for the ICIQ-UI SF.

We performed a Web-based comparison of two questionnaires assessing the severity of UI, the ICIQ-UI SF vs. the ISI, using the ISI as the gold standard. Altogether 1,812 women completed the entry questionnaire of the WEB-EPI UI. Of these, 343 (19%) declared having any involuntary urinary leakage, and were subsequently branched into the urinary incontinence arm of the study. Mean age (SD) for these women was 36.5 (11) years and the distribution of stress, urge, mixed and other incontinence was 41%, 17%, 39%, and 3%, respectively. We found no statistically significant differences between corresponding variables from the three different Web sites. All data were therefore analyzed as a whole.

Responses (n= 343) to the ISIitem I assessing frequency were 14% “less than once a month,” 34% “a few times a month,” 34% “a few times a week,” and 18% “every day and/ or night.” Responses to the ISI item II assessing volume were 54%

"drops," 42% "small splashes," and 4% "more." The mean ISI score (SD) was 1.82 (0.70). The mean (SD) ICIQ-UI SF total score was 7.4 (3.6) with, and 4.3 (1.7) without the HRQoL item.

There were strong correlations between the four-level ISI and ICIQ-UI SF scores with versus without the HRQoL item; Spearman's rho was 0.62, $P < 0.01$ versus 0.71, $P < 0.01$. By adjusting the intervals for the ICIQ-UI SF total score for the study subjects in the first scale development file to obtain maximum agreement with the four levels of the ISI, we could define the following intervals for the ICIQ-UI SF (n = 171): slight (1-5), moderate (6-12), severe (13-18), and very severe (19-21) (Kappa with quadratic weighting = 0.61). Similarly, for the ICIQ-UI SF without the HRQoL item, we could define the following levels: slight (1-3), moderate (4-5), severe (6-9), and very severe (10-11), (Kappa with quadratic weighting = 0.71). Applying these intervals to the second sample (n = 172) in order to validate our findings, Kappa with quadratic weighting for ICIQ-UI SF with and without the HRQoL item was 0.61 and 0.74, respectively.

Our findings suggest that the ICIQ-UI SF may be divided into the following four severity categories: slight (1-5), moderate (6-12), severe (13-18) and very severe (19-21) UI. Disregarding the HRQoL-item, the four severity grades would be slight (1-3), moderate (4-5), severe (6-9) and very severe (10-11).

5 Discussion of methods

5.1 Study I (Paper I)

5.1.1 Strengths

Triple blinding

The triple blinding used in this study is one of its strengths, and is described in greater detail in section 2.2.1 "Establishing a gold standard." Although the DIS was incorporated into the structured questionnaire the urotherapist used; the urotherapist had no knowledge of it, and it was not a part of the gynaecologist's work-up either, enabling us to avoid incorporation bias. The urotherapist had no knowledge of the final diagnosis set by the urogynaecologist, and vice versa. In addition, independent researchers analyzed the collected data;[6] the urotherapist or the gynaecologist did not perform any coding of data or analysis initially. By this approach, all effort possible was taken to assure an unbiased, blinded analysis of the data. By triple blinding, we secured the study against work-up bias and diagnostic review bias.

The gold standard

Defining the gold standard is important, and should, if feasible, reflect standard clinical practice. Sometimes gold standard methods may include invasive methods that may be stressful for the patient like e.g. UDIs. More seldom, gold standard

[6] Atle Klovning and Steinar Hunskaar

methods may even be harmful methods; e.g. using contrast media may result in anaphylaxis. The gold standard chosen was an "expert opinion," based on an extensive clinical assessment and urodynamic findings, reflecting secondary care practice as it is, which is important for the external validity of our findings. In **Study I**, we actually had the choice between two gold standards; either the UDIs alone, or the "extensive assessment" or "expert opinion" as they often are termed. In our study, we chose the "expert opinion" method based on the UDI, voiding diaries etc. Neither of them included the structured questionnaire, administered by the urotherapist, nor the DIS. We refer to our gold standard as the "expert opinion," which is often the preferred termed in systematic reviews.

Spectrum bias

It is important to consider whether or not the full spectrum of UI had been examined. In **Study I**, all women had some other kind of UI due to the fact that they were all referred to the specialist clinic for assessment, and they reported different levels of severity. Also, one of the aims of the work-up at the clinic was to assess co-morbid DI, as DI was thought to lead to surgical treatment failure. As we pointed out in **Paper I**, those who did not have GSI had other types of UI.

5.1.2 Limitations

Lack of power calculation

The number of patients included was 250; a large number for a clinical study at the time it was conducted; and probably large enough to ensure statistical power, although this should have been assessed by a pre-study power calculation, using e.g. nomograms or area-under-the-curve (AUC) assumptions.[114-116]

For example, if we anticipate achieving an AUC = 0.8 to define the DIS as a useful tool, we could have calculated the sample size needed. Regrettably, this was not done for **Study I**, and as a rule of the thumb should not be performed post-hoc. For the sake of scientific discussion as part of writing this thesis, I have performed this power calculation post-hoc below.

***Power calculation according to Flahault et al.*[114]**

Given that we wanted to offer surgical treatment to patients scoring 0-5 for the DIS, we should have calculated the number of study subjects needed in the study. For a crude, unadjusted prevalence of 45% (see 4.4, Table 2), looking for a SpPin with a specificity ≥80%, so that a positive test (DIS 0-5) would rule in the women for surgical treatment without prior UDI, and accepting a lower CI no less than 65% with a 95% probability, the number of cases (N_{cases}) needed would be 98 according to Flahault's nomogram.[114] The number of controls ($N_{controls}$) needed in the study would be calculated as $N_{cases} \cdot ((1\text{-prevalence})/\text{prevalence}) = 98 \cdot ((100\text{-}45)/55) = 120$. The total number of patients needed would thus be 98 + 120 = 218. In **Study I**, 250 patients were recruited, which is satisfactory, considering that a 20% loss to follow-up is usually accepted.

Although we did not calculate the 95% CIs for sensitivity and specificity, this has been done later on in a systematic review by Martin et al.[117] Based on the numbers in **Paper I**, the sensitivity (95% CI) and specificity (95% CI) were 0.60 (0.52 to 0.68) and 0.77 (0.67 to 0.85), respectively, so that our assumption of a CI ≥0.65 is catered for.

Carley et al.[116] and Jones et al.[115] have also published relevant methodological papers to assist researchers in determining power calculations for diagnostic studies.

Estimation of number of patients not needing preoperative UDI

For many invasive diagnostic procedures, the probability of having methods that use invasive procedures are often assumed to be gold standard methods. For example, all procedures sparing gastroscopy are highly welcomed. For UI, the same assumptions are valid- procedures that make it possible to spare invasive and costly procedures like UDI are just as welcome. We were able to estimate the number of patients that did not need to undergo preoperative urodynamic investigations in **Study I**; 8% had DI when the DIS was set to 0-5. We were not able to select exactly who these patients were, though.

Phrasing of items in the questionnaire

For sake of clarity for non-Norwegian readers, I have constructed Table 1 to show the minor differences in wordings between the study questionnaire and the wording of the DIS as in the paper published by Kauppila in 1982:[101]

Table 1. This table shows the wording of the DIS as in the paper published by Kauppila in 1982.[101] These questions are not fully formulated in the DIS. The equivalent Norwegian formulation as it occurs in the questionnaire is shown.

	The DIS		***Score***		**Corresponding study questionnaire (Appendix 1)**
Q1	*Feeling of urgency to void before involuntary loss of urine.*			*Q18*	*Like før du får lekkasje, hvordan er da trangen?*
		No	*0*		Føler ingen trang
		Mild	*1*		Vanlig
		Strong	*2*		Sterk eller plagsom
Q2	*Involuntary loss of urine during sudden physical stress*			*Q24*	*Kan du lekke ved plutselig anstrengelse?*
		Yes	*0*		Ja
		Also in other circumstances	*2*		Nei
Q3	*Involuntary loss of urine after physical stress*			*Q25*	*Hvis ja, kommer lekkasjen med det samme anstrengelsen skjer, eller etter noen sekunder*
		Immediately	*0*		Samtidig
		After a few seconds	*2*		Noen sekunder
Q4	*Amount of urine escaped*			*Q42*	*Hvor store mengder urin lekker du vanligvis om gangen?*
		Small	*0*		Dråper
		Moderate	*1*		Små skvetter
		Large	*2*		Større mengder
Q5	*Ability to stop voiding*			*Q36*	*Kan du stoppe strålen helt ved å knipe sammen?*
		Yes	*0*		Ja
		No	*2*		Nei
Q6	*Painful sensation during voiding*			*Q12*	*Har du noen gang svie eller smerte ved vannlating?*
		Yes	*0*		Ja
		No	*2*		Nei
Q7	*Urgent need to void in haste, psychic strain or nervousness*			*Q23*	*Kjenner du sterk trang til å late vannet i vanskelige situasjoner, når du er stresset eller nervøs?*
		No	*0*		Ja
		Mild	*1*		
		Strong	*2*		Nei
Q8	*Frequency of diurnal voiding*			*Q17*	*Hvor mange ganger må du på toalettet om dagen?*
		5 times or less	*0*		Inntil 5 ganger
		6-7 times	*1*		6-7
		8 or more	*2*		8 eller flere
Q9	*Frequency of nocturnal voiding*			*Q16*	*Hvor mange ganger må du på toalettet om natten?*
		0-1	*0*		Ingen, eller 1 gang
		2-3 times	*1*		2 eller 3
		4 or more	*2*		4 eller flere
Q10	*Previous urinary infections necessitating chemotherapy*			*Q06*	*Har du hatt urinvegsinfeksjon eller blærekatarr etter at du fikk problemer med urinlekkasje?*
		0-1	*0*		Nei, eller 1 gang
		2 or more	*1*		2 eller flere
		chronic infection	*2*		Kronisk

In Table 1 it is obvious that the wording of the 10 items of the DIS is up to the researchers to formulate, thus making it difficult to cross-validate different studies using the DIS. It may be that the Finnish researchers meant that these questions were

to be formulated in their own languages by health personnel and researchers using them, so that the precise wording was not necessary, contrary to the PRO-questionnaires, where the exact wording is vital. Since the DIS questions are not fully worded, this is problematic with respect to content validity.[25]

It is problematic that there originally were three options to item 7 of the DIS: "No," "Mild," and "Strong," while our version only used two options: "Yes" or "No." The recoding in SPSS we used for item 7 was like this (from my disk copy dated 30JAN1995):

```
RECODE SP23 (1=2) (2=0) into KAUP07.
FORMATS KAUP07 (N2).
VARIABLE LABELS KAUP07 'Får du sterkt behov for å tømme blæra '+
'når du er nervøs eller stresset?'.
VALUE LABELS KAUP07
0 'Nei'
2 'Ja'.
Missing values KAUP07 (9).
```

The coding is correct as the "Yes" (2 points) or "No" (0 points) are scored reversely in our questionnaire compared to the DIS, but we loose the ability to grade "Mild" or "Strong" as options (1 or 2 points). To clarify this, our study assigned 0 or 2 points to item 7 of the DIS, which originally had 0, 1 or 2 points.

5.2 Study II (Paper II)

5.2.1 Strengths

The gold standard

One of the strengths of **Paper II** is that we were able to compare our results with results from a large epidemiological study on UI, the EPINCONT study, using the results from the EPINCONT as a gold standard. Although postal surveys have their own methodological problems, web-based surveys introduce others.

Our study could not document whether we might get higher response rates from the younger population than when using postal surveys. The largest web-based study we found in our literature search (n=47,859 women) concluded that the bias associated with collecting information using web questionnaires was not greater than that caused by paper questionnaires. This finding was based on a stronger design than we used in our study, as they randomised respondents to either a postal or web-based questionnaire, or a combination, thus being in greater control of bias. The authors concluded that web-based questionnaires may be a feasible tool for data collection in large population-based epidemiological studies in Sweden.[84]

Newcombe's method for comparing proportions and differences between proportions

Comparative analysis of results of the corresponding variables used in the WEB-EPI UI and the EPINCONT studies was done by calculating the 95% CIs with the CIA software,[1] using the Newcombe method for comparing independent proportions.

Single asterisks (*) were placed in Table 1 of **Paper II** to mark where the point estimate of one variable was not an element of the 95% CI of the corresponding variable, thus indicating a statistically significant difference. Double asterisks (**) were placed to mark where the 95% CI for the difference between the independent proportions did not contain zero, indicating a statistically significant difference.

Although this way of comparing two independent samples is the best method we found, this design introduces biases and confounders. A much stronger design would have been to randomise the respondents to postal or web-based questionnaires. The strength of randomisation is that it reduces bias, as all other variables and confounders apart from the intervention would be evenly distributed between the groups.

The study software

We chose to use Inquisite; a commercial solution for Web research, experiencing that branching was one important and advantageous feature exclusive for web-based questionnaires, making it possible to bypass respondents on their way through the questionnaire.

Today, we could have chosen even more sophisticated and powerful solutions like www.SurveyMonkey.com or Open Source solutions, like Joomla! with its enormous amounts of extensions. Joomla! has a front-end and administrator level back-end that is easy to use, with no need for HTML-coding, Perl or CGI-scripting for end-users. It is based on MySQL databases, and many ISPs have preinstalled Joomla! at their Web hotels.

Web versus postal questionnaires

In a randomised comparison of Web versus mailed questionnaires Ritter et al.,[118] studied 397 volunteers randomly assigned to fill in questionnaires online or via paper-and-pencil versions. With this apparently stronger RCT design, they found that out of 16 instruments, none showed statistically significant differences; Web-based test-retest reliability was high, and Web questionnaires required fewer follow-ups to achieve a slightly (non-significant) higher completion rate compared to mailed questionnaires. From my point of view, the ease of constructing Web forms and applications will hopefully lead to an increase in Web-based research.

Anonymity

We used no cookie technology, no IP-address tracing or other efforts to identify the respondents in order to secure privacy. The survey data were safely hosted at UNI•C, The Danish IT Centre for Education and Research.

5.2.2 Limitations

Representativity

One of the limitations of **Study II** was its selection bias- whether our target population was underrepresented on the Web. The women in our Web study were younger and had more UI than in the EPINCONT study, the main finding in **Paper II**, affecting the external validity. UI is a condition that increases with age, and in our study the number of respondents decreases with increasing age. This selection bias is of importance in discussing our findings in **Paper III**.

However, this finding may also contrast the ideas of social desirability bias, where respondents might want to present themselves as “better” than they are. In our

study it could have been that the respondents were be more open/frank about their UI than in the EPINCONT study.

5.3 Study II (Paper III)

5.3.1 Strengths

Correlation strategy

One strength of **Paper III** might be the strategy to use correlation to assess the **relationship** between the ICIQ-UI and the ISI, and then weighted Kappa to assess the degree of **agreement**. Touvier et al.,[119] commented that agreement for continuous variables are best quantified by the ICC, whereas Kappas are best used for categorical variables.

Bland & Altman argued against the use of correlations when comparing two measurements.[120] They argued that *r* measured the strength of a relationship between two variables, not the agreement between them, and that a change in scale of measurement did not affect the correlation, but certainly affected the agreement. Further, correlation depended on the range of the true quantity in the sample. If this was wide, the correlation would be greater than if it was narrow. Also, the authors stated that significance testing was irrelevant to the question of agreement. Finally, data that seemed to be in poor agreement could produce quite high correlations. Bearing these cautions in mind, it seemed scientifically acceptable to use correlation to check for a relationship between the ICIQ-UI SF and the ISI.

In short, since correlation is different from agreement, Bland & Altman recommended the use of the Bland-Altman plot when comparing e.g. two instruments,[120] for example for urodynamic testing. We were not able to use the

Bland-Altman plot, introduced in the Lancet in 1986,[120] since we were not measuring the same scale with two different instruments, and we measured on two different scales, although both measured severity. This is actually a substantial problem with all the different scales that have been developed for assessing UI, and one of the challenges the ICI now has decided to resolve by suggesting that only the ICIQ-UI modular scales should be used. But this was not the case in our study. We therefore had to use Kappa statistics, and used the Vassar Stats weighted Kappa statistics module[113] to enter the 4x4 contingency tables produced by SPSS.

Kappa discussion

Major criticism towards the use of weighted Kappa statistics was early formulated by Malcolm Maclure and Walter C. Willett in the American Journal of Epidemiology, focusing on the misinterpretation and misclassification of the Kappa statistics.[121] They claimed that Kappa was originally proposed to be a measure of agreement between two observers classifying subjects into two nominal categories. The problem arose when Kappa was applied to multicategory classifications, and used not only to assess reproducibility, but also validity. The authors pointed out that for continuous data grouped into ordinal categories for the mere convenience of the researcher, Kappa would be so arbitrary that it would be virtually meaningless. For naturally ordinal data, they claimed that the intraclass correlation coefficient (ICC) was superior to Kappa. And, for polytomous nominal data, the use of several Kappas for different combinations of dichotomies might be more informative than an overall Kappa for the polytomy. Finally, when assessing the validity, the authors pointed to better alternatives than Kappa, e.g. sensitivity, specificity, or positive and negative predictive values for nominal data, or the mean and standard deviation of a new

measurement and the valid reference measurement, or the product-moment (interclass) correlation coefficient (ICC).

Although weighted Kappa was developed to address the limitations of unweighted Kappa values, it has its own potential weakness as it allows weights to be arbitrary in relative magnitude, which means the magnitude of weighted Kappa may be arbitrary. To avoid this arbitrariness, they suggested that standard weights should be used. It turned ought, however, that a logical choice of standard weights in fact converted weighted Kappa equivalent to the intraclass coefficient.[121]

However, Altman seemed to have no objections to the use of weighted Kappa in his well-known textbook ”Practical Statistics for Medical Research,”[122] where Altman points out that where the categories are ordered, as is often the case, it may be preferable to give different weights to disagreements according to the magnitude of the discrepancy.

Weighted Kappa seems to be the right type of approach for our data, and I consider our strategy to use weighted Kappa to assess the agreement a correct one methodologically, due to two issues:

- Both the ISI and ICIQ-UI SF are ordinal scales
- Weighted Kappa is equivalent to the intraclass correlation

In our study, we plotted four levels of the ISI against the ICIQ-UI SF total sum-score with and without the HRQoL dimension. The association between the ISI and ICIQ-UI SF scores was investigated by Spearman’s rank correlation coefficient (rho), and the agreement was assessed by means of weighted Kappa.

Other methods for developing severity grades (Rasch analysis)

Handa and Massof pointed out that psychometric instruments may not necessarily have interval characteristics.[123] For example, numerical values like ”zero” or ”null”

may mean different things to researchers and patients. Secondly, the distance between score 0 to 50 and 50 to 100 may be perceived differently. Thirdly, summed scores may not be meaningful, and change in a summed score may be difficult to interpret.

As an example, they wanted to test the hypothesis that incontinence-related disability is a variable that the grade A questionnaire IIQ with its 30 items would be able to measure. They used Rasch analysis,[124] a technique applying logistic regression analysis based on two mathematical assumptions. Firstly, the response given to any item by each respondent is a function of that individual's disability level, and secondly, to the inherent difficulty of that item. The thinking behind this procedure is that a woman with slight SUI would only report difficulties performing the most strenuous tasks like jumping on a trampoline. By using WINSTEPS[125] they were able to iteratively estimate a scale for the IIQ, thereby demonstrating the spacing and hierarchical ordering of the 30 items comprising this score. After this initial phase, a goodness-of-fit analysis and separation-reliability analysis was performed. Further details of this are described in their paper.[124]

By using the Rasch analysis, the researchers were able to rank and define interval characteristics of the IIQ along a continuum, permitting meaningful comparisons of change, e.g. before and after surgery. Examples of other scored questionnaires that have undergone Rasch analysis are well-known scales like the SF-36 and Beck Depression Inventory.

Albeit Handa and Massof's study being underpowered (n=27), this validation method is interesting and feasible for questionnaires with many items. However, the ISI has two items and the ICIQ-UI SF has three scored items, and I question whether it is necessary for these scored questionnaires to be submitted to Rasch analysis,

having few items, and the fact that the first two items of both assess frequency and volume for both instruments.

5.3.2 Limitations

The concealment procedure

Another weakness with **Study II** was that our concealment procedure reduced the total number of respondents from 1,812 respondents to 343 (19%), namely the responses from women with UI. This inadvertently led to loss of power, especially when we were defining the ”very severe” category, since we randomly split the UI sample into two halves to create samples for developing and validating the scaling of the ICIQ-UI SF.

Skewed target population age

Also, there was a skewed age distribution, our Internet population being younger than in most other epidemiological samples. Consequently, the severity categories we identified would probably not be valid for an elderly population. A study with a higher number of participants is necessary to clarify the ICIQ-UI SF levels for very severe incontinence, since our study had limited statistical power in the category “very severe,” thus affecting the external validity of our findings.

Power calculations

We did not perform any *a priori* power calculation, and this is a weakness with **Study II**, as it was for **Study I**. We did not know the prevalence of UI in a Web-recruited population *a priori*, although we could have assumed it to be the same as in the EPINCONT study, 25%. For the sake of statistical power more women should have

been included, since there were 2 women in the development sample (n=171) and 5 women in the validation sample (n=172) belonging to the "very severe" category.

We should have performed a power calculation to determine the appropriate sample size. We recruited 1,812 respondents, and 343 of these (19%) had UI and were branched to the UI-part of the study. By data splitting into two samples of 171 and 172 respondents, even more power was lost.

According to a posting at the MedStat discussion forum, power calculations for weighted Kappas cannot be done.[7] Instead, we would have to use a confidence interval approach where we model the inputs in the contingency tables, also catering for a skewed distribution. We could have used Vassar Stats[113] for this purpose. For example, we could use the same skewness or distribution of UI severity as we found for the ICIQ-UI SF with the QoL-item in the development sample as in Table II in **Paper III**: 39% slight (63 of 163 women), 54% moderate, 6% severe and 1% very severe UI, and for the ISI the distribution would be 32% mild, 56% moderate, 10% severe and 2% very severe UI. Given that we would consider a 95% CI for Kappa with quadratic weighting to be no wider than 0.20, and not 0.62 (0.30 to 0.92) as in Table II of **Paper III**, we would have to enter the different scenarios into Lowry's web calculator until we achieved an appropriate 95% CI.[113]

To calculate the number of persons in each table cell for a sample size of 500, this is done as follows:

$$(\text{sample size} \cdot \text{row percentage}) \cdot (\text{cell\#/row sum}) = (500 \cdot 0.39) \cdot (38/63) = 118$$

and so forth. If our validation sample size were e.g. set to 500, given the same distribution as in Table II, the result would be as follows:

---

[7] http://groups.google.com/group/MedStats/browse_thread/thread/775b3aaea60cf292

Data Entry

		B								
		1	2	3	4	5	6	7	8	Totals
A	1	118	46			----	----	----	----	164
	2	77	190	12		----	----	----	----	279
	3		34	15		----	----	----	----	49
	4			3	5	----	----	----	----	8
	5	----	----	----	----	----	----	----	----	----
	6	----	----	----	----	----	----	----	----	----
	7	----	----	----	----	----	----	----	----	----
	8	----	----	----	----	----	----	----	----	----
Totals		195	270	30	5	----	----	----	----	500

Reset Calculate

Unweighted Kappa

Observed Kappa	Standard Error	.95 Confidence Interval	
		Lower Limit	Upper Limit
0.3908			
Method 1	0.0376	0.3171	0.4645
Method 2	0.0376	0.3171	0.4645

0.8902 maximum possible unweighted kappa, given the observed marginal frequencies

0.439 observed as proportion of maximum possible

Kappa with Linear Weighting

Observed Kappa	Standard Error	.95 Confidence Interval	
		Lower Limit	Upper Limit
0.4755	0.035	0.407	0.544

0.8292 maximum possible linear-weighted kappa, given the observed marginal frequencies

0.5734 observed as proportion of maximum possible

Kappa with Quadratic Weighting

Observed Kappa	Standard Error	.95 Confidence Interval	
		Lower Limit	Upper Limit
0.5987	0.0914	0.4195	0.7779

0.7387 maximum possible quadratic-weighted kappa, given the observed marginal frequencies

0.8105 observed as proportion of maximum possible

These calculations show that the Kappa with quadratic weighting is 0.60, and the width of the 95% CI is 0.36 (0.42 to 0.78). In order to have a development sample of 500, using the same results as in our study, we would need 1000 women with UI. We found that 19% of our web respondents had UI, and would need a total sample of 1000/0.19 = 5,263 women. Still, we only have 5 women in the "very severe" category of both the ICIQ-UI SF and the ISI.

Similarly, if we recruit twice the number of women, 10,526, we would have 10 women in the "very severe" category of both tests, unweighted Kappa (95% CI) would be 0.39 (0.34 to 0.44), while Kappa with quadratic weighting (95% CI) would be 0.60 (0.47 to 0.73), meaning that the width of the 95% CI is 0.20.

By including 3,000/0.19 = 15,789 women, our validation sample would consist of 1,500 women, with 15 women in the "very severe" category. The unweighted Kappa (95% CI) would be 0.39 (0.35 to 0.43), and the Kappa with quadratic weighting (95% CI) would be 0.60 (0.50 to 0.70). We would then have achieved our aim of a 95% CI no wider than 0.20.

However, if we aim at having 50 women in the "very severe" category, we would need to recruit 10,000/0,19 = 52,631 women. The Kappa with quadratic weighting (95% CI) would then be 0.60 (0.54 to 0.66).

6 Discussion of results

6.1 Paper I

Will this paper tell us how many women may proceed directly to surgery for their GSI without preoperative UDI?

Given a cut-off level set to 3 for the DIS, the LR+ is 11.7, which tells us that it is an ”excellent” test. These 56 women may proceed directly to surgery as the false positive rate, (1-specificity), is only 3% (1-0.97). This is not the same as a failure rate of 3%, probably due to “occult” DI (now termed DO) as termed by Kauppila et al.[101] The clinical challenge is which false positive rate is acceptable. Another way of thinking is that these 56 women would be the ones needing preoperative UDI.

The ROC

Simplifying preoperative assessments for patients with UI has been a challenge for a long time. For busy clinicians, valid and reliable assessment schemes are of great interest, especially if they can replace invasive examinations like urodynamic examinations. The NICE have already proposed that UDIs are not a mandatory procedure before starting many of the treatment options, especially the conservative, non-surgical options.[30]

Using the ROC-curve gives a graphical view of the data. The point that is closest to the upper left (North-Western corner)[108] is the point that gives the most optimal cut-off levels. But, we also have to take into account the risk of the treatment.

Technically, the ROC-curve is just a plot of the true positive rates against the false positive rates, which is the same as plotting the sensitivity against the (1-specificity) rate.

The ROC curve

Using the ROC-curve to display the different cut off values is also a methodological strength, allowing us to discuss the optimal, theoretical cut-off point, versus the clinical challenges of surgical and conservative treatments. In addition, this is regarded as a far better approach than just looking at sensitivity, specificity, PPV, NPV and accuracy. By using the ROC-curve, it is easy to discuss the optimal cut-off level, as it is usually the point closest to the upper left corner of the diagram. Also, the closer that point is to that corner, the greater the test is accuracy of the test is, as measured by the area under the curve (AUC).

SpPins and SnNouts

In Sackett et al.'s book on practising and teaching EBM,[126] two acronyms may be helpful: SpPins and SnNouts. In our discussion in **Paper I**, we actually mixed around these concepts. We wrote: "Using a rule in sensitivity at 0.80, we register this at a DIS cut-off point at 7. A rule out level for the specificity at 0.80 would yield a cut-off level around 5." According to Sackett et al., this should have been formulated as SpPins and SnNouts. The sentence should instead have been phrased like this: " For a cut-off point set to 5 for the DIS, a positive test rules in the patient as not having GSI. For a cut-off point for the DIS set to 7, a negative test rules out the patients as having GSI. In general, in secondary care setting, we need SpPins, while we need SnNouts in primary care settings.

Likelihood ratios – an even better approach to understanding tests?

Likelihood ratios may be calculated as:[108, 126]

LR+ = sensitivity/(1-specificity)

LR- = (1-sensitivity)/specificity

The likelihood ratios indicate by how much a given diagnostic test result would raise or lower the posttest probability of the target disorder. A likelihood ratio of 1.0 means the posttest probability is the exactly the same as the pretest probability, and thus a useless test. The closer a test is to 1, the less useful it is.[126] As a rough guide, I have set up a table on the interpretation suggested by central textbooks on clinical epidemiology and EBM:[108, 126, 127]

Table 3.2.1 Interpreting likelihood ratios

Change in post-test probability	English	Norwegian	LR+	LR-
Large	Excellent	Utmerket	≥10	≤0.1
Moderate	Good	God	5-10	0.1-0.2
Small, but sometimes important	Fair	Middels god	2-5	0.2-0.5
Small, and rarely important	Poor	Dårlig	1-2	0.5-1

Bearing this in mind, so-called excellent tests would be the ones with either an LR+ ≥10 or an LR- ≤0.1, while tests with LR+ between 1 and 2 or LR- between 0.5 and 1 are not useful tests, as they provide almost no change of the post-test probability. In this way, I have chosen terms characterise tests as excellent, good, fair and poor.

JAMA has for many years published a series of papers "The rational clinical exam," where the authors calculate pooled LRs for symptoms, signs and investigations. Two relevant examples are relevant for this thesis: "What type of UI does this woman have?"[128] and "Does this woman have an acute uncomplicated UTI?"[129] Both of these papers demonstrate how it is possible to combine pooled LRs for symptoms, signs and findings in an elegant manner. I would have chosen to

present the likelihood ratios as well if we should write this paper today. The data in the paper have been recalculated and presented as such in the HTA-report.[130] In Table 2, I present how I would have chosen to present the data today.

Table 2. Data from **Paper I** recalculated.

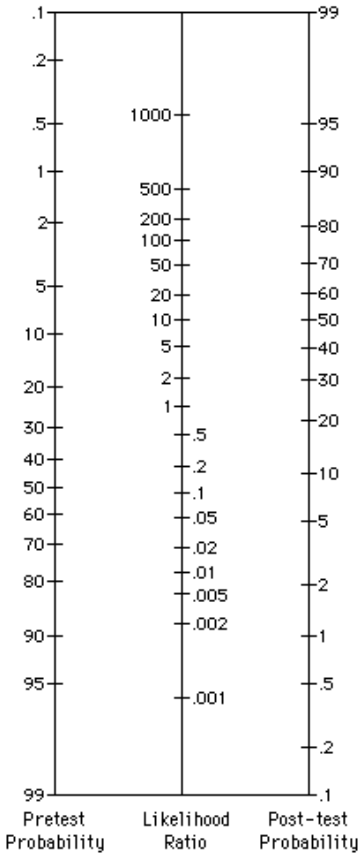
DIS	N=	Sensitivity (%)	Specificity (%)	LR+	LR-
DIS 0-1	16	10	99	10,0	0,9
DIS 0-2	39	25	99	25,0	0,8
DIS 0-3	56	35	97	11,7	0,7
DIS 0-4	78	45	90	4,5	0,6
DIS 0-5	112	60	77	2,6	0,5
DIS 0-6	140	69	59	1,7	0,5
DIS 0-7	159	77	52	1,6	0,4
DIS 0-8	183	85	38	1,4	0,4
DIS 0-9	204	92	26	1,2	0,3
DIS 0-10	219	94	13z	1,1	0,5

The recommended manner of interpreting these LRs is that an excellent test has an LR+ ≥10, a good test 5-10, while a fair test lies in the range of 2-5. Tests between 1-2 are poor, and of no use, as they hardly change the post-test probability. Consequently, this approach shows that the DIS is a poor test if the cut-off level is set to 7. Even so, an LR+ = 2.6 for a cut-off value set to 5 for the DIS only makes it a “fair” test, defining 112 women (44.8%) as having GSI.

Fagan’s nomogram

Fagan’s nomogram[131] is another way of modelling diagnostic reasoning. Most often, working as a GP means handling patients with mostly low-prevalent issues. UI, on the other hand, is actually a high-prevalent condition in general practice (20-25%). This means that positive findings are actually more often true positive than false positive findings, as is not the case in low-prevalent conditions. This means that we should be able to safely diagnose and initiate conservative treatment in general practice.

Fagan's nomogram also enables us to model the sequencing of diagnostic tests. Ideally, these sequential tests should be statistically independent tests. The principle is that that post-test probability of test number one is the pretest probability of test number two, and so forth. In this way, it is possible to increase the post-test probability of a poor or fair to a threshold before e.g. more invasive and potentially harmful treatment. For example, if the pre-test probability of detrusor instability is 10%, and a woman scores DIS = 5, the post-test probability is found by drawing a line through 5% and LR+ = 2.6, yielding a post-test probability of approximately 25%, thereby being a fair test. Similarly, for a woman who scores DIS = 7 (LR+ = 1.6), the post-test probability is about 12%. These figures can be determined more precisely by converting from probabilities to odds ratios and back again, as I show in my Web presentation at http://www.uib.no/isf/people/atle/diagnosis/.



There are many different approaches to presenting findings from diagnostic studies, each of them shedding a different light over the presentation of the results: A systematic review and evaluation of methods of assessing urinary incontinence advocates the use of diagnostic odds ratios, DORs.[117]

External validity

According to Professor David Sackett, external validity is more an issue of **particularising** to the individual, than of **generalising** to all.[108]

The outpatient setting of **Study I** reduces the external validity of the findings in primary health care, and because of this, the prevalence is very high, 45%. This limits the particularising of the findings to primary health care settings. Hunskaar et al. reported in a paper published in 1996[132] that the GP's gatekeeper function results in specialists having a higher prevalence of condition than primary care physicians. The consequence of this is that findings at one level of care are not necessarily valid at the other level.

6.2 Paper II

Which direction does social desirability bias take us in Study II?

Paper II, Fig. 1 shows that the WEB-EPI UI respondents were younger than the EPINCONT respondents, while Fig. 2 shows that they were more incontinent as well. According to the epidemiology section of the 3rd ICI report[24] (p. 280), the prevalence of UI seems to vary over a range of 20%-30% in young women, 30%-40% around menopause and thereafter increasing to 30%-50% in the elderly female population. Albeit this, the prevalence of severe UI seems to range between 6% and 10%.

Criticism against this wide variation in prevalence estimates for UI was raised by Fultz and Herzog,[133] who claimed that this was caused by many different biases where coverage and sampling bias, nonresponse bias, measurement bias and social desirability bias were the most important.

Our major finding was that the Web users we attracted were younger and had more severe incontinence than in the EPINCONT study. It is not surprising that they were younger, as UI increases with increasing age; but the fact that they had more severe UI is surprising. This could be interpreted in two ways. UI is by many considered to be a taboo theme, making it easier for women to state the severity of their incontinence online than in real life, or that the self-selection bias introduced by convenience sampling led to the recruitment of women with more severe UI.

External validity

The findings in **Study II** presented in this paper show that the age bias greatly affects the external validity; the fact that we attracted a younger population with more severe UI than the postal method it was compared with.

Mick Couper discusses the concept of "non-observation," sampling, coverage and non-response as errors of non-observation. In **Paper II**, we had no intention to check for non-observation, or the different aspects of it as Couper states: [88]

> *"While the Internet offers a lot of promise for eHealth research, it also suffers some notable limitations. Key among these are the challenge of drawing representative samples of the population, of dealing with the issue of people without Internet access, and of minimizing the potential for nonresponse bias in Web surveys. On the other hand, the cost of Web surveys relative to alternative modes, the speed with which they can be conducted, and the ability to combine the power of computerised survey instruments with the advantages of self-administration, make this a valuable research tool under certain conditions."*

Although the Web-based sample we recruited was age-biased, thus affecting the external validity of our findings, this does not imply that Web-based methods are not useful for epidemiological studies. Rather, emphasis should be on using the Web as one of several modes of recruiting participants for studies.

6.3 Paper III

First study to grade the ICI recommended PRO-questionnaire: The ICIQ-UI SF

The ICIQ-UI SF has been translated to 38 languages,[25] validated in many studies, and is now the recommended initial assessment tool with validated responsiveness making it useful as an outcome measure as well. Hopefully this will make cross-comparisons between different studies more feasible. This study is the first to propose a scale for the ICIQ-UI-SF. As described in **Paper III**, we were thus able to propose a 4-level severity scale for the ICIQ-UI SF, with and without the HRQoL-dimension:

ICIQ-UI SF	With the QoL item	Without the QoL item
Slight	1-5	1-3
Moderate	6-12	4-5
Severe	13-18	6-9
Very severe	19-21	10-11

HRQoL vs. severity vs. bother

The QoL-concepts of bother and severity need some clarification. It is important to remember that severity says something about the frequency and amount of urinary leakage, while bother is the woman's perception of how UI affects her life. Severity was thus a seemingly objective measure, while bother is a subjective measure. According to the current ICS definition, UI is defined as "any leakage, " and the "bother" item was left out of the definition in 2001. This may be problematic for the ICIQ-UI SF, which has built in the bother scale as a VAS. In **Paper III** we have challenged this, and calculated the correlation between the ISI and the ICIQ-UI SF

with and without the QoL dimension. We found that the correlation was greater without the QoL dimension.

Haltbakk et al. [134] found that patients with prostatic problems were more bothered by irritative than obstructive symptoms. The authors argue that although single-item QoL-scores show a satisfactory responsiveness to change for all treatment methods, and more complex measures show diverging results, they question whether one-item instruments are suitable for research purposes. As QoL often encompasses various aspects like sexual function and quality of sleep, multidimensional instruments would be needed to assess these issues.

If we take a closer look at Fig. 1 in Espuña-Pons and Puig-Clota's paper,[135] we see that the total score is given for the ICIQ-UI SF for women without UI during sexual intercourse is 12.1 (moderate), while it is 14.1 (severe) for women with coital UI. The authors make a point of the fact that it seems like the "bother"-item contributes most to this difference.

However, if item 3 is subtracted, the ICIQ-UI SF scores are 6.2 and 7.4, respectively, thus making both of them "moderate" according to our grading without the QoL-item. What are the consequences of this finding? I am tempted to think that the ICIQ-UI SF actually mixes/blends the "objective" severity grading with the "subjective" bother VAS-item represented as a VAS-score. Both dimensions are important, but putting them into one sum score seems to create some problems. It will be interesting to see how future researchers will cope with this issue.

External validity

In **Paper III** we demonstrated a high correlation between the ISI and the ICIQ-UI SF, and were able to propose a four-level scale for the ICIQ-UI SF. This finding needs reassessment in future studies, as our study had limited power in the "very severe"

group, and limitations due to the age bias found in **Paper II**. However, the Web-based method functioned well for validating the ICIQ-UI SF against the ISI.

7 Conclusions and implications

7.1 Conclusions

7.1.1 Study I (Paper I)

In conclusion, findings from **Study I** in this thesis suggest that lowering the cut-off value for the DIS to 5 is necessary when the choice of treatment is surgery, since fewer patients with false positive findings may be tolerated because they may have other types of UI, compared to non-surgical and harmless treatment. On the other side, treating patients with false positive diagnostic findings by increasing the cut-off value to 7 for harmless treatments like pelvic floor exercises may be deemed acceptable. Using scored diagnostic questionnaires requires balancing test characteristics like sensitivity and specificity or likelihood ratios on one side, and setting, purpose and consequence on the other side. This is essential when considering surgical versus conservative treatment options for UI.

7.1.2 Study II (Paper II and III)

Web-based recruitment for epidemiological studies of conditions functioned well, technologically. But for conditions where the prevalence rises with age, it may be problematic, while it may be useful for studying conditions that affect the younger part of the population. We found that we recruited a younger population with more severe UI in the WEB-EPI UI study than in the EPINCONT study. Using the World Wide Web for validating questionnaires was feasible, and probably all types of self-administered questionnaires may be possible to deploy.

We found that the ICIQ-UI SF may be graded into the following four severity grades: slight (1-5), moderate (6-12), severe (13-18), and very severe (19-21). Without the QoL-item, the grades were slight (1-3), moderate (4-5), severe (6-9), and very severe (10-11).

7.2 Implications for research

This PhD-thesis is based on two studies, with data collected from 1988 to 1992 for **Study I**, and in 2002 for **Study II**. As described in the acknowledgement section, my research has been done as combined work as researcher and GP in several phases, **Study I** in 1993-1996, and **Study II** from 2001-2003, and after moving to Oslo, from 2006 and onwards. Often, PhD-theses are submitted when the first papers are accepted and the third paper is submitted, long before the papers are cited. I therefore thought it would be interesting and relevant to see the implications of my own research, as time has gone in the finalising of this PhD. By using Scholar Google and ISI Web of Knowledge,[8] I found that **Paper I** is cited by ten papers,[117, 128, 136-143] **Paper II** by one,[119] and that **Paper III** is cited by two papers.[135, 144]

7.2.1 Citations: Paper I

1. Amundsen et al. (1999): A criterion validity study

The first paper to cite **Paper I** was a *criterion validity study* by Amundsen et al.[136] published in 1999, validating whether urinary symptoms correlate with video urodynamic findings. This study was not designed to establish a validated questionnaire for urinary incontinence.

The authors claimed that UI is a common problem affecting up to 40% of the female population, a much higher prevalence estimate than the 20-25% large descriptive epidemiological studies like the EPINCONT show. The authors argue that

[8] Search date: 19 May 2010

it is well known that symptoms of urgency incontinence frequently are not easy to demonstrate by urodynamic investigations, and they were unable to identify by the questionnaire which patients would have detrusor instability, and to confirm this with video urodynamics. The authors found instead that patients with urgency incontinence, urgency and frequency rarely had DI.

Citing **Paper I**, the authors found that even using a questionnaire designed to screen for urge symptoms (the DIS), stress incontinence was not accurately predicted and the rate of false-positive results for detrusor instability was high. The authors conclude that in their study, only questions about leaking during activity were helpful in differentiating the aetiology of incontinence, and no questions were helpful in predicting abnormalities other than stress incontinence.

2. Gray et al. (2001): A prediction model for motor UUI

In 2001, Gray et al.[137] developed *a prediction model for motor urge urinary incontinence* to improve the diagnostic strength of a scored questionnaire in detecting urgency or mixed UI. The study researchers used logistic regression analysis to develop the prediction model, based on urodynamic findings and clinical diagnosis in 148 patients. This model was consistent with previous findings, showing that the medical history predicts SUI better than it predicts UUI and MUI.

Paper I is described as part of the authors' *non-systematic review* of relevant literature in their section on "Related Literature," and the authors argued that we observed that a diagnostic instrument must have a high specificity when surgical treatment, with its risk of complications, was used to manage UI. Because the DIS at a cut-off level of 7 yielded a specificity of only 52%, they concluded that it lacked sufficient predictive power to alter a decision to perform surgery in women with UI.

Pursuing this argument, they should have commented that we also studied the consequences of lowering the DIS to 5, turning the DIS into a SpPin.

3. Graham et al. (2002): Non-systematic review

In an *unsystematic review* of the literature from 2002, Graham et al.[138] cites **Paper I** as a validity study of the Detrusor Instability Score (DIS), quantifying the extent of DI symptoms to discriminate stress from urge as the cause of incontinence.

They describe the wording of several items to be clumsy, possibly suffering from translation from Finnish, but nonetheless that it had been validated for use in an English language population. The DIS has been compared with urodynamic findings in two studies by the creators of the score[101, 102] and in **Paper I**, and in both studies; a low DIS-score (0-5) had reasonable predictive value for the absence of a hypertonic bladder or detrusor instability. Although sensitivity and specificity were considered marginal, the DIS has a positive predictive value of 0.82 in an outpatient setting for determining which patients would not demonstrate detrusor instability on urodynamic evaluation. The authors did not present Kujansuu's validation study.[102]

4. Sveen et al. (2004): Quality of life study after stroke

In 2004, a *HRQoL-study* after stroke was published by Sveen et al.,[139] where UI was recorded by a scale for scoring the urological history as in **Paper I**, incorporating the DIS, and administered by a specially trained nurse. The score of the DIS is not presented. The authors concluded that demographic variables and selected impairment tests measuring UI and aphasia did not contribute significantly to explaining well-being after stroke. More important factors were place of residence and leisure activities.

5. Bradley et al. (2005): Development and validation of a new questionnaire

As part of *developing and validating a new scored questionnaire* for UI, the QUID, Bradley et al. [140] refer to the original study by Kauppila[101] and **Paper I**, but not to Kujansuu's validation study.[102]

Apart from describing the test characteristics, their main conclusion was that in the 250 Norwegian urogynaecology patients studied in **Paper I**, the diagnoses of the detrusor instability scores were 66% accurate for stress urinary incontinence and 68% accurate for urge incontinence, and that an English version of the detrusor instability score had not been tested to their knowledge, and reliability characteristics of the survey were unknown. The authors incorrectly cite **Paper I,** as we did not find an accuracy of 68% for urge incontinence. What we showed was that the accuracy for DIS 0-5 was 66% and for DIS 0-7 it was 68%.

6. Brown et al. (2006): A criterion validity study

In 2006, Brown et al.[141] conducted a *criterion validity study* of the 3IQ questionnaire, a simple, quick, and non-invasive test with acceptable accuracy for classifying urge and stress incontinence among middle-aged women, with an extended urogynaecological evaluation as the gold standard. They found that the LR+ and LR- with 95% confidence intervals for the 3IQ compared with the extended evaluation were 3.29 (2.39-4.51) and 0.32 (0.24-0.43) for UUI, and 2.13 (1.71-2.66) and 0.24 (0.16-0.35) for SUI, respectively. They concluded that their findings should be replicated in other primary care clinical settings. In addition, they claimed that clinical outcomes should be assessed in a trial comparing treatments based on the 3IQ and the extended evaluation.

Paper I is one of 7 papers described in Table 5 as one of the published studies evaluating the accuracy of questionnaires to classify type of urinary incontinence in women, and as such is a validation study.

7. Gustafsson et al. (2006): A 3-year cohort study after hysterectomy

Gustafsson et al.[142] published a paper in 2006 using the DIS in a three-year observational cohort study on UI after hysterectomy. **Paper I** was cited as a validation study of the DIS. In this study, the DIS was used preoperatively, postoperatively, after one and three years. They found that total hysterectomy was not associated with increase in UUI or SUI.

8 and 9. Martin et al. (2006): A Cochrane Review and a systematic review

Martin et al. [117] performed a Cochrane review and published a synopsis of this in the N&U.[143] Among the identified 6,009 studies originally found in the literature, 1,479 were duplicates. After reading all the 4,620 abstracts, 490 studies were found to be potentially relevant. After reading the full papers, only 121 studies met the inclusion criteria of the review, and **Study I** was one of the studies that qualified for inclusion. Being the only validation study included, data could of course not be pooled.

10. Holroyd-Leduc et al. (2008): The Rational Clinical Examination

The JAMA-series on "The Rational Clinical Examination" published in 2008 "What type of urinary incontinence does this woman have?"[128] and used findings from **Paper I** in Table 3; Questionnaires used to diagnose urge and stress incontinence. With a cut-off set to 5, the DIS yielded an LR+ and LR- of 2.6 and 0.52, respectively. This paper is actually a clinical scenario to be solved by data from a systematic review of the literature accompanied by a meta-analysis yielding pooled likelihood ratios.

In this paper, the authors argue that in light of the controversy of urodynamics in diagnosing urinary incontinence, they chose to include studies that used urodynamics, expert opinion or both as the gold standard. In order to check whether including expert opinion as part of the gold standard had any effect, they performed a sensitivity analysis by removing altogether 6 of the studies (including **Paper I**) from the meta-analysis, finding that this did not alter the main findings. The authors thus concluded that including these 6 studies using the combined "expert opinion with UDI" did not result in biased conclusions.

7.2.2 Citations: Paper II

1. Touvier et al.[119] A comparison between Web-based and paper versions of a self-administered questionnaire

In this comparison of a Web-based and paper version of a self-administered questionnaire, the NutriNet-Santé anthropometric questionnaire, data concerning 17 questions divided into subquestions (55 variables in all) dealing with height, weight, hip and waist circumferences, weight history, restrictive diet and weight self-perception were collected. Both versions of the questionnaire were filled in by 147 volunteers (paper version first, N = 76, or Web-based version first, N = 71).

Agreement was assessed by intraclass correlation coefficients (ICCs) for continuous variables and Kappas for categorical variables. Agreement between the two versions was high. ICCs ranged from 0.86 to 1.00. Kappas ranged from 0.69 to 1.00 for comparable variables. A total of 82 data entry mistakes (1.5% of total entries), 60 missing values (1.1%), 57 inconsistent values (1.1%) and 3 abnormal values (0.1%) were counted in the paper version (non-existent in the Web-based version due to integrated controls). The Web-based version was preferred by 92.2% of users.

In conclusion, the quality of information provided by the Web-based anthropometric questionnaire used in the NutriNet-Santé Study was equal to, or better than, that of the paper version, with substantial logistic and cost advantages.

A key question in large-scale internet-based studies concerns their capacity to reach a sufficiently broad and diversified population. Although use of computer-based questionnaires may exclude segments of the population without access to, or the capacity to use, computers (notably the elderly [as in **Paper I**]), Internet access is constantly increasing throughout the World.

7.2.3 Citations: Paper III

Paper III has at the time of writing (July 2010) been cited twice, by Espuña-Pons[135] and by Novara et al.[144] The four-level scale we developed for the ICIQ-UI SF will probably be of interest for researchers of future studies using the ICIQ-UI SF. Still, we think it may be in need of reliability testing by other researchers and in different settings like first and secondary care, and for different kinds of diagnostic procedures and treatments. Although authors publish the ICIQ-UI SF total score and its three items, the four-level grading might be more meaningful. As stated in 4.1.3, the ICIQ-UI SF may be divided into the following four severity categories: slight (1-5), moderate (6-12), severe (13-18), and very severe (19-21).

1. Espuña-Pons and Puig-Clota (2009): Cross-sectional study on sexual UI

In a cross-sectional, epidemiological multicentre study on UI during sexual intercourse,[135] Espuña-Pons and Puig-Clota used the King's Health Questionnaire (KHQ) and the Spanish version of the ICIQ-UI SF, and found a prevalence rate of 29.4% in 1,292 sexually active women that had UI during sexual intercourse, and that these women not only had lower quality of life (higher UI-SF total score), but also

greater severity. In Figure 1 of this paper they cite the severity scaling we developed in **Study II** and published in **Paper III**. This figure shows the ICIQ-UI SF total, and also the frequency, quantity and impact items, and conclude that it is the impact item that contributes most to the ICIQ-UI SF total score for these women. They found that women with and without UI during sexual intercourse had an ICIQ-UI total score of 14.1 (severe UI) and 12.1 (moderate UI), respectively. Of women with coital UI vs. not coital UI, more had SUI (38% vs. 28%) and fewer had UUI (17% vs. 27%) ($P<0.001$). Of women with SUI, 36% had coital UI, while only 20% of the women with UUI had coital UI.

2. Novara et al. (2010): 44-month cohort study on radical cystectomy

In a 44-month follow-up study of 113 patients who had undergone radical cystectomy (RC) for bladder cancer and were alive and disease-free,[144] the four-level scale devised in **Paper III** was used to assess the severity of UI. Novara et al. found that 20 (18%) were continent (score 0), and that 32%, 35% and 15% had slight, moderate and severe UI according to the grading we developed as part of **Study II**. None had very severe UI. They also used other validated questionnaires to evaluate LUTS, UI and erectile function, and strongly recommended the use of validated questionnaires after surgery.

As an example, they argued that despite the large number of papers in their research field, none hade ever used validated questionnaires to assess the LUTS function. And, by using previous criteria, like defining continent patients either not using pads/condoms, or in some case series those using a safety pad for occasional leakage. Using these two criteria, 69% and 91% were defined to be continent during daytime, while 25% and 85% were continent during night time. By applying the

Italian ICIQ-UI SF the found fewer to be continent, and judged this to be more realistic.

7.3 Possible implications for future research

The ICI has suggested researchers and clinicians use the ICIQ-UI SF and from the other ICIQ modules for all research on UI; this might imply that the severity grading in **Paper III** might be a paper that could be useful in many future studies. Still, I think there might be other researchers that would be interested in challenging this grading of the ICIQ-UI SF.

Also, many researchers might find the statement on page 368 in the chapter on "Initial Assessment of Urinary and Faecal Incontinence in Adult Male and Female Patients" of the 4th ICI controversial or even provoking.[26] The ICI state that the current Fourth Consultation represents a departure from the recommendation scheme of the previous reviews, and although questionnaires will still be graded A, B, or C, the recommendation is to preferably use questionnaires from the ICIQ modules. Should none of the modular questionnaires be appropriate for research or clinical purposes, this recommendation of the ICI is to use an earlier Grade A questionnaire, or if no suitable instrument exists, a Grade B or C questionnaire.

Findings from the two studies in this thesis have suggested that there has been a need for research that explores the usefulness of simple methods for better assessment of UI. But, with the finalising of the ICIQ-modules, its translation into 38 languages, and in light of the sound validation work that has been undertaken, it is high time that UI researchers acknowledge that no more questionnaires need to be developed in the future, as researchers would now have a common, complete platform

in the ICIQ-modules. Future research would then enable cross-comparisons and meta-analyses.

8 References

1. Gardner SB, Winter PD, Gardner MJ. Confidence Interval Analysis. 1.0 ed. London: BMJ; 1989.
2. Abrams P, Andersson KE, Birder L, Brubaker L, Cardozo L, Chapple C, Cottenden A, Davila W, de Ridder D, Dmochowski R, Drake M, DuBeau C, Fry C, Hanno P, Hay Smith J, Herschorn S, Hosker G, Kelleher C, Koelbl H, Khoury S, Madoff R, Milsom I, Moore K, Newman D, Nitti V, Norton C, Nygaard I, Payne C, Smith A, Staskin D, Tekgul S, Thuroff J, Tubaro A, Vodusek DB, Wein A, Wyndaele JJ. 4th International Consultation on Incontinence. Recommendations of the International Scientific Committee: Evaluation and Treatment of Urinary Incontinence, Pelvic Organ Prolapse and Faecal Incontinence. Paris. 2009.
3. Abrams P, Blaivas JG, Stanton SL, Andersen JT. The standardisation of terminology of lower urinary tract function. The International Continence Society committee on standardisation of terminology. *Scand J Urol Nephrol Suppl* 1988;114:5-19.
4. First report on the standardisation of terminology of lower urinary tract function. *Br J Urol* 1976;48:39-42.
5. First report on the standardisation of terminology of lower urinary tract function. Incontinence, cystometry, urethral closure pressure profile, and units of measurement. *Scand J Urol Nephrol* 1977;11:193-6.
6. Second report on the standardisation of terminology of lower urinary tract function. *Br J Urol* 1977;49:207-10.

7. Second report on the standardisation of terminology of lower urinary tract function. International Continence Society committee on standardisation of terminology, Copenhagen, August 1976. *Eur Urol* 1977;3:168-70.

8. Second report on the standardisation of terminology of lower urinary tract function. Procedures related to the evaluation of micturition: flow rate, pressure measurement. *Scand J Urol Nephrol* 1977;11:197-9.

9. Third report on the standardisation of terminology of lower urinary tract function. Procedures related to the evaluation of micturition: pressure-flow relationships. Residual urine. *Scand J Urol Nephrol* 1978;12:191-3.

10. Third report on the standardisation of terminology of lower urinary tract function. Procedures related to the evaluation of micturition: pressure-flow relationships. Residual urine. Produced by the International Continence Society, February 1977. *Br J Urol* 1980;52:348-50.

11. Third report on the standardisation of terminology of lower urinary tract function. Procedures related to the evaluation of micturition, pressure-flow relationships, residual urine. Produced by the International Continence Society committee on standardisation of terminology, Nottingham, February 1977. *Eur Urol* 1980;6:170-1.

12. Fourth report on the standardisation of terminology of lower urinary tract function. Terminology related to neuromuscular dysfunction of the lower urinary tract. Produced by the International Continence Society. *Br J Urol* 1981;53:333-5.

13. Fourth report on the standardisation of terminology of lower urinary tract function. Terminology related to neuromuscular dysfunction of the lower

urinary tract. The International Continence Society Committee on Standardisation of Terminology. *Scand J Urol Nephrol* 1981;15:169-71.

14. Sixth report on the standardisation of terminology of lower urinary tract function. Procedures related to neurophysiological investigations: electromyography, nerve conduction studies, reflex latencies, evoked potentials and sensory testing. The International Continence Society Committee on Standardisation of Terminology, New York, May 1985. *Int Urol Nephrol* 1986;18:349-56.

15. Abrams P, Blaivas JG, Stanton SL, Andersen J, Fowler CJ, Gerstenberg T, Murray K. Sixth Report on the Standardisation of Terminology of Lower Urinary Tract Function. Procedures related to neurophysiological investigations: electromyography, nerve conduction studies, reflex latencies, evoked potentials and sensory testing. The International Continence Society. *Br J Urol* 1987;59:300-4.

16. Seventh report on the standardisation of terminology of lower urinary tract function: lower urinary tract rehabilitation techniques. International Continence Society Committee on Standardisation of Terminology. *Scand J Urol Nephrol* 1992;26:99-106.

17. Haylen BT, de Ridder D, Freeman RM, Swift SE, Berghmans B, Lee J, Monga A, Petri E, Rizk DE, Sand PK, Schaer GN. An International Urogynecological Association (IUGA)/International Continence Society (ICS) joint report on the terminology for female pelvic floor dysfunction. *Neurourol Urodyn* 2010;29:4-20.

18. Haylen BT, de Ridder D, Freeman RM, Swift SE, Berghmans B, Lee J, Monga A, Petri E, Rizk DE, Sand PK, Schaer GN. An International

Urogynecological Association (IUGA)/International Continence Society (ICS) joint report on the terminology for female pelvic floor dysfunction. *Int Urogynecol J Pelvic Floor Dysfunct* 2010;21:5-26.

19. Foldspang A, Mommsen S. The International Continence Society (ICS) incontinence definition: is the social and hygienic aspect appropriate for etiologic research? *J Clin Epidemiol* 1997;50:1055-60.

20. Holtedahl K, Hunskaar S. Prevalence, 1-year incidence and factors associated with urinary incontinence: a population based study of women 50-74 years of age in primary care. *Maturitas* 1998;28:205-11.

21. Hunskaar S, Arnold EP, Burgio K, Diokno AC, Herzog AR, Mallett VT. Epidemiology and natural history of urinary incontinence. In: Abrams P, Khoury S, Wein A, editors. Incontinence. 1st International Consultation on Incontinence (Monaco 1998). Plymouth: Healh Publication Ltd. 1999. p. 216-7.

22. Hunskaar S, Burgio K, Diokno AC, Herzog AR, Hjälmås K, Lapitan MC. Epidemiology and Natural History of Urinary Incontinence (UI). In: Abrams P, Cardozo L, Khoury S, Wein A, editors. Incontinence. 2nd International Consultation on Incontinence (Paris 2001). Plymouth: Health Publication Ltd. 2002. p. 193-5.

23. Abrams P, Cardozo L, Fall M, Griffiths D, Rosier P, Ulmsten U, van Kerrebroeck P, Victor A, Wein A. The standardisation of terminology of lower urinary tract function: report from the Standardisation Sub-committee of the International Continence Society. *Am J Obstet Gynecol* 2002;187:116-26.

24. Hunskaar S, Burgio K, Clark A, Lapitan MC, Nelson R, Sillen U, Thom D. Epidemiology of Urinary (UI) and Fecal (FI)Incontinence and Pelvic Organ Prolapse. In: Abrams P, Cardozo L, Khoury S, Wein A, editors. Incontinence. 3rd International Consultation on Incontinence (Monaco). Plymouth: Health Publication Ltd. 2005. p. 255-312.

25. Coyne K, Kelleher C. Patient reported outcomes: the ICIQ and the state of the art. *Neurourol Urodyn* 2010;29:645-51.

26. Staskin D, Kelleher C, Avery K, Bosch R, Cotterill N, Coyne K, Emmanuel A, Yoshida M, Kopp Z. Initial Assessment of Urinary and Faecal Incontinence in Adult Male and Female Patients. In: Abrams P, Cardozo L, Khoury S, Wein A, editors. Incontinence. 4th International Consultation on Incontinence (Paris, 2008) 4th ed. Plymouth: Health Publications Ltd. 2009. p. 331-412.

27. Donovan J, Naughton M, Gotoh M. Symptom and quality of life assessment. In: Abrams P, Khoury S, Wein A, editors. Incontinence. 1st International Consultation on Incontinence (Monaco 1998). Plymouth: Health Publication Ltd. 1999. p. 295-332.

28. Donovan JL, Badia X, Corcos J, Gotoh M, Kelleher C, Naughton M, Shaw C, Lukacs B. Symptom and Quality of Life Assessment. In: Abrams P, Cardozo L, Khoury S, Wein A, editors. Incontinence. 2nd International Consultation on Incontinence (Paris, 2001). Plymouth: Health Publication Ltd 2002. p. 267-316.

29. Donovan JL, Bosch R, Gotoh M, Jackson S, Naughton M, Radley S, Valiquette L, Batista JE, Avery K. Symptom and Quality of Life Assessment. In: Abrams P, Cardozo L, Khoury S, Wein A, editors. Incontinence. 3rd

International Consultation on Incontinence (Paris, 2004). Plymouth: Health Publication Ltd. 2005. p. 519-84.

30. NICE. Urinary incontinence (full version). The management of urinary incontinence in women. 1 ed. London, UK: RCOG Press; 2006.

31. NICE. Urinary incontinence (short version). The management of urinary incontinence in women. 1 ed. London, UK: RCOG Press; 2006.

32. NICE. Urinary incontinence (search strategy). The management of urinary incontinence in women. 1 ed. London, UK: RCOG Press; 2006.

33. NICE. Urinary incontinence (quick ref pullout). The management of urinary incontinence in women. 1 ed. London, UK: RCOG Press; 2006.

34. NICE. Urinary incontinence (quick ref). The management of urinary incontinence in women. 1 ed. London, UK: RCOG Press; 2006.

35. Dmochowski RR, Dmochowski R. Editorial Comment. *J Urol* 2009;74:281-2.

36. Digesu GA, Hendricken C, Fernando R, Khullar V. Do Women With Pure Stress Urinary Incontinence Need Urodynamics? *J Urol* 2009:278-81.

37. Avery KN, Bosch JL, Gotoh M, Naughton M, Jackson S, Radley SC, Valiquette L, Batista J, Donovan JL. Questionnaires to assess urinary and anal incontinence: review and recommendations. *J Urol* 2007;177:39-49.

38. Abrams P. A critique of scoring systems. *Prog Clin Biol Res* 1994;386:109-23.

39. Abrams P, Avery K, Gardener N, Donovan J. The International Consultation on Incontinence Modular Questionnaire: www.iciq.net. *J Urol* 2006;175:1063-6.

40. Avery K, Donovan J, Peters TJ, Shaw C, Gotoh M, Abrams P. ICIQ: a brief and robust measure for evaluating the symptoms and impact of urinary incontinence. *Neurourol Urodyn* 2004;23:322-30.

41. Tubaro A, Zattoni F, Prezioso D, Scarpa RM, Pesce F, Rizzi CA, Santini AM, Simoni L, Artibani W. Italian validation of the International Consultation on Incontinence Questionnaires. *BJU Int* 2006;97:101-8.

42. Araki I, Beppu M, Kajiwara M, Mikami Y, Zakoji H, Fukasawa M, Takeda M. Prevalence and impact on generic quality of life of urinary incontinence in Japanese working women: assessment by ICI questionnaire and SF-36 Health Survey. *J Urol* 2005;66:88-93.

43. Espuña Pons M, Rebollo Alvarez P, Puig Clota M. [Validation of the Spanish version of the International Consultation on Incontinence Questionnaire-Short Form. A questionnaire for assessing the urinary incontinence]. *Med Clin (Barc)* 2004;122:288-92.

44. Hashim H, Avery K, Mourad MS, Chamssuddin A, Ghoniem G, Abrams P. The Arabic ICIQ-UI SF: an alternative language version of the English ICIQ-UI SF. *Neurourol Urodyn* 2006;25:277-82.

45. Rotar M, Trsinar B, Kisner K, Barbic M, Sedlar A, Gruden J, Vodusek DB. Correlations between the ICIQ-UI short form and urodynamic diagnosis. *Neurourol Urodyn* 2009:501-5.

46. Seckiner I, Yesilli C, Mungan NA, Aykanat A, Akduman B. Correlations between the ICIQ-SF score and urodynamic findings. *Neurourol Urodyn* 2007;26:492-4.

47. Sandvik H, Hunskaar S, Seim A, Hermstad R, Vanvik A, Bratt H. Validation of a severity index in female urinary incontinence and its implementation in an epidemiological survey. *J Epidemiol Community Health* 1993;47:497-9.

48. Sandvik H, Seim A, Vanvik A, Hunskaar S. A severity index for epidemiological surveys of female urinary incontinence: comparison with 48-hour pad-weighing tests. *Neurourol Urodyn* 2000;19:137-45.

49. Gavira Iglesias FJ, Caridad y Ocerin JM, Perez del Molino Martin J, Valderrama Gama E, Lopez Perez M, Romero Lopez M, Pavon Aranguren MV, Guerrero Munoz JB. Prevalence and psychosocial impact of urinary incontinence in older people of a Spanish rural population. *J Gerontol A Biol Sci Med Sci* 2000;55:207-14.

50. Hannestad YS, Rortveit G, Sandvik H, Hunskaar S. A community-based epidemiological survey of female urinary incontinence: the Norwegian EPINCONT study. Epidemiology of Incontinence in the County of Nord-Trøndelag. *J Clin Epidemiol* 2000;53:1150-7.

51. Hägglund D, Walker-Engström ML, Larsson G, Leppert J. Quality of life and seeking help in women with urinary incontinence. *Acta Obstet Gynecol Scand* 2001;80:1051-5.

52. Rortveit G, Daltveit AK, Hannestad YS, Hunskaar S. Vaginal delivery parameters and urinary incontinence: the Norwegian EPINCONT study. *Am J Obstet Gynecol* 2003;189:1268-74.

53. Rortveit G, Daltveit AK, Hannestad YS, Hunskaar S. Urinary incontinence after vaginal delivery or cesarean section. *N Engl J Med* 2003;348:900-7.

54. Yu HJ, Wong WY, Chen J, Chie WC. Quality of life impact and treatment seeking of Chinese women with urinary incontinence. *Qual Life Res* 2003;12:327-33.

55. Hannestad YS, Lie RT, Rortveit G, Hunskaar S. Familial risk of urinary incontinence in women: population based cross sectional study. *BMJ* 2004;329:889-91.

56. Espuña-Pons M, Guiteras PB, Sampere DC, Bustos AM, Penina AM. Prevalence of urinary incontinence in Catalonia, Spain. *Medicina Clinica* 2009;133:702-5.

57. Whitehead WE, Borrud L, Goode PS, Meikle S, Mueller ER, Tuteja A, Weidner A, Weinstein M, Ye W, Pelvic Floor Disorders N. Fecal Incontinence in US Adults: Epidemiology and Risk Factors. *Gastroenterology* 2009;137:512-7.

58. Anifantaki S, Filiz TM, Alegakis A, Topsever P, Markaki A, Cinar ND, Sofras F, Lionis C. Does urinary incontinence affect quality of life of Greek women less severely? A cross-sectional study in two Mediterranean settings. *Qual Life Res* 2009;18:1311-9.

59. Nygaard I, Barber MD, Burgio KL, Kenton K, Meikle S, Schaffer J, Spino C, Whitehead WE, Wu J, Brody DJ, Pelvic Floor Disorders N. Prevalence of symptomatic pelvic floor disorders in US women. *JAMA* 2008;300:1311-6.

60. Jahanlu D, Qureshi SA, Hunskaar S. The Hordaland Women's Cohort: A prospective cohort study of incontinence, other urinary tract symptoms and related health issues in middle-aged women. *BMC Public Health* 2008;8:10.1186/471-2458-8-296.

61. Daneshgari F, Imrey PB, Risendal B, Dwyer A, Barber MD, Byers T. Differences in urinary incontinence between Hispanic and non-Hispanic white women: a population-based study. *BJU Int* 2008;101:575-9.

62. Chiarelli P. Urinary stress incontinence and overactive bladder symptoms in older women. *Contemporary Nurse* 2007;26:198-207.

63. Diez-Itza I, Ibanez L, Arrue M, Paredes J, Murgiondo A, Sarasqueta C. Influence of maternal weight on the new onset of stress urinary incontinence in pregnant women. *Int Urogynecol J* 2009;20:1259-63.

64. Hendriks EJM, Bernards ATM, Berghmans BCM, de Bie RA. The psychometric properties of the PRAFAB-questionnaire: A brief assessment questionnaire to evaluate severity of urinary incontinence in women. *Neurourol Urodyn* 2007;26:998-1007.

65. Sung VW, Glasgow MA, Wohlrab KJ, Myers DL. Impact of age on preoperative and postoperative urinary incontinence quality of life. *Am J Obstet Gynecol* 2007;197:10.1016/j.ajog.2007.08.076.

66. Arya LA, Banks C, Gopal M, Northington GM. Development and testing of a new instrument to measure fluid intake, output, and urinary symptoms: the questionnaire-based voiding diary. *Am J Obstet Gynecol* 2008;198:10.1016/j.ajog.2008.01.049.

67. Giberti C, Siracusano S, Gallo F, Cortese P, Ciciliato S. Transvaginal bone-anchored sling procedure: 4 years of follow-up on more than 200 consecutive patients. *J Urol* 2008;72:313-7.

68. Barber MD, Spino C, Janz NK, Brubaker L, Nygaard I, Nager CW, Wheeler TL, Pelvic Floor Disorders N. The minimum important differences for the

urinary scales of the Pelvic Floor Distress Inventory and Pelvic Floor Impact Questionnaire. *Am J Obstet Gynecol* 2009;200:10.1016/j.ajog.2009.02.007.

69. Gil KM, Somerville AM, Cichowski S, Savitski JL. Distress and quality of life characteristics associated with seeking surgical treatment for stress urinary incontinence. *Health Qual Life Out* 2009;7:10.1186/477-7525-7-8.

70. Liebergall-Wischnitzer M, Hochner-Celnikier D, Lavy Y, Manor O, Shveiky D, Paltiel O. Randomized Trial of Circular Muscle Versus Pelvic Floor Training for Stress Urinary Incontinence in Women. *J Womens Health* 2009;18:377-85.

71. Sung VW, Joo K, Marques F, Myers DL. Patient-reported outcomes after combined surgery for pelvic floor disorders in older compared to younger women. *Am J Obstet Gynecol* 2009;201:10.1016/j.ajog.2009.07.024.

72. Trabuco EC, Klingele CJ, Weaver AL, McGree ME, Lightner DJ, Gebhart JB. Medium-term comparison of continence rates after rectus fascia or midurethral sling placement. *Am J Obstet Gynecol* 2009;200:10.1016/j.ajog.2008.10.017.

73. Wei J, Nygaard I, Richter H, Brown M, Barber M, Xu X, Kenton K, Nager C, Schaffer J, Visco A, Weber A, Pelvic Floor Disorders N. Outcomes following vaginal prolapse repair and mid urethral sling (OPUS) trial-design and methods. *Clin Trials* 2009;6:162-71.

74. Seim A, Sivertsen B, Eriksen BC, Hunskaar S. Treatment of urinary incontinence in women in general practice: observational study. *BMJ* 1996;312:1459-62.

75. Arya LA, Jackson ND, Myers DL, Verma A. Risk of new-onset urinary incontinence after forceps and vacuum delivery in primiparous women. *Am J Obstet Gynecol* 2001;185:1318-23.

76. Indrekvam S, Sandvik H, Hunskaar S. A Norwegian national cohort of 3198 women treated with home-managed electrical stimulation for urinary incontinence-effectiveness and treatment results. *Scand J Urol Nephrol* 2001;35:32-9.

77. Chiarelli PCJ. Promoting urinary continence in women after delivery: Randomised controlled trial. *BMJ* 2002;324:1241-4.

78. Kerschan-Schindl K, Uher E, Wiesinger G, Kaider A, Ebenbichler G, Nicolakis P, Kollmitzer J, Preisinger E, Fialka-Moser V. Reliability of pelvic floor muscle strength measurement in elderly incontinent women. *Neurourol Urodyn* 2002;21:42-7.

79. Melville JL, Miller EA, Fialkow MF, Lentz GM, Miller JL, Fenner DE. Relationship between patient report and physician assessment of urinary incontinence severity. *Am J Obstet Gynecol* 2003;189:76-80.

80. Hanley J, Capewell A, Hagen S. Validity study of the severity index, a simple measure of urinary incontinence in women. *BMJ* 2001;322:1096-7.

81. Sandvik H, Espuña M, Hunskaar S. Validity of the incontinence severity index: comparison with pad-weighing tests. *Int Urogynecol J Pelvic Floor Dysfunct* 2006;17:520-4.

82. Bell DS, Kahn CE, Jr. Health status assessment via the World Wide Web. *Proc AMIA Annu Fall Symp* 1996:338-42.

83. Eysenbach G, Diepgen TL. Epidemiological data can be gathered with World Wide Web. *BMJ* 1998;316:72.

84. Ekman A, Dickman PW, Klint A, Weiderpass E, Litton JE. Feasibility of using Web-based questionnaires in large population-based epidemiological studies. *Eur J Epidemiol* 2006;21:103-11.

85. Ekman A. The use of the World Wide Web in epidemiological research. Stockholm: Karolinska Instituttet; 2006.

86. Couper MP. Web surveys: a review of issues and approaches. *Public Opin Q* 2000;64:464-94.

87. Couper MP. Web survey design and administration. *Public Opin Q* 2001;65:230-53.

88. Couper MP. Issues of representation in eHealth research (with a focus on Web surveys). *Am J Prev Med* 2007;32:S83-9.

89. Bell DS, Mangione CM, Kahn CE, Jr. Randomized testing of alternative survey formats using anonymous volunteers on the World Wide Web. *J Am Med Inform Assoc* 2001;8:616-20.

90. Birnbaum MH. Human research and data collection via the Internet. *Annu Rev Psychol* 2004;55:803-32.

91. Braithwaite D, Emery J, De Lusignan S, Sutton S. Using the Internet to conduct surveys of health professionals: a valid alternative? *Fam Pract* 2003;20:545-51.

92. Bälter O, Bälter KA. Demands on Web survey tools for epidemiological research. *Eur J Epidemiol* 2005;20:137-9.

93. Bälter KA, Bälter O, Fondell E, Lagerros YT. Web-based and mailed questionnaires: a comparison of response rates and compliance. *Epidemiology* 2005;16:577-9.

94. Berger M, Wagner TH, Baker LC. Internet use and stigmatized illness. *Soc Sci Med* 2005;61:1821-7.

95. Powell J, McCarthy N, Eysenbach G. Cross-sectional survey of users of Internet depression communities. *BMC Psychiatry* 2003;3:19.

96. Reed BD, Crawford S, Couper M, Cave C, Haefner HK. Pain at the vulvar vestibule: a Web-based survey. *J Low Genit Tract Dis* 2004;8:48-57.

97. McCabe SE, Couper MP, Cranford JA, Boyd CJ. Comparison of Web and mail surveys for studying secondary consequences associated with substance use: evidence for minimal mode effects. *Addict Behav* 2006;31:162-8.

98. McCabe SE, Boyd CJ, Couper MP, Crawford S, D'Arcy H. Mode effects for collecting alcohol and other drug use data: Web and U.S. mail. *J Stud Alcohol* 2002;63:755-61.

99. Klovning A, Hunskår S, Eriksen BC, Vanvik A. Spesialistpoliklinikk for utredning av urininkontinens hos kvinner. *Tidsskr Nor Legeforen* 1994;114:3068-70.

100. Dahl FA, Grotle M, Saltyte Benth J, Natvig B. Data splitting as a countermeasure against hypothesis fishing: with a case study of predictors for low back pain. *Eur J Epidemiol* 2008;23:237-42.

101. Kauppila A, Alavaikko P, Kujansuu E. Detrusor Instability Score in the evaluation of stress urinary incontinence. *Acta Obstet Gynecol Scand* 1982;61:137-41.

102. Kujansuu E, Kauppila A. Scored urological history and urethrocystometry in the differential diagnosis of female urinary incontinence. *Ann Chir Gynaecol* 1982;71:197-202.

103. Bø K, Hunskår S, Laake K, Vinsnes A, Steenbuch I. Inkontinens. Om ufrivillig vannlating hos kvinner og menn. Steenbuch I, editor. Oslo: Universitetsforlaget; 1992.

104. Voigt R. [Agreement of anamnestic data with the results of urogynecologic examinations in the diagnosis of urinary incontinence in women]. *Cesk Gynekol* 1985;50:170-4.

105. Ulmsten U. Urininkontinens hos kvinnor. Underlag til vårdprogram. Stockholm: Socialstyrelsen. (Swedish National Board of Health and Welfare)1988.

106. Rosier PF, Gajewski JB, Sand PK, Szabo L, Capewell A, Hosker GL. Executive Summary: The International Consultation on Incontinence 2008-Committee on: "Dynamic Testing"; for urinary incontinence and for fecal incontinence. part 1: Innovations in Urodynamic Techniques and Urodynamic Testing for signs and symptoms of urinary incontinence in female patients. *Neurourol Urodyn* 2009.

107. Dmochowski R. Evaluating the effectiveness of therapies for urinary incontinence. *Rev Urol* 2001;3 Suppl 1:S7-S14.

108. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology. A basic science for clinical medicine. Boston/Toronto/London: Little, Brown and Company; 1991.

109. Walter SD, Sinuff T. Studies reporting ROC curves of diagnostic and prediction data can be incorporated into meta-analyses using corresponding odds ratios. *J Clin Epidemiol* 2007;60:530-4.

110. Simon S. ROC. [WWW] 1999 [updated 2008]; Available from: http://childrensmercy.org/stats/ask/roc.asp.

111. Altman D, Machin D, Gardner MJ. Confidence Interval Analysis. 2.0 ed. London: BMJ Publishing; 2000.

112. Chan YH. Biostatistics 104: correlational analysis. *Singapore Med J* 2003;44:614-9.

113. Lowry R. Cohen's unweighted Kappa. Kappa with linear weighting. Kappa with quadratic weighting. . 2008 [cited 2008 01.02]; Available from: http://faculty.vassar.edu/lowry/kappa.html.

114. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol* 2005;58:859-62.

115. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. *Emerg Med J* 2003;20:453-8.

116. Carley S, Dosman S, Jones SR, Harrison M. Simple nomograms to calculate sample size in diagnostic studies. *Emerg Med J* 2005;22:180-1.

117. Martin JL, Williams KS, Abrams KR, Turner DA, Sutton AJ, Chapple C, Assassa RP, Shaw C, Cheater F. Systematic review and evaluation of methods of assessing urinary incontinence (Structured abstract). Health Technology Assessment [serial on the Internet]. 2006; (6): Available from: http://www.mrw.interscience.wiley.com/cochrane/cldare/articles/DARE-12006008291/frame.html.

118. Ritter P, Lorig K, Laurent D, Matthews K. Internet versus mailed questionnaires: a randomized comparison. *J Med Internet Res* 2004;6:e29.

119. Touvier M, Mejean C, Kesse-Guyot E, Pollet C, Malon A, Castetbon K, Hercberg S. Comparison between web-based and paper versions of a self-administered anthropometric questionnaire. *Eur J Epidemiol* 2010:287-96.

120. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307-10.

121. Maclure M, Willett WC. Misinterpretation and misuse of the Kappa statistic. *Am J Epidemiol* 1987;126:161-9.

122. Altman DG. 14.3 Inter-rater agreement. Practical Statistics for Medical Research: Hapman & Hall/CRC 1991. p. 403-9.

123. Handa VL, Massof RW. Measuring the severity of stress urinary incontinence using the Incontinence Impact Questionnaire. *Neurourol Urodyn* 2004;23:27-32.

124. Bond TG, Fix CM. Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah, N.J.: Jawrence Erlbaum Associates, Inc.; 2001.

125. Linacre JM, Wright BD. WINSTEPS: Multiple-choice, rating scale, and partial credit Rasch ANALYSIS. Chicago: MESA Press; 2001.

126. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine: how to practice and teach EBM. Edinburgh & New York: Churchill Livingstone; 2000.

127. Guyatt G, Rennie D. Users' guides to the medical literature. USA: AMA Press; 2002.

128. Holroyd-Leduc JM, Tannenbaum C, Thorpe KE, Straus SE. What type of urinary incontinence does this woman have? *JAMA* 2008;299:1446-56.

129. Bent S, Nallamothu BK, Simel DL, Fihn SD, Saint S. Does this woman have an acute uncomplicated urinary tract infection? *JAMA* 2002;287:2701-10.

130. Martin JL, Williams KS, Abrams KR, Turner DA, Sutton AJ, Chapple C, Assassa RP, Shaw C, Cheater F. Systematic review and evaluation of methods of assessing urinary incontinence (Structured abstract). Health Technology

Assessment [serial on the Internet]. 2006: Available from: http://www.mrw.interscience.wiley.com/cochrane/clhta/articles/HTA-32006000204/frame.html.

131. Fagan TJ. Letter: Nomogram for Bayes theorem. *N Engl J Med* 1975;293:257.

132. Hunskaar S, Seim A, Freeman T. The journey of incontinent women from community to university clinic; implications for selection bias, gatekeeper function, and primary care. *Fam Pract* 1996;13:363-8.

133. Fultz NH, Herzog AR. Prevalence of urinary incontinence in middle-aged and older women: a survey-based methodological experiment. *J Aging Health* 2000;12:459-69.

134. Haltbakk J, Hanestad BR, Hunskaar S. Use and misuse of the concept of quality of life in evaluating surgical treatments for lower urinary tract symptoms. *BJU Int* 2003;91:380-8.

135. Espuña-Pons M, Puig-Clota M. [Coital urinary incontinence. Associated symptoms and severity of incontinence]. *Actas Urol Esp* 2009;33:801-5.

136. Amundsen C, Lau M, English SF, McGuire EJ. Do urinary symptoms correlate with urodynamic findings? *J Urol* 1999;161:1871-4.

137. Gray M, McClain R, Peruggia M, Patrie J, Steers WD. A model for predicting motor urge urinary incontinence. *Nurs Res* 2001;50:116-22.

138. Graham CW, Dmochowski RR. Questionnaires for women with urinary symptoms. *Neurourol Urodyn* 2002;21:473-81.

139. Sveen U, Thommessen B, Bautz-Holter E, Wyller TB, Laake K. Well-being and instrumental activities of daily living after stroke. *Clin Rehabil* 2004;18:267-74.

140. Bradley CS, Rovner ES, Morgan MA, Berlin M, Novi JM, Shea JA, Arya LA. A new questionnaire for urinary incontinence diagnosis in women: development and testing. *Am J Obstet Gynecol* 2005;192:66-73.

141. Brown JS, Bradley CS, Subak LL, Richter HE, Kraus SR, Brubaker L, Lin F, Vittinghoff E, Grady D. The sensitivity and specificity of a simple test to distinguish between urge and stress urinary incontinence. *Ann Intern Med* 2006;144:715-23.

142. Gustafsson C, Ekstrom A, Brismar S, Altman D. Urinary incontinence after hysterectomy - Three-year observational study. *J Urol* 2006;68:769-74.

143. Martin JL, Williams KS, Sutton AJ, Abrams KR, Assassa RP. Systematic review and meta-analysis of methods of diagnostic assessment for urinary incontinence. *Neurourol Urodyn* 2006;25:674-83.

144. Novara G, Ficarra V, Minja A, De Marco V, Artibani W. Functional Results Following Vescica Ileale Padovana (VIP) Neobladder: Midterm Follow-up Analysis with Validated Questionnaires. *Eur Urol* 2010;57:1054-51.

Appendixes

Appendix 1

The questionnaire used **Study I**. The DIS was embedded into this questionnaire (**Paper I**).

Appendix 2

The EPINCONT questionnaire which was embedded into the Web-based questionnaire used in **Study II** (**Paper II and III**).

Appendix 3

The ICIQ-UI SF questionnaire, which was embedded into the Web-based questionnaire used in **Study II** (**Paper III**).

Appendix 4

The Web-based questionnaire, which was used in **Study II** (**Paper II and III**). Logos, banners, Web pages and the interview published on the Web.