University of MILANO-BICOCCA

STATISTICAL METHODS FOR MASS SPECTROMETRY DATA ANALYSIS AND IDENTIFICATION OF PROSTATE CANCER BIOMARKERS

Tutors: Prof. Rino Bellocco
Prof. Mario Plebani

Andrea Padoan

To my wife Samantha and our son Luca . . .

Abstract

BACKGROUND
Prostate Cancer (PCa) is the most common cancer among males in Europe. Patients developing early PCa sometimes refer non-specific symptoms, namely lower urinary tract symptoms (LUTS), and they usually undergo medical investigations based on Prostate Specific Antigen (PSA) and Digital Rectal Examination (DRE). Suspicious results of one or both testings are prerequisite to Prostate Biopsy. However, due to PSA low sensitivity/specificity in predicting positive prostate biopsy, the identification of new PCa biomarkers is actually a real need. MALDI-TOF/MS protein profiling could be a valuable technology for biomarkers identification. However, up to now its use is laden with lack of reproducibility that confounds scientific inferences and limits its broader use.

AIMS
Goal of this study is to analyze urine collected after prostatic massage in patients referring LUTS, to identify candidate biomarker for PCa, by using MALDI-TOF/MS. We considered important aspects of MALDI-TOF/MS label-free proteomic profiling, in order to assess features reproducibility and to propose appropriate strategy to handle both measurement error and limit of detection (LOD) problems. The study results should aid in reducing the number of worthless first-biopsied and assist Urologists on differential diagnosis of PCa.

METHODS
In a cross-sectional study, we collected urine obtained after DRE from 205 patients that referred LUTS to consultants at the Urological Unit at University of Padova. All patients undergone to prostate biopsy for suspicious PCa. Urines were dialyzed and analyzed by MALDI-TOF/MS in reflectron mode. For the MALDI-TOF/MS reproducibility evaluation, we analyzed a urine pooled from 10 reference samples, spiked with 12.58 pmol of a 1589.9 m/z internal standard (IS) peptide. For the inter-run variability assessment, 14 aliquots were dialyzed by MALDI-TOF/MS. For the intra-run study, an aliquot was divided into 26 separate sub-aliquots and analyzed by MALDI-TOF/MS. To estimate the signal detection limit (sLOD), serial dilution up to 1/256 of a urine pool were analyzed in triplicate. We evaluated the sLOD and adjusted the data appropriately to reduce its variability. We investigated six data normalization approaches - the mean, median, internal standard, relative intensity, total ion current and linear rescaling normalization.

Between-spectrum and the overall spectra variability were evaluated by the coefficient of variation (CV). An optimized signal detection strategy was also evaluated to overcome peak detection algorithms errors. Measurement errors and with-in subject variances were evaluated by an external dataset, made of urine repeatedly collected from 20 reference subjects. Intra class correlation coefficient (ICC), Regression Calibration (RCAL) and SIMEX analyses were used to estimate unbiased logistic regression coefficients relating MALDI-TOF/MS features with Patients biopsy outcome. Monte Carlo simulations were used to estimate influence of different LOD adjustment methods on ICC and RCAL.

RESULTS

Initially, we evaluated the intra- and inter-run on data obtained from automatic peak detection. Normalization methods performed almost similarly in both studies, except IS, which resulted in an increased CV. Calculated sLOD varied with spectra m/z. After sLOD adjustment, raw and normalized data showed a reduction in CVs, while median and mean normalizations performed better, especially in the intra-assay study. However, by optimizing the peak signal detection, the overall features variability drastically decreased. Median normalization with sLOD correction remained the preferable choice for further analyses. Evaluating the external dataset, we found that most of the MALDI-TOF/MS variability is intrinsic to the biological matrix. By using substitution of below LOD values by LOD/2, simulation studies showed that ICC estimations were poorly affected by LOD, when measurement error σ is less that 0.36 and values below LOD are less that 50 %. Comparing results from *naïve* logistic regression, RCAL and SIMEX, measurement error appeared to cause a "bias toward the null". However, SIMEX estimations seemed to correct for a smaller amount of bias than RCAL. Overall, we found eight MALDI-TOF/MS features associated with positive biopsy results.

CONCLUSION

Findings from the reproducibility study showed that the major contributing factor for MALDI-TOF/MS profiling variability is the peak detection process. So, a new algorithm suited for MALDI-TOF reflectron mode is desirable for its applications in profiling studies. However, normalization strategies aid in increasing MALDI-TOF/MS label-free data reproducibility, especially with sLOD correction. Despite urine does not seem to be a promising biological fluid for proteomic biomarker discovers, RCAL and SIMEX appeared valuable approaches to obtain regression coefficients adjusted for biological and instrumental errors on MALDI-TOF/MS features.

Contents

List of Figures

List of Tables

1

INTRODUCTION

In the United States and Europe prostate cancer (PCa) is the most common cancer among males, and is estimated to account for 28% of new cancer cases in US and 22.8% in Europe (1, 2). In Italy prostate cancer represents 14.4% of all the diagnosed cancers and it is the second cause of cancer-related mortality, with 8.1% of all cancer deaths (3). Therefore, annually 238,590 men in the United States and 416,700 men in Europe are estimated to be newly diagnosed with PCa and around 29,720 US and 92,200 European men will probably die from this disease in 2013.

Early PCa stages diagnosis is extremely complicated because this tumor is often indolent. However, patients sometimes suffer of non specific symptoms as weak urinary stream, painful urination or feeling of incomplete voiding, which are shared across some non-cancer conditions like benign prostatic hyperplasia or prostatitis. Nowadays, early detection of PCa relies on serum prostate-specific antigen (PSA) testing and digital rectal examination (DRE). Since its first clinical application, serum PSA has been a valuable tool in the detection, staging and monitoring of this disease. Although the routine use of serum PSA testing has undoubtedly increased PCa detection, one of its main drawbacks has been its lack of specificity resulting in a high negative biopsy rate (4). Moreover, the early detection of many indolent PCa has resulted in treatment of tumors that would not have become life-threatening to a patient. For this reasons, the identification of new biomarkers for prostate cancer diagnosis is actually a real need in the clinical practice. What we really expect from a new biomarker is to improve, without further testing, the “rule in” of testing patients with cancer and the “rule out” of testing patients without cancer. However, a new sensitive and specific biomarker can also ameliorate patient’s management, reducing anxiety and discomfort for really negative patients and enhancing cancer detection rate and cure. Overall these improvements will benefit not only men with lower urinary tract symptoms but also, more generally, to the National Health Care System.

1.1 Prostate cancer risk factors

In a study evaluating the risk of cancer among 44,788 pairs of twins in Sweden, Denmark, and Finland, 42 % of cases of PCa (95 % confidence interval, 29 to 50 %) was attributed to inheritance (5). However, other epidemiological evidences support a major contribution of environmental factors to the development of prostate cancer. Nowadays, the established risk factors for prostate cancer can be subdivided in demographic, life-style and genetic factors.

1.1.1 Demographic and life-style risk factors

Study of age-specific incidence curves of PCa reveals that prostate cancer risk begins to rise sharply after 55 years of age and peaks at age 70–74, declining slightly thereafter. In fact, PCa is diagnosed in very few people aged younger than 50 years ($< 0.1\%$ of all patients). The mean age of patients with this disorder is 72-74 years, and about 85% of patients are diagnosed after age 65 years (6). Autopsy studies confirm that prostate cancer has a long induction period, and that many men have incipient lesions in their 20s and 30s. Moreover, that studies underline that most men aged older than 85 years have histological prostate cancer.

Many epidemiological studies underlined that race/ethnicity are other important risk factors for PCa. For example, it is well known that African American men have higher prostate cancer incidence rates than White men. However, the reasons of this racial difference are not yet completely understood, mainly because this associations could vary by race due to differences in prostate tumor biology, risk factor prevalence, and characteristics between racial groups (5). On this topic, Mordukhovich et al. reviewed 37 epidemiological studies, published between January 1970 and December 2008, that reported PCa race-specific effect estimates between African American and White men. They found no evidences of racial differences in associations between alcohol intake, tobacco use, and family history of PCa and PCa. Further, they found that it was not possible to identify clear patterns among studies evaluating associations between prostate cancer and physical activity, weight/BMI, dietary factors, occupational history, sexual behavior and other health conditions (7).

Regarding the life-style, it is well known that PCa incidence and mortality vary greatly in different geographic regions of the world, with low risks of PCa mortality characteristic of Asia and high risks of PCa mortality characteristic of the US and Western Europe. Furthermore, the fact that incidence rates increase significantly in groups who immigrate to North America indicates that life-style factors may be the major cause of life-threatening PCa in the US. Wilson et al. summarized lifestyle and dietary factors for the prevention of lethal PCa. They showed that obesity is an important risk factor (15% increase in the risk of fatal PCa for each 5 kg/m^2 increase in BMI), while physical activity and smoking have not been consistently associated with PCa incidence. Further, coffee could be associated with a lower risk of lethal prostate cancer (8). Rota et al. performed a meta-analysis on the relationship between alcohol drinking and prostate cancer. They found that the overall relative risk for any alcohol drinking compared with non/occasional drinking was 1.06, and that this risk increase up to 1.08 in heavy alcohol drinker(9).

1.1.2 Genetic risk factors

Epidemiological studies conducted as far back as the 1950s determined that having a first-degree relative (brother or father) with prostate cancer increased risk for an individual by approximately two- to three-fold, on average. Risk is further increased by early age at onset in relatives and multiple relatives with the disease. More recently, segregation studies have identified familial clustering patterns of prostate cancer that are consistent with the presence of high penetrance genetic mutations that confer a Mendelian pattern of inheritance. However, PCa genetic risk factors can be briefly divided epidemiologically into **hereditary** and **sporadic forms**, but these two groups can't be distinguished at a molecular level. In fact, with the exception of *BRCA1/BRCA2* genes and rare cancer predisposition syndromes (hereditary breast ovarian cancer syndrome), highly penetrant inherited genes conferring the prostate cancer phenotype have not been clearly identified (10). Possible candidate genes accounting for inherited prostate cancer were identified as the Hereditary prostate cancer 1 (HPC1)/2'-5'-oligoadenylate (2–5A) dependent ribonuclease L (*RNASEL*), *HPC2/ELAC2*, macrophage scavenger receptor 1 gene (*MSR1*), *BRCA2* and *CHEK2*, the proportion of cases attributable to germline mutations in these loci is small, less than 10% (11). So, it is possible that a big amount of genetic variability is most likely mediated by more common genetic variants or single polymorphisms (SNPs) that have relatively weak effects on prostate cancer risk when singularly considered (12). Polymorphisms involving genes coding for the androgen receptor (AR), 5α-reductase type II and vitamin D receptor have been associated with a variable risk of disease. Other genes for which polymorphisms were found to be associated with an increased risk of prostate cancer include *HSD3B1*, *PSA* and phase I/II enzymes, like *CYPS* (13).

The hereditary prostate cancer genetic controversies have been recently re-opened by Sun et al., who found that the currently established PCa risk-associated SNPs and family history are informative in differentiating an individual's risk for PCa. More interestingly, they claimed that inherited genetics are able to identify men who have considerably elevated risk for PCa (two- and threefold times the population median risk) (14). Many genome wide association studies (GWAS) have assessed the impact of common genomic variation (SNPs) on prostate cancer disease and overall they have allowed to identified >30 common variant alleles that increase prostate cancer risk. Surprisingly, these variants are supposed to explain an estimated 20% to 25% of inherited prostate cancer risk, which is relatively large when compared with breast (5%) and colon cancer (6%) (10, 15). By this studies the region 8q24 has been now identified at least five distinct prostate cancer susceptibility regions, with an increased risk for prostate cancer in men < 50 years (16), but also *KLK3* which encode for PSA and *KLK2* genes have also been found as associated with prostate cancer risk. Another chromosomal region of interest for potentially harboring PCa susceptibility variants is 17q12 and, in particular the polymorphism rs4430796, which increased the risk for PCa in young men

(16).

Epigenetic modifications

In the late 1980s, a functional link was identified between DNA methylation of the cytosine residue of CpG dinucleotides and aberrant transcriptional gene silencing in cancer. This and other epigenetic changes, such as histone modification have been studied in the context of prostate cancer, although DNA hypermethylation probably remains the diagnostically most advanced and biologically most relevant and stable alteration. It is also interesting to note that epigenetic events, and more specifically gene silencing through DNA hypermethylation, are stable, frequent, and abundant. The methylation status of GSTP, the gene encoding the glutathione S-transferase-P$_i$-1 protein involved in detoxification reaction, has been firstly identified as associated with prostate cancer, and today it is the most frequently evaluated epigenetic biomarker for prostate cancer diagnosis (17). Van Neste et al., profusely review the studies that evaluate methylation of GSTP in prostate cancer showing that aberrant methylation is present not only in PCa, but also in its pre-neoplastic legions (PIN). However, the information of DNA methylation is often acquired from sampled tissues, namely an invasive method like prostate biopsy. So, to make biomarkers based on DNA hypermethylation attractive candidates, sampling should be less invasive and more comfortable for patients. Payne et al. evaluated urine and plasma for hypermethylation of some candidate gene by Real Time PCR (18). They found that *RASSF2* measured in urine DNA, achieved 74% sensitivity for PCa with a specificity of 95% for young asymptomatic males and confirmed that GSTP1 hypermethylation can be encountered in up to 60-70% of pre-neoplastic lesions. Moreover, they stated that when evaluating young asymptomatic males as the negative class, measurement of the biomarkers in urine DNA (urine were collected after DRE) was more sensitive than for plasma DNA.

Alternative splicing genes

Alternative splicing, the process by which exons of pre-mRNAs are spliced in different arrangements, plays a major role in the functional diversity of expressed gene transcripts. Alternative splicing arrangements are not evaluable by DNA studies because they are post-trascriptional modifications. After DNA has been trascribed in mRNA, splicing variants of this original mRNA may occurs, generating different proteins isoforms. A number of alternatively spliced genes have been associated with prostate cancer, including *KLK3*, which encode for PSA, member of the kallikrein gene family located at chromosome locus 19q13.3–19q13.4. In fact, it is well known that PSA is present in the serum as a mixture of several molecular species. Other members of the kallikrein gene family (namely *KLK2*, *KLK11*, and *KLK15*) have at least one splice variant, some of which could potentially be used as diagnostic markers because found to be up regulated in prostate cancer (19). Heuze-Vourch et al. studied the protein variants of the PSA gene and they found 12 *hKLK3* transcripts produced by multiple splicing or polyadenylation. Overall, these transcripts code for at least eight proteins, all of them might be found in human sera (20). These "alternative proteins", called PSA related proteins (PSA-RPs) present a conserved N-terminal part of PSA, including the secretion signal peptide and the pro-peptide, suggesting that all the PSA-RPs were synthesized as pre-pro proteins. Some of this PSA-RPs differ in the C-terminal region and some have deletions with respect to the *wild type* PSA protein. Despite that evidences, it is not completely understood yet if this "splicing variants" could be associated with an increase risk in PCa or if they could have an alternative biological function. However, these studies underline how proteomic may have a role in improving PCa diagnosis.

1.2 Usefulness of a Screening program for prostate cancer

The utility of PSA in prostate cancer screening is currently being evaluated in two large clinical trials, the prostate arm of the Prostate, Lung, Colon and Ovary Cancer Screening Trial (PLCO) in the USA and the European Randomized Study of Screening for Prostate Cancer (ERSPC). The PLCO cancer screening trial randomly assigned 76693 men to receive either annual screening with PSA and DRE or standard care as control. After a follow-up of 7 years, the incidence of PCa per 10 000 person-years was 116 (2820 cancers) in the screening group and 95 (2322 cancers) in the control group (rate ratio: 1.22). The incidence of death per 10 000 person-years was 2.0 (50 deaths) in the screened group and 1.7 (44 deaths) in the control group (rate ratio: 1.13). The ERSPC trial included a total of 162 243 men between 55 and 69 years of age. The men were randomly assigned to a group offered PSA screening at an average of once every 4 years or to an unscreened control group. During a median follow-up of 9 years, the cumulative incidence of PCa was 8.2% in the screened group and 4.8% in the control group. The absolute risk difference was 0.71 deaths per 1000 men. This means that 1410 men would need to be screened and 48 additional cases of PCa would need to be treated to prevent 1 death from PCa. However, there is a high risk of patients overtreatment (21).

Currently, subjects over 50 years of age with a life expectancy of at least 10 years or men over age 45 who have at least one first-degree relative (father, brother, or son) with prostate cancer, are considered “at risk”. On the contrary, men with no prostate cancer symptoms with a year life expectancy lower than 10 years, should not be offered testing for prostate cancer since they are not likely to benefit from it because prostate cancer grows slowly. Based on these two large randomized control trials results, most if not all of the major urologic societies have concluded that, at present, widespread mass screening for PCa is not appropriate (21, 22). Therefore, National Public Health System, that might offer screening programs for people “at risk” of prostate cancer, should not offer or encourage a widespread health screening program PSA and DRE based for this type of cancer.

1.3 Biomarkers for prostate cancer

Biological markers (biomarkers) have been defined as "cellular, biochemical or molecular alterations that are measurable in biological media such as human tissues, cells or fluids". More recently, the definition has been broadened to include biological characteristics that can be objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention (23). Therefore, biomarkers can be considered as tools that can help not only in disease prediction, but also in understanding causes, screening, progression or regression of a disease. However, most of biomarkers are usually not recommended for all these topics, although they are widely used by clinicians. PSA is one of the most studied biomarkers and, until now, one of the most debated, probably for its advantages demonstrated in the past. In the late 1980s the introduction of prostate-specific antigen (PSA) testing led to the steepest increase in reported cancer incidence that has ever been observed for any cancer type. In 1991 alone, reported incidence increased by nearly 30% as substantial numbers of men had PSA tests for the first time. First-time and repeat PSA testing continued to increase during the mid-1990s, but since 1992 PCa incidence rates declined steadily for a few years. This decline may be interpreted as PSA testing raised incidence by "pulling forward in time" the date of diagnosis for many prostate cancers, but eventually exhausted the pool of prevalent cancers within reach of its sensitivity. It is noteworthy that since 1996, incidence began to increase again and appears to have resumed the linear trend established before PSA testing. This implies that the risk of "pseudo-disease" (PCa detected by PSA that would otherwise never have been manifested as a diagnosis) is not as large as some have suspected. Since the use of surgery for BPH has declined substantially with the advent of improved drug therapy, the reasons for the persistent upward trend in PCa incidence remain unidentified (24). In this scenario, a number of PSA derivatives, such as percent free PSA (f/t PSA), [-2] proenzyme PSA, PSA density and PSA velocity among others, have been developed in an attempt to address these issues. In fact, there is a clear evidence that many men with very low levels of PSA can harbor prostate cancer (a rate of 6.6% has been indicated in men with a PSA level of $\leq$0.5 ng/ml (25)). Table 1.1 gives the rate of PCa in relation to serum PSA with normal PSA values, as reported from the latest European Association of Urology guidelines (EAU). Moreover, some new molecular test has been developed in the last years, with the aims of improve the diagnostic performance of PSA and DRE.

1.3.1 Prostate specific antigen

The physiological function of PSA protein is to liquefy the clotted semen, in order to facilitate the motility of spermatozoa. Therefore, small amounts of PSA are normally detectable in sera, while PSA levels increase in case of tumor. However, while higher serum PSA levels are often noted in men with prostate cancer, PSA elevation is not specific for prostate cancer causing tested false positive patients. Other factors, particularly benign prostatic hyperplasia (BPH) and prostatitis, cause elevation of serum

PSA level, ng/mL	Risk of PCa, %
0-0.5	6.6 %
0.6-1	10.1 %
1.1-2	17.0 %
2.1-3	23.9 %
3.1-4	26.9 %

Table 1.1: Risk of prostate cancer in relation to low prostate-specific antigen values (21)

PSA and distinguishing PSA elevations due to carcinoma of the prostate from BPH or prostatitis remains problematic. Furthermore, the PSA test sensitivity, achieved for early stage tumors, is low and among patients with non-metastatic prostate cancer only about 50 % is usually correctly identified.

The percentage of free PSA ($free/total$ PSA $\cdot$ 100) has been used to stratify the risk of prostate cancer in men with total PSA levels of 4–10 ng/mL (the so-called grey zone) and a negative DRE. Based on this ratio, a reflex test has been introduced in clinical chemistry laboratories to improve the specificity and sensitivity of tPSA alone (26). And example of application of Reflex testing has been described by Artibani (25, 26). A test positive flow chart is: 1) suspicious DRE or, 2) total PSA of $>$ 10 ng/mL or, 3) total PSA of $\leq$2.5 ng/mL with %free PSA <15% or, 3) total PSA of 2.6–4 ng/mL with % free PSA of <20% or 3) total PSA of 4.1-10 ng/mL with %free PSA <25%. As PSA can change with time, two PSA measurements have been introduced in order to account for this variation: PSA velocity and PSA density. PSA velocity records the change per year while PSA density is calculated by dividing PSA level by the size of the prostate. PSA velocity, on a multivariate analysis including age and PSA, has been shown to significantly improve the ability to detect high-risk prostate cancer unlike PSA density. Nevertheless, a systematic review published in 2009 of 87 papers suggested that there was scant evidence that PSA velocity or PSA density provided predictive information that was better than PSA level alone. However, the current European Association of Urology guidelines state that PSA velocity and PSA density have limited use in the diagnosis of prostate cancer due to background noise (total volume of prostate, BPH), the variations in interval between PSA determinations, and acceleration/deceleration of PSA velocity and PSA density over time (25).

1.3.2 The [-2]proenzyme PSA, PHI and PCA3

For the [-2]proenzyme PSA (p2PSA) and the Beckman Coulter Prostate Health Index (PHI), a mathematical combination of total PSA, free PSA, and p2PSA, some retrospective and prospective studies have suggested that both them may significantly improve the accuracy of total PSA and %free PSA to predict the presence of prostate cancer (25, 27). However, since evidences are still insufficient to assess the utility these new biomarkers, only few laboratories have currently implemented these tests. One of the latest identified markers (1999) is the prostate cancer specific gene *PCA3*. *PCA3* is a noncoding RNA that is highly over-expressed in prostate cancer tissues. A sensitive, urine-based, quantitative test for PCA3 was developed by Gen-Probe Incorporated and released commercially in Europe in 2006 under the brand name PROGENSA® PCA3. Testing for PCA3 involves a collection of urine samples after a DRE is conducted. This gentle pressure on the prostate gland causes the release of prostate juice in the urethra together with some tumour cells if present. Samples are so centrifuged and the level of PSA mRNA and PCA3 mRNA measured; the ratio of the two normalizes for the variable number of prostate cancer cells collected. The sensitivity of the optimal PCA3/PSA ratio has been shown to be 67% and the specificity 83%. Since PCA3 is claimed to be a prostate cancer specific marker, it should not be affected by prostate volume and benign prostatic hyperplasia (BPH) (28) or other non-cancerous prostate conditions such as prostatitis (29). Despite that, the determination of PCA3 still remains experimental because up to now the performance of PCA3 has not been completely validated by independent studies. EAU guidelines note that, at a population level, PCA3 appears to be helpful, but its impact at a single-patient level still remains highly questionable (21).

1.3.3 Genetic markers associated with PSA levels and PCa

Only few studies evaluated the potential clinical information gained from discovered genetic markers associated with early-onset PCa, among entire population or, more interestingly, in high-risk men undergoing prostate cancer screening. Hughes et al. studied 6 genetic markers reported to be associated to early-onset of PCa, showing that only rs6983561 in chromosome 8q24 is informative in predicting time to prostate cancer diagnosis, but only among African American men. In addition, this marker seems to influence PSA prediction for prostate cancer and increases PSA accuracy for longitudinal prediction of prostate cancer (16). However, many PCa related SNPs have shown to be associated with PCa risk. Thus, many researches have started to analyze the triple association between serum PSA levels, SNPs genetic markers and PCa risk. In fact, it would be desirable to enhance this biomarker performance in PCa early detection, including its free form or truncated forms performance. Some important considerations should be taken in advance on analyzing SNPs related with PCa or serum PSA concentrations. In fact, find a further association between one of this candidate SNPs and PCa risk could solely be due to detection bias, generated by PSA based patients accrual, a condition which is difficult to deal with. In fact, because a patient is defined at risk mainly based on his PSA levels, only long prospective cohort studies or Randomized control trials can clearly assess if some particular SNPs will be associated with PSA levels and/or PCa, without falling in biased estimation. Gudmundsson et al., which accrued men aged 50-69 in the ‘Prostate Testing for Cancer and Treatment’ trial (ProtecT), performed a GWAS and found that 6 loci were associated with PSA levels and three of them (10q26, 5p15 and 12q24) were not described yet. Unlike the variants previously identified as associated with PSA in other studies, in the new variants they found, two of the new loci, 12q24 and 10q26, do not associate with prostate cancer risk and the third locus, at 5p15, has only a moderate effect on prostate cancer. Interestingly, they shown that the variants rs10788160-A on 10q26 and rs11067228-A on 12q24 are associated with a greater probability of having a normal prostate biopsied. They concluded that these new markers, rather than predisposition to prostate cancer, primarily predict the outcome of PSA-based prostate cancer screening, that is the decision to perform a biopsy, and the outcome of the biopsy (30). Parikh et al., performed a fine mapping of the KLK3 locus on chromosome 19q13.33 by genotyping tag SNPs in close up 7000 patients draw from five prostate cancer studies. They found that three SNPs were associated with PCa, and that these SNPs were associated with lower serum PSA levels. However, as Authors stated, their findings suggest that these three SNPs are associated with either developing or being diagnosed with nonaggressive PCa, potentially due to differential case identification related to PSA level. Again, accurate study design are necessary to estimate genetic PCa associated locus, due to PSA interference in patients accrual (31).

1.4 Urine as a source of biomarkers

A valuable source for biomarker measurement should be easy to collect, invasiveness and inexpensive. These considerations are specially true when researchers aspire to recommend their new findings for screening of the general population. For example, prostate biopsy is considered the "gold standard" for PCa diagnosis. However, extracting tissue is an unsuitable method for biomarker testing because of its invasiveness, expensiveness and because it may cause patient discomfort and clinical complications. Therefore, testing of disease-related biomarkers in body fluids obtainable in a non-invasive manner is a desirable choice for PCa diagnosis during screening. Urines, e.g., is an optimal source of collection. With respect to blood, they can be more organ-specific. Most of of urinary proteins originates from normal glomerular filtration of plasma proteins (the glomerular barrier only restricts passage of larger proteins, small to middle-molecular-weight proteins can still pass through), but some proteins are generated by secretion of proteins from renal tubular epithelial cells, shedding of whole cells along urinary passage and of apical membranes of renal tubular epithelial cells, and also from exosome secretion. Nevertheless, it must be considered that disruption of the glomerular barrier and/or tubular injury can result in an increased proportion of the plasma proteins in the urine (32). On the contrary, urine protein content can widely vary not only between subjects, but also within the same subject, especially with hydration and disease status. A recent study of Nagaraj et al. has analyzed the daily intra- and inter-variability of urine protein composition. They concluded that the intra-day and inter-subject variabilities are virtually the same (about 45%) and both equally contribute to the total variability observed in the samples sets. Moreover, they found that almost no proteins appeared to be unique to a single person and that Serum albumin, Kininogen-1, prosaposin, zinc-α-2-glycoprotein, Apolipoprotein D are the first four more abundant proteins in urine and account for 40% of total urinary proteins, whereas the top 20 abundant proteins contribute to more than 60% of the proteome. All of these proteins, with the except of Serum Albumin, **are secreted glycoproteins** (33).
Some studies have evaluated the possibility of using urine as a source of new molecular biomarkers PCa-related. Roupret et al., who studied ten genes (GSTP1, RASSF1a,

ECDH1, APC, DAPK, MGMT, p14, p16, RARb2, and TIMP3) using quantitative real-time methylation-specific PCR, found that aberrant methylation can be detected in cells from post-prostate massage voided urine specimens from radical prostatectomy patients with early prostate cancer. The sensitivity in diagnosis of PCa they found, by combining all of the 10 gene loci, was 93% and specificity 74% (34). Meid et al., studied telomerase activity in 36 specimens of cells after prostatic massage in the fresh voided urine of 16 patients who subsequently underwent radical prostatectomy and after urethral washing in 20 who underwent prostate needle biopsies. They found that telomerase activity was present in 14 of 24 samples from patients with prostate cancer (sensitivity 58%), while in the 12 specimens from patients with benign prostatic hyperplasia telomerase activity was negative, confirming no or low enzyme activity in hyperplastic prostatic tissue (35). PCA3, as said above, is another example of biomarker evaluation in urine for PCa detection.

Most authors agree that urine for PCa diagnosis should be collected after prostate massage (4, 35, 36). As collecting this type of samples is more invasiveness than collecting the void urine, the advantages should be carefully evaluated before introducing new tests. However, the PCa specificity of urines collected after DRE is supposed to be greater than which obtainable by blood samples, because during DRE prostate cells (and so malignant cells too) are directly released into the urethra through prostatic ducts, allowing their identification with an appropriate and specific assay, by molecular biology or by proteomic analysis.

1.5 The "proteomic approach"

Most of the physiological changes in cancer are mediated by alterations which occur at DNA level. In fact, it is well know that many cancers share pattern of gene mutations at some *oncogenes* or *onco-suppressor genes* (e.g. *p53, RAS, PTEN, BRCA1, BRCA2*, etc.). Currently, **mRNA expression profiles are often used as surrogates for protein expression**. This will incur little problem for genes that are regulated at the transcriptional level. Many examples in literature reported that mRNAs transcripts (as so called **transcriptomics**) may explain at most 40% of the differential expression of proteins (37). This effect is not surprising. mRNA translation is fine regulated by many intracellular mechanisms, like post-trascriptional regulation, which is the control of gene expression at the RNA level. Moreover, produced proteins may undergo many **post-translation** modification (one of which has been described above as alternative splicing) like phosphorylation or glycosilation. In this scenario, proteomics typically gives us a better understanding of an organism than genomics.

Proteomic studies dating from the 1970s utilized the technique of one or two-dimensional gel electrophoresis to display a large number of proteins from a given cell-line or organism. Following, with the widely introduction of the mass spectrometer instrumentation, researchers switched to a "combinatorial approach", which include a combination of gel electrophoresis with mass spectrometry techniques (Figure 1.1). Up to now, combinato-

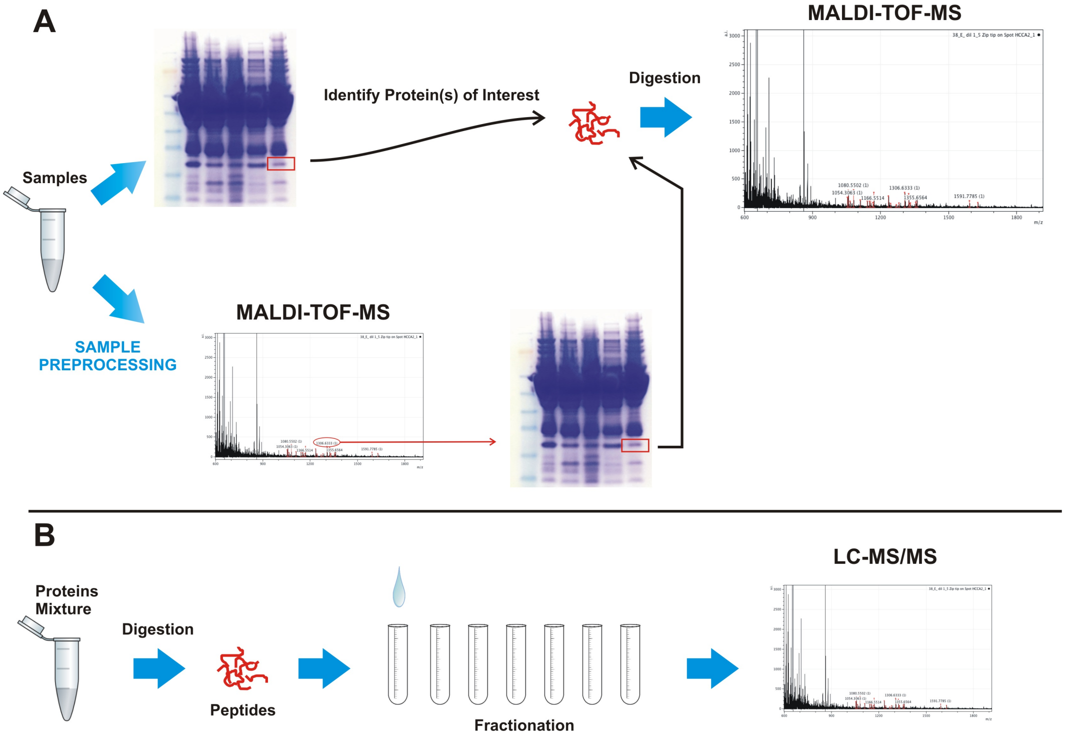

Figure 1.1: The two major approaches used in clinical proteomics. **A**: in the top-down proteomics, whole proteins are prefractionated via various gel and non-gel based techniques and enzymatically digested to obtain fingerprinting for protein identification; **B**: in the bottom-up proteomic, firstly proteins are digested and after non-gel based fractionations of peptides are perfomed to identify proteins. Generally, bottom-up proteomic allows to identify a large number of protein/peptides from a restric number of samples, while the top-down allows to easly analyze and compare a large number of samples.

rial approaches have been utilized for many proteomic analysis, mainly cancer biomarker identification. The combinatorial approach can be subdivided in **top-down** or **bottom-up** methodology. In top-down proteomic, proteins of interest are initially identified in 1D or 2D gel electrophoresis. Following, the gel bands corresponding to the candidate proteins, are excises from gel and in-gel digested by trypsin to retrieve peptides. Therefore samples are evaluated in mass spectrometry, usually MALDI-TOF/MS, to identify protein name. This process is also know as peptide mass fingerprinting (PMF). In bottom-up proteomic, once samples are purified, they are in solution digested by trypsin to obtain peptides; following proteins are fractionated by chromatographic techniques and each fractions analyzed by LC-MS/MS to obtain protein names. The fractionation process allows to detect a larger number of peptides, with respect to that be obtained. Alternatively, candidate proteins can be directly identified by mass spectrometry analysis (a process called **protein profiling**) and therefore evaluated by the in-gel digestion procedure. Protein profiling is high-throughput and allows to analyze and compare a large number of samples from different patient's type.

1.5.1 Urinary proteomic

The urine proteomic biomarker discovery research has undergone to a rapid expansion during the last decade, mostly due to the need of prognostic and diagnostic markers improvement. Urine proteomic analysis is complicated not only by low protein concentration and high salt content (mainly Na^+, K^+ and Ca^{2+}) but also by pH variations, urea or other urine small components (like metabolites) that can interfere with the analysis. Despite that, due to the glomeruli filtration, urines don't normally contain high molecular weight proteins (more than 60 kDa - 70 kDa) and the dynamic range of proteins are not as wide as in plasma/serum (38). In fact, Adachi et al. identified more than 1000 proteins/peptides in the urine, mainly extracellular proteins, plasma membrane proteins, and lysosomal proteins concluding that this high number of protein found is probably due to the more comparable concentration of proteins in the urinary proteome with respect to serum/plasma proteome (38). More interestingly for the disease diagnosis, urine contains many cleaved products of plasma protein that can freely pass the glomeruli to the tubules.

Recently, MALDI-TOF/MS profiling has been used to study the urinary proteome, with different purposes (39, 40, 41, 42, 43, 44, 45, 46, 47, 48). Rehman et al., used a top-down approach for the identification of PCa related proteins. They evaluated urine collected after DRE in 6 PCa and 6 patients with BPH. By using 2D-gel electrophoresis, coupled by MALDI-TOF/MS, they found that calgranulin B/MRP-14 expression was present in a major proportion of patients with prostate cancer with respect to controls (47). M'Koma et al. studied two series of patients, the first composed by 89 men with PCa and 125 negative controls, and the second composed of 103 PCa and 38 control patients; urines were collected before prostate biopsy or by catheter in PCa patients and from normal stream in negative controls. Firstly, they desalted urine by C8 and C18 resins and following elutions were analyzed by MALDI-TOF/MS. Indeed, in order to identify the peptides sequences of the identified features, they used a bottom-up approach: samples were fractionated with HPLC and following analyzed by MALDI-TOF/MS. They found four distinguishing m/z values (1373.1, 1433.5, 2236.3, and 2484.6) able to discriminate between PCa and control patients. By peptide mass fingerprinting, only one of this feature was identified as the semenogelin-I-isoform β-preproprotein (45). Okamoto et al., studied a total of 113 patients, 57 cases with PCa and 56 patients with BPH by using urine collected after DRE. Patients diagnosis was confirmed by prostate biopsy. They don't use specific pre-analysis clean-up or purification steps. Indeed, samples were on-spot purified by ProteinChip technology, which briefly consist on simple affinity chromatography applied directly on the target plate. Finally, peptides/proteins were evaluated by SELDI-TOF. They detected 49 mass peaks that were significantly up-regulated and 23 peaks that were significantly down-regulated, compared with peaks obtained from benign lesion samples. Moreover they performed a hierarchical clustering analysis, reporting they were able to discriminate PC from benign lesions with 91.7% sensitivity and 83.3% specificity. However, they did not performed cross-validation to ensure over-fitting over estimation problems (46). Calvano et al. evaluated middle stream early morning urine from healthy non-smoking individuals and PCa patients,

proved by prostate biopsy. They performed an in deep analysis of pre-treatment strategies reproducibilities for MALDI-TOF/MS proteomics, but also a comparative analysis from case and controls urine, reporting a list of several features, ranging from m/z 931.64 to 2942.48 being different in PCa and control subjectes.
Some studies have also evaluated urine by LS-MS and more than 2000 proteins have been detected in urine using the new generation of LC-MS/MS instruments (36). However, label free LC-MS analysis requires an high level of expertise and is a costly technique. Up to now, nobody have still evaluated urine as source of biomarkers by mass spectrometry labeled methods for quantitative analysis, e.g. by using iTraq technology.

1.6 MALDI technology overview

MALDI is the acronym for Matrix-Assisted Laser Desorption/Ionization, a soft ionization technique that often is improperly confused with the MALDI mass spectrometry (MALDI-MS) instrumentation. Indeed, the MALDI-MS is composed by three main components, namely the Ion source generator, the Mass analyzer and the Detector.

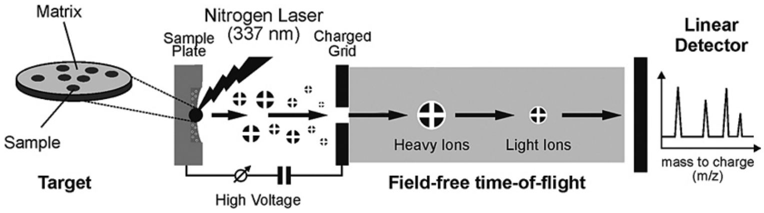

Figure 1.2: MALDI-TOF mass spectrometry analyzer

MALDI and electrospray ionization (ESI) are commonly said soft ionization techniques, that allow the analysis of many kinds of biomolecules like proteins, peptides, sugars and DNA. Ionization is called *soft* because large molecules, when analyzed, don't suffer of fragmentation problems, whose phenomenon happens in conventional ionization techniques that generally break molecules in small tiny parts. ESI and MALDI were developed independently but concurrently, in the first years of the 80s. After the MALDI discovery, people were particularly impressed for the possibility of analyze a wide range of high molecular weight biomolecules. However, what really made the difference with ESI, in particular for the biologists, was the stunning sensitivity which, for the first time, made mass spectrometry compatible with sample preparation techniques used in these fields. For MALDI, the minimum amount of proteins evaluable became rapidly of few femtomoles. Further, MALDI is particularly useful to evaluate biomolecules in solution because often it doesn't necessary require additional pre-treatment sample purification steps. The first step for MALDI-MS analysis requires that samples are spotted on a plan steel surface, named **target plate**, which generally has 384 or more spotting positions. At this point a matrix solution of energy absorption material is mixed together to each sample. Once the sample/matrix mixture are dried, the target plate is placed in the source chamber and an UV laser beam triggers the desorption process, mainly because matrix material heavily absorbs energy from the UV laser light. At this point analyte molecules are ionized in the hot plume. After, charged analytes are accelerated by an electric field of know intensity generated by thin round metallic plaques, called **electric lens**. This electric field can be positive or negative, reflecting the MALDI working condition, said positive or negative mode. This acceleration results in ions with the same charge to have the same kinetic energy. Accelerated ions are therefore evaluated in the mass analyzer. This latter component of the mass spectrometry instrumentation can be of different types, however in MALDI-MS the Time of Flight (TOF) or triple quadrupole (Q) (or combination of them for tandem mass spectrometry) are the most commonly used. The final goal is to detect the abundances and the mass-to-charge (m/z) of all the ionized analytes, which are shown in the mass spectrum, the output of the instrument.

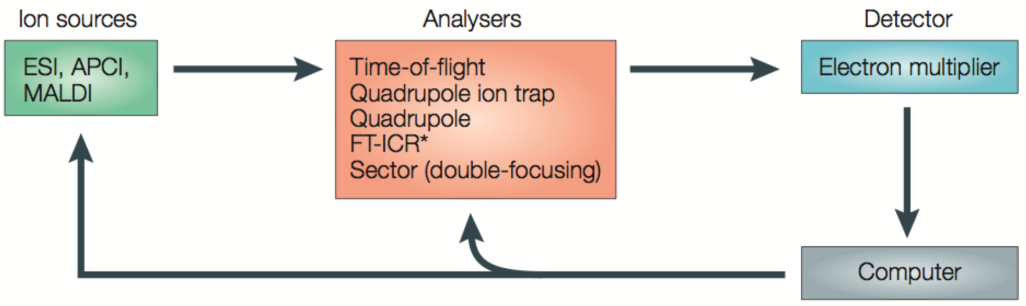

Figure 1.3: Basic components of a typical mass spectrometer.

As MALDI is a relatively young discovery in the mass spectrometry field, many aspects are still being debated. A detailed description of some aspects of this technology are reported following.

1.7 Matrix analytes incorporation, energy absorbtion and sample ablation

Samples can be prepared and spotted by different techniques, namely **dry droplet**, **thin-layer** and **sandwich** methods. However, the most widely used is the dry droplet method, mostly because it allows automation. In a typical dry droplet MALDI sample preparation, a small volumes of sample (about $10\mu L$) containing the analytes and a saturated or semi-saturate solution of the matrix are mixed in equal volumes. The matrix-analyte droplet of typical 1 μL volume is deposited on the target plate and then slowly dried in air. Upon the solvent evaporation, the matrix crystallizes to form a bed of small crystals that sized a few hundred micrometers, depending on the matrix and the preparation method (49). The most common MALDI matrices (2,5-dihydroxybenzoic acid (2,5-DHB), sinapinic acid and 4-hydroxy-α-cyanocinnamic acid (HCCA) and 3-hydroxy-picolinic acid) incorporate the analytes in the crystals quantitatively and in

a homogeneous distribution. However, in some cases surface tension leads to a non homogeneous distribution of the individual crystals and therefore the best performance is achievable only shooting at certain locations, which often requires manual control by the experimenter. A rule of thumb for the identification of this effect is that the addition of the analyte solution should not noticeably change the crystallization behavior of the neat matrix. This effect generally can be avoided by improving sample purification. Another factor that can deeply contribute to crystal formation and therefore to MALDI achievable results is the humidity (50). In particular, Umemura et al., not only found that humidity has a profound effect on sample preparation for the dry droplet method, but also that strictly controlled humidity conditions maximizes the number of peaks found and the overall instrumental reproducibility. They claimed that the best results can be achievable by using 40% humidity.
The further important role of the matrix is to optical absorb laser energy and transfer this energy to the samples. This process is governed by Beer's law (51):

$$H = H_0 \cdot e^{-\alpha z} \tag{1.1}$$

where H is the laser fluence at depth z into the sample, H_0 is the laser fluence at the sample surface, and α is the absorption coefficient. The wavelength-dependent molar absorption coefficient α is a property of the matrix compound. *Viceversa*, the inverse of α is called penetration depth (δ), and has values of only from 20 to 200 nm. It represents the depth into the sample, at which the fluence has decreased to about 30 % of the value at the surface. It is also an order of magnitude estimate of the depth of material ablated (desorbed) per single laser pulse in MALDI. After the laser pulse, energy is transferred more or less uniformly to a macroscopic sample volume and eventually, each laser pulse transfers an amount of energy to the sample, close to the sum of all bond energies in the solid (equivalent to the sum of the heat of fusion and evaporation) causing, for each laser exposure, the desorption/ablation of removal of a bulk volume.

1.7.1 Cationization and adducts formation

During the crystallization process, the solution charge state of the analytes are maintained upon incorporation in the matrix (52). Thus, counter ions (typical basic compounds of matrix or TFA) have to be included in the crystals to account for charge balance. Consequently, cation and anion adducts should show-up in MALDI mass spectra according to the number and polarity of charged groups of the analyte (53). Cationization is a secondary ionization associated process by which analytes acquire a positive charge because associated with metal ions and it is a common phenomenous in MALDI. In fact, cationization by Na$^+$, K$^+$, or other metal cations is usually observed especially for low proton affinity analytes such as carbohydrates and many synthetic polymers (54). Moreover, metal salts are often added to MALDI samples for polymer analysis, in order to enhance the analyte signal. Ion-adduct formation is not restricted to neutral molecules. Peptides/proteins with basic aminoacidic residual can be usually detected in positive ion mode not only as $[M-H]^+$, but also as $[M-H-Cat]^2+$, mainly due to single charged cations adducts formation (e.g., Na$^+$ or K$^+$). The cationization in all likelihood takes place in the expanding plume. It requires a co-desorption of the analyte and the cations, and the prominent effect are obtained from sample locations where both species exist in close neighborhood, such as in the center of DHB-dried or SA-dried droplet preparations (54). Differently, with α-Cyano-4-Hydroxycinnamic acid only a few adducts were observed, presumably mainly arginine residues are complexed (53). However, cation and anion adducts should show-up in MALDI mass spectra according to the number and polarity of charged groups of the analytes even if anionic adducts were difficult to analyze. Especially, the positive adduct formation are believe to depend to the number of basic sites observable both for peptides and proteins (53). Better understanding of the cationization mechanisms should be provided for better control of peptides cationization and ion formation in MALDI, in order to optimize empirical theory on sample preparation procedures (53). Further, the peptide cationization process is important because can significantly reduce the sensitivity in the analysis, partitioning the ion intensity arising from a single peptides into various adduct cluster peaks. Finally, matrix cationization is an important limitation in peptide mass fingerprinting, for the presence of multiple matrix signals due to [Matrix-H-Cation]$^+$ adducts formation.

1.7.2 The Ionization of analytes

In MALDI, since the analytes are embedded in an excess of matrix molecules, the high density of excited matrix molecules results in a rather high rate of energy pooling in the sample. To explain the ionization phenomena two models have been proposed. The older model assumes that matrix incorporated neutral analytes molecules and than, after the photoionization of the matrix molecules, the charge would transferred to the analytes molecules in the plume. Instead, the new model, called **"lucky survivor"**, assumes that proteins are incorporated into the matrix as charged species. This assumption is based on the observation by Krueger et al. in which pH-indicator molecules retain their color and charge state upon crystal incorporation for acidic, neutral, or basic (52). Based on Kruger et al. results, depending on the pH and the dye, the ionic species incorporated vary from protonated to zwitterionic neutral or zwitterionic charged species to negatively charged ions. In this latter model for the most common acidic matrices, peptides will carry a positive excess of charges, counterions being typically either trifluoroacetate or matrix anions. After, the model assumes a desorption of small crystal, some of them with only a single analyte ion. From these clusters many of them are supposed to carry a positive or negative charge, or an excess of a single counterion. Following the desorbtion, crystals are assumed to lose neutral matrix and solvent molecules as well as counterions. This results in a neutralization of the peptide charges except for the only remaining excess charge, mostly one single charge. Differently from the first model, this lucky survival model are able to explain the observation of mostly singly charged ions and for negative and positive ion formation.

1.7.3 The suppression effect in complex samples

The *suppression effect* regards the ion signal of some analytes, that can be suppressed by another analyte present at higher concentration in the specimen under analysis. A reported example is that only a subset of peptides generated by a tryptic digest are observed in a MALDI spectrum (55). Despite the phenomenon is not well understood, the concentration dependence effect may probably be due to direct analytes competition for protonated matrix (55). The suppression effect is particularly important in complex samples, where one/some specific peptides or molecules, present at higher concentration than others, can mask partially or completely the detected intensities of all other ionic species.

1.8 Matrix compounds

The choice of matrix and the sample preparation are two crucial points in MALDI analysis. Unfortunately, there is no single MALDI matrix or sample preparation protocol which is suited to all analytical problems and analytes in MALDI-MS. Matrixes deeply influence not only the ionization behavior but also the fragmentation of the analytes and also have practical implications on the performance of the experiments (56). Almost 20 years after the invention of MALDI by the pioneer work of Tanaka et al., the processes of desorption and ionization have not been fully described yet. As a result, most of the effort spent in search for better matrix has still remained empirical and countless substances have been tested and applied as matrices. However, matrixes should present some features like:

1. solubility in solvents applied for matrix preparation.
2. absorption behavior at the laser wavelength applied.
3. inertness of the matrix and vacuum stability.
4. for low molecular weight compounds potential overlaps of matrix and analyte signals should be avoided.
5. less adducts formation between salts and analytes and salt and matrix.

1.8.1 α-Cyano-4-Hydroxycinnamic acid

This matrix is commonly used for peptides in the lower mass range and it is not soluble in water but well soluble in organic solvents like Acetonytrile. It is supposed to be an high absorbing energy matrix, which means the analyte molecules get a lot of internal energy during desorption and ionization. This leads to a considerable amount of ion fragments in the drift tube (post source decay). The solution for small molecular weight peptides is to work with low laser energy. However, increasing analytes molecular weights, the probability of fragmentation can increases until almost all of the analyte ions undergo

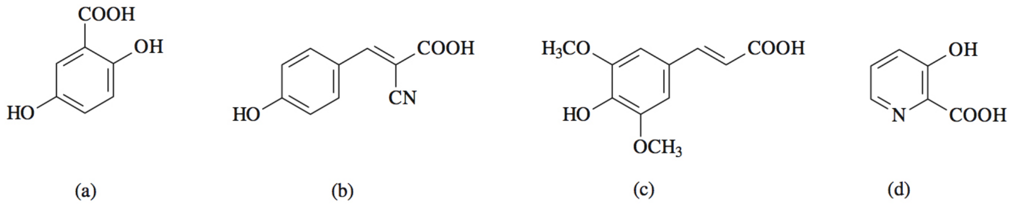

Figure 1.4: Frequently used matrices. a) 2,5-dihydroxybenzoic acid (DHB), b) α-Cyano-4-Hydroxycinnamic acid, c) Sinapinic Acid, d) 3-hydroxy-picolinic acid (3-HPA).

fragmentation. Therefore, α-Cyano is commonly used for reflectron positive ion mode, to analyze peptides with a low molecular weight range. The main advantage of α-Cyano is the ability of this matrix to form small homogenous crystals. Since geometric non homogeneity relates directly to decreased resolution in the MALDI-analysis, α-Cyano preparations usually yield good resolution. Since HCCA is insoluble in water, the samples can be washed on the target (e.g. thin layer sample preparation). It is the most used matrix for low molecular weight peptide profiling and for peptide-mass-fingerprint, generated by enzymatic digestions.

1.8.2 Sinapinic Acid

Sinapinic Acid (SA) is most commonly used in the analysis of high mass proteins. Like HCCA, SA is not soluble in water but well soluble in organic solvents. Compared to α-Cyano it is a less absorbing energy matrix. The analyte ions get less internal energy and the amount of fragmentation is smaller, making this matrix more suitable for measurement of proteins. SA also can form small crystals, which generally have a needle shape. However, it tends to form adducts with the analytes ions.

1.8.3 2,5-dihydroxybenzoic acid

This is the Matrix of choice for the preparation of glycoproteins and glycans, but also for protein analysis. Unlike α-Cyano and Sinapinic Acid it is soluble in water as well as organic solvents. The main disadvantage of 2,5-dihydroxybenzoic acid (DHB) is the fact that it forms big crystal needles (ca. 100 μm sized). This means that the geometry of the sample changes from spot to spot on a preparation. If spectra are summed up from different spots on the sample preparation, the resolution is considerably lower than spectra obtained from an α-Cyano preparation. On a steel target, DHB preparations will form a crystalline ring. Good peptide spectra are usually only obtainable from the rim of that preparation. The main advantage of DHB for MALDI of peptides is the fact that this matrix is more tolerant towards contaminations such as salts and/or detergents than other matrices, because crystals incorporate the proteins, but exclude the majority of common contaminants.

1.8.4 3-hydroxy-picolinic acid

3-hydroxy-picolinic acid (3-HPA) is a water insoluble matrix, usually chosen to analyze the oligonucleotides (DNA) or glycoproteins. 3-HPA is known to be a very soft matrix and is normally used in negative ion model.

1.9 Mass analyzers

There are some different types of mass analyzers (Figure 1.3), which measure charged ions in different ways. These analyzers can be divided into two groups: beam analyzers and trapping analyzers. In beam analyzers, the ions leave the ion source in a beam and pass through the analyzing field until the detector. In trapping analyzers, the ions are trapped in the analyzing field, after being formed in the analyzer itself or being injected from an external ion source (57). On of the most important parameters in MALDI mass spectrometry is the precision of mass measurement, that is strictly related to the resolution, that is, the ability to resolve two adjacent peaks. In general, resolution can be defined as m/Δm, where m is the mass of the peaks being resolved and Δm is the mass difference between the two peaks.

1.9.1 The Time of flight

Conceptually, the simplest mass analyzer is probably the time-of-flight (TOF) (Figure 1.5). A TOF spectrometer separates ions based on their velocity. The ions, formed in the source, are accelerated through a fixed potential (for example, 26 kV) into the TOF drift tube. As all the ions with same charge obtain the same kinetic energy after acceleration, the lower m/z ions achieve higher velocities than the higher m/z ions. In fact, ion velocities are inversely related to the square root of m/z. After the ions are accelerated they travel through a fixed distance, typically 0.5–2.0 meters, before hitting the detector. Thus, by measuring the time, the m/z of the ion can be determined. Normally, the same molecules can hold many different charges (tipically $+2+3, -2$), acquired during the ionization process. This effect, which is more pronounced in ESI mass spectrometry, can be found also in MALDI and is commonly seen in MS data on peptides and proteins. In TOF analyzers, multi charge molecules will have different kinetic energy and therefore the same molecules will be able to reach the detector with different times (multiply charge species). Interestingly, the multiply charge molecules will appear as a mass-to-charge fraction of the original mass, named parental mass or parental ionic species. The process by which the parental mass are found based on the multiply charge detected ions is called **deconvolution**.

1.9.2 Reflectron mode analysis for accurate mass peptides identification

The Time of Flight analysis explained above is normally so called **linear** analysis, while another one, the **mass reflectron** analysis can be used to acquire ions abundances in

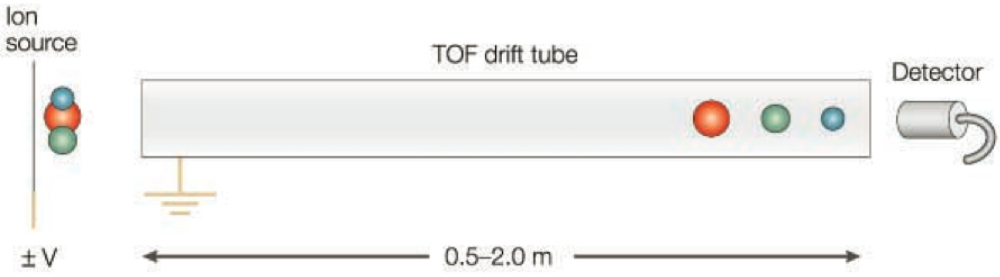

Figure 1.5: Pictorial diagrams of TOF mass analysers. Mass analysis in time-of-flight (TOF) spectrometry is achieved because ions of different mass-to-charge (m/z) values have different velocities and therefore reach the detector at different times.

low molecular weight, typically between 0 and 5000 Da. In the reflectron mode analysis, a ion mirror, place before the detector at the end of the TOF analyzer, is used to reverse the direction of travel of the ions. As result, ions will be forced to enter in a second tube, which is an extension of the TOF analyzer (Figure 1.6). So ions will take more time to reach the reflectron detector. Using the reflectron, the spread of flight times of the ions with different mass-to-charge ratio (m/z) will be augmented, dramatically improving mass resolution, with respect to the linear analysis.

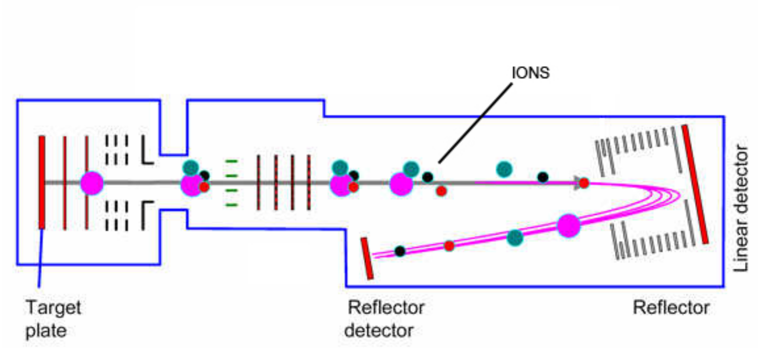

Figure 1.6: Reflectron mode analysis. Ions are accelerated and reversed in direction by the reflectron electric lens. Therefore, reversed ions will hit the reflectron detector rather than linear detector.

1.9.3 The Mass spectrum

The mass spectrum is a plot, obtained by a mass spectrometry analysis, which graphs the *signal intensities vs m/z* (mass-to-charge ratio). Therefore, the mass spectrum of a sample is a pattern representing the distribution of detected ions. The x-axis of a mass spectrum show the relationship between the mass of a given ion and the number of elementary charges that it carries. The y-axis of a mass spectrum show signal intensity of the ions. The intensity of ion current measured by the spectrometer does not accurately represent relative abundance, but strictly correlates with it. So, it is common to label the y-axis with *"arbitrary units"*. In mass spectra, especially when acquired in **reflectron mode**, as individual ions are measured an important result, often overlooked by those not familiar with MS, is the effect of **isotopes**. Each peptide/protein is present not only with a single ion but with multiple ions, each for every isotopes. Another common feature in mass spectrum is the presence of a baseline noise. Therefore, a *peak* can be briefly defined as bunch of signal intensity points that go beyond the noise level, or more precisely as a bunch of signals beyonds a defined signal-to-noise ratio (58, 59). This noise is mainly composed by:

1. Johnson noise due to the electrical system,
2. shot noise or Poisson noise due to the discrete nature of the ion signal
3. chemical noise due to matrix ions, produced during the desorption and ionization of peptides, matrix and impurities in the sample.

. Due to the high efficiency of the electronic device generally included in the instrumentation, the raw mass spectrum contains normally an enormous number of couple of points, mainly more than 100 thousand in a mass range from 1000 to 4000 m/z. However, the number of points in a hypothetical window of the same m/z range decrease when the m/z increase, and therefore the resolution decrease when the m/z increase. In a typical research experimental study, peaks present in the mass spectrum are labeled only with their m/z, **while no information regarding the peptide/protein name are reported, simply because identified peptides/proteins are unknown**.

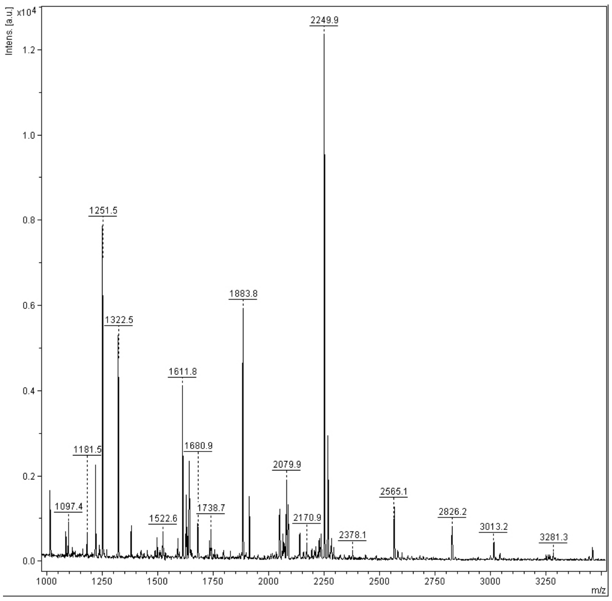

Figure 1.7: Illustrative spectrum obtained from a urine sample analyzed by MALDI-TOF/MS set in positive reflectron mode.

1.10 MALDI-TOF/MS quantitative analysis

In proteomic studies it is generally required to compare the levels of each individual peptide in different experimental conditions or in different disease conditions. Nowadays, many quantification methods are used in proteomic, mainly based on dyes, fluorophores and radioactivity. They have provided very good sensitivity, linearity, and dynamic range, but they suffer from two important shortcomings: first, they require high-resolution protein separation typically provided by 1D or 2D gels, which limits their applicability to abundant and soluble proteins; and second, they do not reveal the identity of the underlying proteins (60). Despite mass spectrometry are currently supposed to be one of the reference methods for analytes quantification in Clinical Chemistry, this notoriety has fallen with the approaching of proteomic studies. In fact, rather that evaluating quantities of a single analytes, proteomic aims to identify as many peptides/proteins as possible, with the maximum accuracy or within an accuracy range. Therefore, different quantification strategies were studied and implemented to overcome the quantification issues in proteomic.
Currently, **relative protein quantification** by MALDI mass spectrometry is based either on measurement of signal intensities for **samples whose proteins are labelled with different stable isotopes tags** (especially by **isobaric tags for relative and absolute quantitation** technology) or on **label-free methods** (60, 61, 62). Differently, **absolute quantification** of single or a few analytes can be successfully performed by adding labeled internal standards, chemically exactly alike the molecules under analysis, except for the isotope tag. As expected, MALDI-TOF/MS quantitation and reproducibility have generated a big debate in the scientific community in the last decade. Both MALDI-TOF/MS labelled and label free analyses have several advantages and disadvantages, which have been profusely reviewed by Benk et al (62). Follow a briefly description.

1.10.1 Labeled methods

The most common used methods for quantification in mass spectrometry proteomic studies are the isobaric tags for relative and absolute quantitation (iTRAQ) technology. The method is based on the covalent labeling of the N-terminus and sidechain amines of peptides from protein digestions with tags of varying mass. Despite the reagents are patented and formulation unknow, some information are available to describe the process. Different tags are used to label different samples/treatments. Then these samples are pooled together (Figure 1.8). The efficiency of the protocol mainly encompass the complete protein digestion and the label efficiency of the reagents. After peptides from the pooled samples are analyzed by tandem mass spectrometry (MS/MS). A database search is then performed using the fragmentation data to identify the labeled peptides and hence the corresponding proteins. The fragmentation of the attached tag generates a low molecular mass reporter ion that can be used to relatively quantify the peptides and the proteins from which they originated. Because tags are of limited number, currently no more than 16-plexes can used. In a typical experiment using 4-plex tag, e.g., up to four different kind of samples can be analyzed, each one generally obtained by pooling samples from the same type. Therefore analyte levels represents the mean analytes levels for each group.

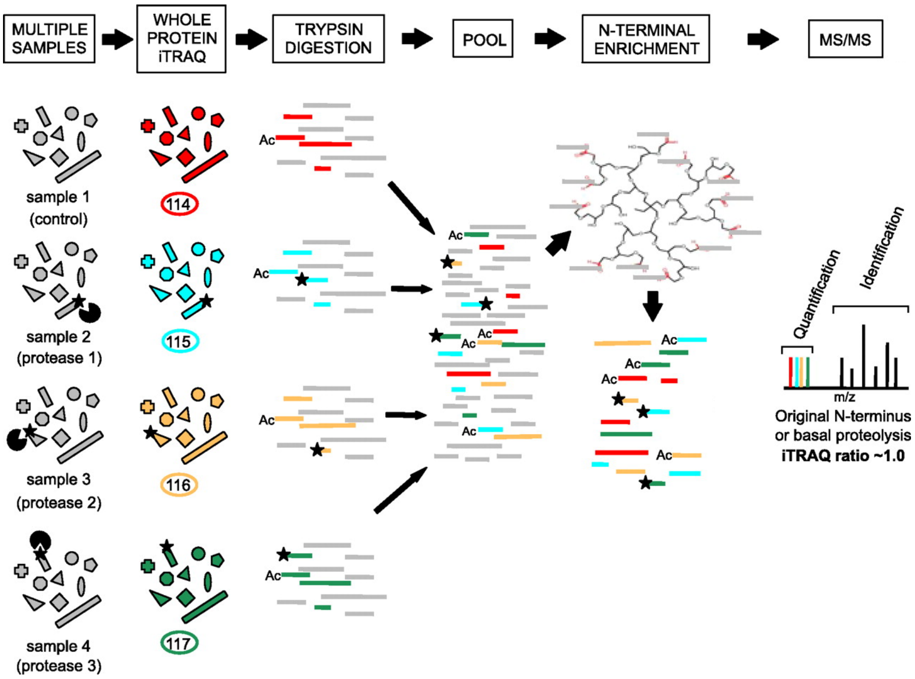

Figure 1.8: Multiple samples can be co-analyzed and compared using iTRAQ. After enzymatic digestion (i.g. Trypsin), peptides in each sample are labeled with distinct iTraq labels (114, 115, 116, 117 reporter ions). This not only distinguishes proteins derived from each sample (indicated by different colors) but can also be used for conventional iTRAQ quantification of the proteins present after trypsinization.

1.10.2 Unlabeled methods

Label free methods (or unlabeled methods) do not use isotopes or tags to quantify analytes. Proteomic profiling using MALDI-TOF/MS is an example of label free analysis. Most of the published studied on MALDI-TOF/MS profiling evaluated the reproducibility of this instrumentation in serum or plasma by different sample pre-treatment. Albrethsen et al. summarized the results reported from different investigators (63). They shown how the peak intensities in intra-experiment reproducibility in MALDI protein profiling vary dramatically between individual protein peaks (2%-40%), and how evaluation studies based on a few selected mass peaks may bias the imprecision estimate downward. On considering the entire spectra and not only a set of reference peaks de Noo et al. calculated the overall inter-measurement CVs. Their findings were consistent to what reported above, ranging from 14% to 23% and was calculated on four different measurements of 4 different days (64). In urinary proteomic, Fiedler et al. evaluated the within-day and between-day reproducibility by using magnetic beads purification and the relative peak intensities of nine characteristic signals of the urine sample, obtaining CVs in the ranges of 1%–14% (within days) and 4%–16% (between days) (41). Benkali et al. evaluated two samples prepared from C2-extract of a healthy volunteer's urine aliquot, spiked with increasing concentrations of three peptides. By a new pre-treatment approach, a solid phase extraction followed by a nanoLC-MALDI-MS analysis, they obtained CVs ranging from 10% to 20% (39). Calvano et al., reported that, using Spin-coater during crystallization, between sample variability is less that which obtained by using the Dry droplet methods (40). Despite the differences were significant, only 10 peaks were used as references for the comparison.

1.11 MALDI-TOF/MS workflow for protein profiling

MALDI-TOF/MS protein profiling consist of a series of experiments developed to identify single or patterns of peptides/proteins which are differently present in different groups of samples. In this scenario, MALDI-TOF/MS offers a valid and rapid method to determine molecular mass of small amounts of peptides and proteins. In fact, as said above, the soft ionization allows intact molecular ions formation, providing accurate determination of polypeptide mass. Non-covalently bound subunits of proteins generally dissociate into individual polypeptide chains, whereas peptide chains connected by covalent bonds such as disulfides remain attached. These consideration underline the strength of MALDI-TOF/MS, because, for many aspects, it can be considered as an analogous to electrophoresis under denaturing conditions without reduction of disulfide bonds (65). As said above, during desorption/ionization of proteins, both positive and negative molecular ions are formed but usually with singly charged ions predominating. The yield of positive ions is greater for most proteins and peptides, and therefore they are usually analyzed in the positive ion mode. Further, peptides and proteins ionize mainly as singly charged protonated molecular ions, despite a small amounts of molecular ions paired with sodium or doubly charged can form. However, a typical MALDI-TOF/MS proteomic profiling workflow is made of many steps (Figure 1.9), which mainly are:

1. Sample collection and storage
2. Sample pre-treatment to remove salts and contaminants
3. Sample deposition on MALDI plate
4. Analysis
5. Data pre-processing
6. Data Normalization
7. Statistical data analysis

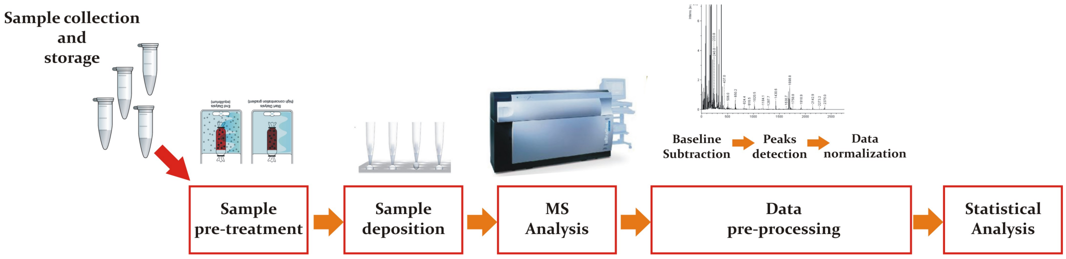

Figure 1.9: Pictorial diagrams of a typical profiling workflow. The analysis requires some pre- intra- and post-analytical steps.

1.11.1 Sources of errors in the protein profiling workflow

All steps of profiling workflow can potentially be a **sources of errors** (Table 1.2). The possibly introduced errors can be both random and/or systematic error (also called bias). Random error can be caused by unknown and unpredictable changes which can arise during all the steps of the analysis workflow, error that is shared across all the sample. Bias, which can also arise during all the analysis workflow, is a type of error which occurs non-randomly across the analysis. When bias occur without regarding the outcome or the measurement of the exposure, the so-called **nondifferential bias**, the effects on the study are less serious, but favor the null hypothesis of no association. Alternatively, when bias is associated to the outcome of the exposure measurement (**differential bias**), the effect tends to favor an association in either direction, which may no be the true relationship between the biomarker and the disease. Another aspect to consider is the confounding effect, that may alter the measurement of the biomarkers. The confounding effect can be inherent to the study design (e.g., difference in age, gender or weight of the subjects) or included in the workflow (e.g. differences in analytical condition that occurs during the experimental batch). Patients randomization during accrual is a valid strategy to address the confounders problem inherent to the patient's characteristics at baseline, even if specifying proper inclusion or exclusion criteria is necessary. Another aspect to consider is the **blinding** of the study. If the subject's group is know under analysis, investigators might be prone to analyze the specimen or interpret the assay result in a different way. If blinding cannot be done during these steps, then rigorous explicit operating procedures might help to minimize biases.
During the analysis phase, many errors or bias can be introduced. For example, if the spectroscopy pattern 'wanders' over time, it may be that the machine can inadvertently introduce a signal into the data. In this case, generated errors can be both random or systematic. If samples are analyzed randomly between cases and non cases, bias may be of random type; *vice-versa* if cases and non-cases are analyzed in bunch, introduced bias may be of differential type. Another example that can generate bias is the different specimens storage procedures and/or time. Suppose that cases samples have been stored for 10 years, while non-cases samples for less that 1 year; biases resulting from changes caused by storage conditions in samples can be generated. On summary, differential bias in analytical phase can occur if the cancer and non-cancer groups are

Possible source of bias or random errors	
Sample collection	In urine, first void versus mid-stream void Type of tube and/or anticoagulant Location of sample collection (single or multicentric)
Sample handling	Time before collection and handling Centrifugal time and temperature Storage temperature and time
Instrumental Analysis	
Sample processing	Fractionation and/or depletion Other sample processing like desalting
Experimental protocols	Freeze-Thaw cycles Deposition methods (Dry droplet, etc.) Matrix type Temperature and Humidity during crystallization Quality of crystallization Instrumental setting Calibration Manual or automatic sample deposition Manual or automatic acquisition
Data analysis	Spectra pre-processing Peak detection Low abundant peak identification Feature selection

Table 1.2: Source of errors or bias in protein profiling workflow. Modified from De Bock et al., (66).

handled in systematically different ways, introducing an apparent ‘signal’ into one group but not the other. Using a large sample size does not directly address biases, although it can reduce statistical uncertainty by providing a smaller confidence interval around a result and so reducing random error. Small studies done well can effectively answer important questions and demonstrate a ‘proof of principle’ about a molecular marker. The essential feature of such a study is design that minimizes problems from chance and bias and discussion that appropriately considers possible shortcomings.

Another aspect to consider, mainly during statistical analysis, is the over-fitting problem. In fact, proteomic studies often consider a multitude of variables, as so called features. In other words, researcher should answer the question to produce convincing evidences: "Does chance explain results?". To avoid over-fitting statistical data analysis should include multiple testing or adequate cross-validation methods, like the **leave one out cross-validation**.

1.11.2 Sample collection and handling

A successful biomarker research program starts not only with a careful study design, but also with the preparation of a detailed protocol for standard operation procedures (SOP) containing the definition of: a) the sample source and type and b) sample collection, storage and processing procedures. In fact, it has been demonstrated that small differences in sample collection and processing could have large impact on the results of the study. Further, researchers need to accurately avoid that clinical data may be site-, study-, population-, or sample-dependent (66). For example, Karsan et al. studied serum sample for breast cancer biomarker profiling by SELDI-TOF/MS and found that specimen collection and processing introduce significant biases in the spectral pattern, such that machine learning algorithms can differentiate between sample source, days of sample preparation, and days that they were read (67). They obtained even more surprisingly results by studying the multi-center collected samples. In fact, findings underline how there were distinct spectral features that the statistical algorithms successfully found, classifying the clinics from which the samples were acquired rather than the cancer patients itself (67). There are many pre-analytical steps that should be standardized. For example, types of tubes/or anticoagulant used for collection should be specified in SOP (68), like the fasting condition for venipuncture. For urine, the collection modality (first void versus mid-stream urine) has been demonstrated to be of primary importance for room temperature sample stability in proteomic analysis (48, 69). Other crucial aspects for plasma specimens are sample centrifugation, storage temperature and time, and exposure to freeze-thaw cycles (70). For example, if centrifugal speeds are too low and/or care is not taken in removing the plasma layer, contamination with platelets may occur and subsequently affect profiles.

Some changes may occur with storage conditions but most of them are analytes specific, because some proteins are stable for some freeze–thaw cycles, whereas other can show decreases after three such cycles, with similar differences in stability depending on storage temperature (68). Calvano et al. studied urine samples pre-analytical biases and stated that : *"it should not be under-stressed that the quite common storage temperature of -20 ^oC, often considered completely safe to the stability of urine samples is unsafe and for longer storages, liquid nitrogen is desirable"* (40). In particular, they found that urine, if not immediately analyzed, should be stored for no more than 35 days at -20 ^{o}C to avoid sample degradation and then the risk to occur in a false positive marker for diagnostic purposes. Differently, Schaub et al. found that storage of the urine samples at -70 ^{o}C is adequate and sufficient because contents did not change the spectra compared with those obtained before freezing and that almost the same spectra could be generated after four freeze-thaw cycles (48). The *storage bias* is one of the best know source of variability and bias in mass spectrometry. Further special care should be taken with tube plastics and the presence of contaminants by which MALDI-TOF/MS is heavily sensitive and for the time that a sample needs to reach the laboratory to be processed and placed in safe conditions.

1.11.3 Sample pre-treatment

As stated above, in MALDI-TOF/MS samples requires only simple pre-treatment steps. However, in order to reduce intra-sample variability, it is desirable to remove as well as possible samples contaminants and metal salts that affect the subsequent analysis. However, complicated and elaborated sample treatment have clear disadvantages, mainly because 1) they are time-consuming and 2) may be under-coming to loose samples analytes, potentially important for the analysis. In fact, as reported by Calvano et al., the proteins/peptides observed can be strongly and directly dependent on the sample pre-treatment method chosen, indicating that it has to be carefully selected and optimized for the specific disease under investigation (40). Another example was reported by Hu et al. They evaluated three different fractionation protocols to evaluate if splitting a sample into three fractions can better highlight different subsets of the proteins. Interestingly, they found that two distinct clusters clearly identified in each fraction. However, further exploration shown that the cluster found matched very closely with the day on which the sample collection protocol had been changed midway through the experiment rather than the fractionation of samples (71). Most of the paper that investigate sample pre-treatment for urinary biomarkers profiling are based on SELDI-TOF/MS analysis. As described above, SELDI protein-chips allow a direct, on-chip purification step. Therefore, MALDI-TOF/MS, which not encompass this step, needs the choice of a purification protocol before the sample deposition and analysis. One of the most widely used method for urine pre-treatment followed by mass spectrometry analysis is the centrifugal ultrafiltration, especially for the evaluation of the low molecular weight proteome (72). The centrifugation force applied to the sample cause a flow through a semipermeable membrane, able to retain solutes with a molecular weight higher than the nominal molecular weight cut-off (NMWC) (Retentate). *Vice-versa*, lower molecular weight fraction can freely pass through the membrane (filtrate) and collected (73). Further, other approaches may also be used. Thongboonkerd et al. compared Ultrafiltration with Acetone protein precipitation and stated that both provide complementary data for a 2D-page analysis (74). Differently, Calvano et al. compare ultracentrifugation, followed by Zip Tip desalting, with chromatographic urine purification. They used in house packed chromatographic columns, made with HLB micro-solid-phase extraction sample purification and found that it gave better results for linear mode MALDI-TOF/MS protein analysis (40). Another pre-treatment step proposed was the magnetic beads (41). This method uses different chemical chromatographic surfaces on an outer layer of magnetic beads to selectively purify certain subsets of proteins, allowing unbound impurities to be removed by washing with buffers. NanoLC coupled with MALDI-TOF/MS has been also suggested as a valid method for urine biomarker identification (39). Nowadays, urinary proteomes treatment and analysis has not been standardized yet, although the Human Kidney and Urine Proteome Project (HKUPP) working group has been debating this topics since 2006 (75).

1.11.4 Sample deposition and analysis

Once samples are collected and pre-treated for the analysis, they are ready to be deposited on the MALDI target plate together with the matrix. Despite there are three commonly used approaches to spot samples, namely dry droplet, thin-layer and the sandwich method, the ultimate goal is the homogenous co-crystallization of matrix and analytes. A phenomenon frequently observed in MALDI-TOF/MS is a strong variation in intensity and resolution of the signals at different positions of a sample spot (56). This so-called *"hot spot"* or *"sweet spot"* formation leads to poor shot-to-shot and spot-to-spot reproducibility and is therefore a factor strongly increasing measurement times and complicating automated measurements. Again, salts and contaminant can strongly influence the samples-matrix homogenous co-crystallization. Matrix is another aspect to consider. For example, preparations with HCCA often deliver good spot formation, while DHB preparations lead to formation of long needles exhibiting strong hot spot formation (56). Other authors show that 1) humidity (50) and 2) crystallization timing and temperature (76) can be optimized to create homogeneous crystals and increase MALDI-TOF/MS ionization performances. Schaub et al., who analyzed the sample deposition and the *"hot-position"* formation for urine specimens, suggest that the most representative spectra for a given urine sample is achievable by sampling many different spot positions and combining the data by summing acquired intensities (48). The robotic preparation of MALDI target plate has been demonstrated to decrease the variability of acquired spectra, with respect to manual preparation. As demonstrated by Tiss et al., the intra-run CV decrease from 9.6 ± 4.2 % to 7.5 ± 4 % by using an in house modified sample preparation robot, adapted for the ZipTip purification protocol (77). However, robotic equipment are not generally made for sample manipulation (e.g. like ZipTip) and are usually coupled with high performance liquid chromatography (HPLC), which need further steps focusing on sample pre-treatment and standardization.

1.11.5 Data pre-processing

A typical dataset arising in MALDI–TOF/MS protein profiling for candidate biomarker discovery contains tens or hundreds of spectra, with each spectrum containing tens of hundreds of intensity measurements representing an unknown number of protein peaks. Several modeling should be implemented simultaneously to extract valuable information. In fact, each spectrum signal can be approximately described as the following function:

$$y(t) = B(t) + N(t) \cdot S(t) + \epsilon(t)$$

The true signal, $S(t)$, consists of a sum of possibly overlapping peaks, each corresponding to a particular biological molecule, e.g. a protein or a peptide. In fact, two proteins/peptides could have the same mass or the instrumental resolution could not be as sufficient as need to detect different ionic species. Despite a parametrically characterization of the shapes of the peaks are generally not performed, the approximate shapes of peaks can be estimated empirically by simulating the physical process by which TOF mass spectrometers collect data. The normalization factor, N, is a constant multiplicative factor to adjust for spectrum-specific variability, e.g. to adjust for differing amounts of protein ionized and desorbed from each slide. The baseline function, B , represents a systematic artifact commonly seen in mass spectrometry data. This artifact is believed to be attributable to a cloud of matrix molecules hitting the detector in the early part of the experiment, or to detector overload. Normally, error is supposed to be mean-zero Gaussians with the variance a smooth function of t, (i.e. $\epsilon \backsim N[0, \sigma^2(t)]$.) (78) Spectra pre-processing usually consists of individual operations like high frequency filtering, baseline subtraction, peaks detection and intensities normalization, which can be executed in different orders. So far, no optimum operating sequence has been determined. Findings reported by Hu et al., show how spectra calibration is another crucial aspect to consider. They found that offset error in the calibration of the spectra can generate different findings in protein expression, when evaluating different group of patients (71). Although peaks are supposed to be perfectly aligned with respect to their m/z, a further step to align masses are advisable to avoid overcoming in calibration imprecisions.

Nowadays, many algorithms or platforms have been developed for mass spectrometry data pre-processing and handling. Some of them are public because developers have shared the code. However, because MALDI-TOF/MS spectra consists of high resolution data and suffer of the isotopes problem, many are nor directly or indirectly applicable. A detailed list of public peak detection algorithm usable for MALDI-TOF/MS data with their properties are reported in Table 1.3. Follows a briefly explanation of the computational approaches involved.

Smoothing Filters

The most used filtering techniques are: **1) moving average filter, 2) Savitzky-Golay filter and 3) Gaussian filter**. Moving average operates by averaging a number of points from the input signal to produce each point in the output signal. The Saviztky-Golay filtering is like a generalized moving average filter. It performs a least squares fit of a small set of consecutive data points to a polynomial and takes the central point of the fitted polynomial curve as output. Gaussian filter is like a weighted moving average filter, but sets larger weight factors for points in the center and smaller weight factors for points away from the center.

Baseline correction

Baseline correction is typically a two-step process: (1) estimating the baseline and (2) subtracting the baseline from the signal. The most common is the **wavelet function**, where a symmetric wavelet function (usually the Mexican Hat), is used. In fact, continuous wavelet transform (CWT) removes baseline automatically. In the **monotone minimum** two steps are used to estimate baseline. The first step is to compute the difference for adjacent points, which can be used to determine the slope of each point. Then, if the slope of a local point A is smaller than zero, a nearest point B to the right of A, whose slope is larger than zero, is located; differently, if the slope of a local point

A is larger than zero, a nearest points B to the right of A, whose intensity is smaller than A, is located. The intensity of every point on the result baseline between A and B equals to the intensity of A. **Linear interpolation** divide the raw spectrum into small segments and use the mean, the minimum or the median of the points in each segment as the baseline point. **Moving average of minimum** firstly estimates a rough baseline by finding local minimum within a window for each point and then it uses a moving window to smooth the rough baseline obtained.

Peak finding criterion

There are many peak detection methods developed, and most of them are made to detect peaks after smoothing and baseline correction. However, as stated above, CWT does not require baseline correction and, more interestingly, does not need any smoothing too. In the **signal to noise ratio (SNR) methods** a signal above a fixed ratio is considered a true positive. Therefore, how noise is defined take a special role. Noise can be estimated as 95% percentile of absolute continuous wavelet transform (CWT) coefficients of scale one within a local window or as the median of the absolute deviation (MAD) of points within a window. Choosing a **detection/intensity** threshold is helpful to filter out small peaks in flat regions. In fact, in these regions, SNR alone may identify many noisy points as peaks. **In the local maximum** peaks are defined as local maximum of N neighboring points. **Ridge lines**, used in the wavelet method, are obtained by the 2-D coefficient matrix with size of MxN, where M is the number of scales obtained after CWT transformation and N is the length of spectrum. Therefore, local maximal coefficients of adjacent scales are connected to form ridge lines. The distance between two adjacent points on a ridge line is considered a gap and a valid ridge line has gaps below a given threshold. **In the shape ratio method**, peak area is firstly computed as the area under the curve of a candidate peak. Shape ratio is computed as the peak area divided by the maximum of all peak areas. A valid peak has a shape ratio larger than a threshold. **In kernel density method**, the non parametric kernel estimator is used to depict the density function, which maximum is selected by the local maximum method.

Up to now, no peak detection algorithms have been developed specifically for MALDI-TOF/MS reflectron data. The major difficulties in reflectron data evaluation are: a) the presence of isotopic forms of peptides/proteins, b) the high m/z resolution contained in acquired spectrum. However, algorithms advancement for high resolution reflectron MALDI-TOF/MS data will benefit many Research fields, not only proteomic profiling.

1.11.6 Data Normalization

A good normalization strategy for MALDI-TOF/MS data should account for possible instrumental non linearity in peptides/protein quantification. Normalization strategies are generally used in the pre-processing workflow to make comparable different MS spectra. In fact, a well-known key limitation of MS is that the measured abundances of proteins are relative. This affects the calculation of peaks intensities and peaks area.

Program	Smoothing	Baseline	Peak finding criterion
Cromwell	Wavelet based	Monotone minimum	S/N & LM
LIMPIC	Kaiser window	Moving average	S/N & Detection Threshold
CWT	Wavelet based	Loess	S/N & Ridge lines
LMS	Gaussian filter	Lin. int.	S/N & LM
PROcess	Moving average	Lin. int. & Loess	S/N, LM & Shape ratio
Wave-spec	Wavelet Based	Monotone minimum	LM & Kernel Density

Table 1.3: Most used algorithms for peak detection on MALDI-TOF/MS data. S/N = signal to noise ratio; LM = local maximum; Lin. int. = linear interpolation. (79, 80)

The peaks magnitude changes among different samples, being related to the overall protein abundance (81). Recent studies have empirically shown that normalization is a crucial step for comparing mass spectra for biomarker identification (82, 83, 84, 85). Therefore, normalization is usually conducted in order to increase comparability of spectra resulting from different measurements (83).

Normalization of mass spectra typically entails subtracting an (optional) offset and dividing by a scaling factor. Such offset and scaling parameters can be defined and applied on considering all spectra (**global normalization**), or considering single spectrum (**local normalization**). In MALDI-TOF/MS important aspect to consider for ameliorating non-linearity in the detector response are the ionization suppression and the interaction between analytes, all undesirable variation that may get introduced in the MS data. The first intuitive normalization technique consist on normalizing for an analyte considerable as **internal standard**, which quantity is known. Arguably, in a so complicated process like MALDI-TOF/MS protein profiling, normalization with respect to an internal standard would be ideal. With this approach a fixed amount of an exogenous component (usually a peptide or a protein) is added to all samples and peak heights or areas of individual endogeneous analytes are measured relative to it. Signals (discrete m/z values) that increase relative to that of the internal standard when two or more samples are compared can be considered to reflect increases in the amounts of the analyte; decreases correspond to reduced levels. Incorporation of the internal standard would be ideal also because it is possible to adjusts for some of the variability inherent the process of sample preparation, ionization and ion detection (86). However, in proteomic profiling, it is not typical to add a known amount of a known protein to the sample because doing so hampers high-throughput and adds logistical complexity (81). Differently, the usage of isotope modified compounds has already been demonstrated to be successful in MALDI-TOF/MS absolute quantitative determination of urine Hepcidin, obtaining good accuracy (percentage relative error less than 10%) and recovery (more than 80 %) (87). Unfortunately, isotope compounds can not be used in protein profiling, because proteins under investigation **are not *a priori* known**.

The standard approach for normalization is the **total ion current (TIC)**. TIC represents the summed intensity across the entire range of the detected masses, or more accurately the square root of the sum of the squared intensities (86). Its use is preferred by many authors mainly because the TIC is a good surrogate for total protein content in the measured sample and therefore peaks intensities could be normalized by the total sample protein contents. Another common normalization technique is the **relative abundances**, usually called relative intensity (RI). In this method, the tallest peak is called the *base peak* and it will have a relative value of 100%. All other peaks are given values relative to that in terms of percentage. RI is generally used because relative abundances of the same spectrum are considered comparable each other. Moreover, if base peak is shared across spectra, normalizing by relative intensities is the same as normalizing by a reference peak which concentration vary, often with respect to protein content. Other normalization methods have recently been suggested, like mean, median and linear rescaling (81). The formulas corresponding to these normalization, TIC and RI are reported in the following equations (from 1.2 to 1.7) where i and A are the feature and the spectrum under analysis respectively, and Lin means Linear rescaling.

$$A_i^{IS} = \frac{A_i}{A_{IS}} \tag{1.2}$$

$$A_i^{mean} = \frac{A_i}{mean(A)} \tag{1.3}$$

$$A_i^{median} = \frac{A_i}{median(A)} \tag{1.4}$$

$$A_i^{RI} = \frac{A_i}{max(A)} \tag{1.5}$$

$$A_i^{TIC} = \frac{A_i}{\sqrt{\sum_{i=1}^{N}(A_i)^2}} \tag{1.6}$$

$$A_i^{Lin} = \frac{A_i - min(A)}{max(A) - min(A)} \tag{1.7}$$

All these methods are local normalization. In particular local normalizations (based on single spectrum) have the important advantage that do not require to re-calc the normalization factor each time a new spectrum is added to the dataset. Global normalizations, that use coefficients calculated from all spectra in analysis, need to be re-estimated each time a new spectrum is added. Zero-offset mean or median normalization method is calculated by dividing feature intensities by the mean or the median intensity value of the spectrum's features. Linear normalization utilizes the largest and smallest peaks of each mass spectrum. The intensity at the smallest peak (the minimum intensity) of the spectrum is subtracted from the intensity at each mass-to-charge ratio. That value is then divided by the difference between the maximum (largest peak) and minimum (smallest peak) intensities of the spectrum (82).

1.11.7 Statistical data analysis

Currently, there is a big debate in the statistical methods for the analysis of MALDI-TOF mass spectrometry data. This discussion is mostly due to the very high dimensionality of the obtained dataset, which usually contain a large number of variables (also called features) as compared to the relative low number of subjects. This type of dataset motivates the need for computational techniques in data analysis and for suitable methods to assess reproducibility and overfitting (88, 89, 90). Two useful broad categorizations of the techniques used are **supervised learning** techniques and **unsupervised learning** techniques (88). The two techniques are easily distinguished by the presence of external subjects' labels. The unsupervised learning techniques, such as finding those features that are correlated across all the samples, operate independently of any external labels. By the contrary, in the supervised learning techniques those labels and data are used to create a learning method for these labels. These two types of machine learning methods are generally used to answer different types of questions. In supervised learning, the goal is typically to obtain a set of variables (a process also known as features selection) that can be used reliably to make a diagnosis, predict future outcome, predict future response to pharmacologic intervention, or categorize that patient as part of a class of interest. In unsupervised learning, the typical application is to find either a completely novel cluster of peptide/proteins with putative common (but previously unknown) expression or, more commonly, to obtain cluster or group of features that appear to have patterns of similar expression. The major used techniques for unsupervised learning are Principal Component Analysis and Clustering determination. Instead, Decision trees and Support vector Machines are widely used as supervised learning techniques. Also methods for multiple comparisons with error correction like False Discovery Rates method (FDR) can be applied to find features highly correlated with the disease. Despite the employing of this "computational approach" in data analysis, the over-fitting problem remains unsolved. With the term "over-fitting", researchers generally mean the probability of finding a discriminatory pattern of features completely by chance, which can happen when large numbers of variables are assessed for a small number of outcomes. This problem can be partially overcome by splitting the database in a Training set and in a Validaton set. By this way, a discriminatory pattern or prediction rules can be derived by the training set. After that, the validation set is kept totally independent and can be analyzed to test the hypothesis (for example, discriminatory pattern or prediction rule) that is derived from the training set (89, 90).

1.12 MALDI-TOF/MS reproducibility

MALDI-TOF/MS analysis has several aspects that can affect its reproducibility. As illustrated above, technical variability may arise in prior to acquisition steps (like matrix deposition methods, calibration procedures and machine performance during time), during the acquisition steps and in the post-acquisition workflow (data processing, including baseline subtraction, smoothing, peak detection and normalization) (63). Because most of these phenomena are currently poor understood, an optimization could be advantageous not only for the general comprehension but also for the further possible application to the analyses. In each analysis step, only some effects have been investigated, while others still remain unknown. For example, the presence of contaminants in samples can decrease the quality of obtained crystals, which often requires manual control by the experimenter and doesn't allow the automatic sample acquisition. Further, during the ionization process, the adducts formation with metal salts (mainly Na^+ and K^+) and the presence of multiply charge ions can split the parental mass intensity in different peak signals. The ion suppression effect may increase variability of some ionic species when evaluating different samples or in different biological fluids, especially when one (or a bunch of analytes) have higher concentration with respect to the others. Another source of variability is the targeted analyte effect, demonstrated by Toghi Eshghi et al. by spiking a complex mixture of analytes with a single analyte at concentration similar to limit of detection (LOD). The targeted analyte effect enhance sensitivity and decrease LOD behaving as a signal carrier for other analytes (91).

1.12.1 MALDI-TOF/MS instrumental detection limit

The limit of detection for MALDI-TOF/MS depends on many variables, which may change detection sensitivity by orders of magnitude; so any general statements about LOD must be considered as a merely rough approximations. For example, for small peptides, under optimal conditions, MALDI-TOF/MS detection limits can extend up to $< 1\ fmol$. Differently, during analysis of complex samples like plasma, proteins and peptides are detected with lower sensitivity.

An enormous variety of definitions relating to detection limits and to quantitation limits are commonly used in the clinical chemistry literature. Unfortunately, universally accepted procedures for calculating these limits do not exist. MALDI-TOF/MS, like most analytical instruments, produce a signal even when a blank sample (matrix without analyte) is analyzed. This signal is referred to as the instrument background level. After, noise can be defined as the measure of the fluctuation of the background level and it is generally calculated by the standard deviation of a number of consecutive point measurements of the background signal. A common and widely used approach for limit of detection (LOD) estimation is to set *a priori* signal-to-ratio (S/N) as LOD threshold. LOD estimated by this method is generally referred to as instrumental detection limit. Another commonly used LOD definition is calculation of the lowest quantity of a substance that can be distinguished from the absence of that substance (a blank) within a stated confidence limit (generally 1%). By this latter definition, the following formula can be written:

$$signal(LOD) = signal(blank) + 3 \cdot SD \tag{1.8}$$

Therefore, the detection limit is estimated from the mean of the blank, the standard deviation of the blank and some confidence factor. Also in mass spectrometry instrumentation (and in MALDI-TOF/MS) the presence of background results in a nonzero signal even at zero concentration of the analyte. Sub-optimal detection efficiency compromises the output signal. Analyte concentration variations introduced by the analyte-matrix co-crystallization, desorption/ionization, analyzer, and detector add noise to the measurements, thus limiting the threshold as well as the confidence of low-abundance analyte detection. A recent study evaluated the detection limit and sensitivity of MALDI/TOF-MS when analyzing complex samples. The practical calibration curve of MALDI/TOF-MS, which is the measured mass spectral signal versus a given analyte concentration, differs from the ideal curve. In the ideal curve, the analyte concentration is proportional to the intensity, while in MALDI/TOF-MS the latter curve has a sigmoidal shape (91). This results underlined that different analytes, contained the same sample, may behave differently in terms of LOD. Therefore, analyzing complex samples (which contains many proteins), multiple LOD need to be assessed. In protein profiling studies, hundreds of proteins are normally detected, and so the instrumental LOD determination by serial dilution of samples is not possible. One hypothetical solution could be to dilute a reference sample (e.g. a pooled sample) and analyze each obtained dilution by MALDI/TOF-MS. Because as many intensities are measured for each single analyte as many dilution points, the true sigmoidal curve can be depicted, as shown by Toghi Eshghi et al. The background signal information can be calculated and used to determine each analyte-signal LOD.

1.12.2 Data pre-processing as source of variability

As a multiple step process, data pre-processing can heavy influence the data analysis and so the identification, quantification and discovery of disease-related biomarkers. Firstly, if baseline subtraction is not performed for all spectra, peaks detected across samples could be of different intensities only for the presence of an artifact signal, and not for true differences in analytes contents. Because the behavior of this baseline signal is not *a priori* estimable, the only solution for researchers is to analyze each spectrum, singularly estimate the artifact signal and subtract it from the spectrum signal. Secondly, if peak detection is not accurate, some information will be irremediable lost. In fact, after peak finding and spectra alignment, false negative undetected peaks will have zero intensities. This introduce an excess of zeros in the data that should be handled in the subsequent data analysis. Therefore, peak detection appears as the most critical point. A further step in data pre-processing is the smoothing filters. This usually apply traditional signal processing techniques for high frequency filtering and allow to correct for possible imprecision in the detected intensities.

1.12.3 Public peak peaking algorithm performances

Up to now, no specific MALDI-TOF/MS reflectron mode peak detection algorithms have been developed yet, most probably for the high spectra complexity due to the presence of multiple isotopic forms of peptides. Differently, many algorithms have shown to perform quite well for MALDI-TOF/MS linear model. Yang et al. reviewed and compared most of the public peak finding algorithms, usable also for MALDI-TOF/MS linear mode (Cromwell, LIMPIC, CWT, LMS and PROcess) (79). They used one group of simulation data, in a m/z range between 400 Da and 64800 Da and one group of real MALDI-TOF/MS data (obtained from 246 individually purified protein, tryptic digested), ranging from 800 and 3500 Da. So the evaluation they performed are consistent with MALDI-TOF/MS reflectron data. As performance parameters they used **false discovery rate (FDR)** and **sensitivity**. False discovery rate is defined as the number of falsely identified peaks divided by the total number of peaks found, while sensitivity is defined as the number of correctly identified peaks divided by the total number of true peaks. From both the simulation and the real MALDI-TOF/MS data, CWT performed better. They concluded stating that CW optimally characterizes the shape of peaks in mass spectra and that the concept of forming ridge lines in CWT effectively removes false positive peaks. However, by allowing to introduce as maximum a 5% of false peaks (FDR of 0.05 %), the best sensitivity reached was 50% for both analyses (79). Serendipity, half of the peaks were undetected, also considering the CWT algorithm which performed as the best. Unfortunately, there is not independent evaluation of the Wave-spec algorithm. On considering this results and also our experience, all peak detection algorithms are supposed to be *error prone*, and their capability, especially on identify low abundant peaks, may heavily influence not only the statistical analysis results, but also the further finding on possible candidate biomarkers. Because low abundant peaks are also influenced by the detection limit problems, peak finding errors are strictly related to the accurate instrumental detection of this peaks.

1.12.4 Influence of LOD in peaks detection

Unfortunately, only some studied evaluated peak detection sensitivity with respect to peaks S/N ratio, which is an important tuning factor to evaluate performances of peak detection, especially in low abundant ionic species. In fact, it is easier to correctly identify peaks when the difference with background is wide than peaks with low signal. Tracy et al., found that, at least for some ionic species, lower S/N is associated not only with higher uncertainty in the peak detection but also with a greater residual variation after alignment (92). In our experience, this effect is almost present in MALDI-TOF/MS data. On applying different peak detection algorithms an excess of low abundant, low S/N true peaks can be considered undetected. On the other hand, trying to optimize the peak detection for low abundant ionic species may lead to many noisy signals to be considered as true peaks, increasing the overall false discovery rate (FDR). And this effect may probably be algorithm-dependent.

Because this relationship between S/N ratio and sensitivity in peak detection exist, some further consideration are need regarding the instrumental LOD, because these three concepts are closely related. In fact, detection limits can be estimated for MALDI-TOF/MS as well as in other instruments when researchers are working on single o few analytes. However, in protein profiling studies, where many hundreds of analytes are usually detected, a multitude of single detection limits should be calculated, one for each analyte.

Therefore, a different strategy for peaks peaking is desirable. One possible solution could be to include a feed-back process in peaks detection that aids in preserving losing of low abundant ionic species, even if intensities are below LOD. To illustrate this statement we can suppose to compare many spectra from different samples. We can also suppose to obtain, after peaks peaking and alignment, only a number n of detected peaks. On comparing results, probably we will found that some low abundant ionic species are mis-detected (for peaks detection error) in a variable number of spectra (so the detected number of peaks for each spectrum are less of n). Moreover, it will not be a surprise to discover that most of these mis-detected peaks are "true" peaks, namely with intensities above the detection limits. To deal with this problem, a better strategy will be to:

1. detect peaks for all spectra;
2. focus only in peaks m/z (so, discarding the detected intensities);
3. re-evaluate all the spectra at the detected m/z values in order to obtain the signal intensities, even if that specific signal would not be detectable by the peak peaking algorithm.

However, by this methods we will introduce many noisy signals, which are signals below the detection limit threshold. Therefore, a further strategy to deal with the detection limit problem should be desirable.

2

AIMS

Our overall objective is to evaluate urine collected after prostatic massage, to identify candidate biomarkers for PCa by using the MALDI-TOF mass spectrometry instrumentation. Usually, PCa is an indolent form of tumor and, especially in early stages, it causes nonspecific symptoms which are generally referred as lower urinary tract symptoms (LUTS). As these symptoms are indistinguishable from those produced by benign prostatic hyperplasia (BPH) and prostatitis, patients referring LUTS to General practitioner are usually forward to consultants for a urological examination and enter in differential diagnosis. According to the European Association of Urology guidelines, suspicious DRE examination and/or increased PSA levels required a prostate biopsy to exclude the possibility of cancer. Moreover, the last European randomized study of screening for prostate cancer (ERSPC) showed that DRE and PSA based screening leaded to a large part of patients being over-treated (up to 50%). Therefore, we have proposed a cross-sectional study to evaluate the possibility of identify new proteomic biomarkers, taking advantage of urines collected during the urological examination and the prostatic massage, which was performed during the DRE inspection. In this study, two hundred five patients that referred LUTS to consultants at the Urological Unit at University of Padova were collected, from December 2008 to June 2011, and all patients undergone to prostate biopsy for suspicious PCa diagnosis. The study results should aid in reducing the number of worthless first-biopsied and assist urologist on differential diagnosis of patients with LUTS.

Using a high throughput instrumentation like MALDI-TOF/MS, collected urines was profiled in low molecular weight **reflectron mode**, to detect possible candidates biomarkers for PCa. However, as MALDI-TOF/MS reproducibility has been widely debated in literature, two additional evaluations were performed to estimate the MALDI-TOF/MS analytical variability and the urine measurement errors, respectively. Firstly, **Intra- and Inter-run instrumental reproducibilities** were evaluated by using a pool of urines collected and dialyzed before the analyses. Secondly, an **external dataset**, derived from a serial collection of urine on apparently healthy male subjects which did not refer LUTS, was used to estimate **the coefficients of bias** for measurement error and the with-in subject variability of MALDI-TOF/MS features. Moreover, as MALDI-TOF/MS features contain some data below the instrumental signal detection limits, an appropriate method to deal with this problem was suggested.

Our main specific aims are:

- To quantify the intra- and inter-run analytical variability of MALDI-TOF/MS analysis of urine.
- To evaluate the effects on features variability and on technical replicates comparability of six normalization methods, commonly used in proteomic or genetic studies.
- To estimate the signal detection limits (sLOD) thresholds of urine detected MALDI-TOF/MS features and to evaluate the sLOD impact on features variability.
- To evaluate the possibility of spectra optimized signals detection for future development of new peak detection algorithms.
- To inspect the error structures of the external dataset
- **To assess whether estimations of coefficients of bias (ICCs) were affected by left censored data**.
- To determine by logistic regression *Naïve* analyses whether MALDI-TOF/MS features, obtained from urine collected after DRE analyses, were associated with PCa presence at biopsy.
- **To adjust logistic regression *Naïve* coefficients for measurement error, by using both the regression calibration and the SIMEX methods**.

Secondary specific aims include:

- To evaluate the MALDI-TOF/MS features representativeness effect by feature exclusion using the data collected for the reproducibility study.
- To inspect overall informations contained in MALDI-TOF/MS features, obtained from urine collected after DRE analyses, by unsupervised cluster analyses for future application of class prediction machine learning algorithms.
- To assess whether MALDI-TOF/MS urine analyses overall **achieve the minimal Clinical Chemistry analytical desirable performance.**

3

MATERIALS AND METHODS

3.1 Patients

3.1.1 Reproducibility study

Ten healthy subjects (5 men and 5 women), with age ranging from 24 to 49, were selected as urine donor. During the same day, subjects were asked to collect at least 10 ml of urine. Finally a **urine pool** was created by mixing together 10 ml of each sample. Before any sample pre-processing, the pooled urine was centrifuged at 16.000 g for 15 minutes to eliminate cell debris and after, 50 aliquots of 2 ml were prepared and stored at -80 ^{o}C for less that 1 month until any further analysis.

3.1.2 External dataset and measurement error structure

Twenty healthy male, with age ranging from 24 to 56, were selected as urine donor. For each subject, 2 or 3 urine aliquots were collected in a time-window of one week.

3.1.3 PCa patient's biomarker study

In this cross-sectional study we included two hundred five patients that referred Lower urinary tract symptoms to consultants at Urological Unit at University Hospital of Padova, in a time period from December 2008 to June 2011. All patients undergone to prostate biopsy for suspicious diagnosis of PCa. Based on the anamnestic records collected from Urologists, a Patients' database was created which included age, histological results of prostate biopsy, Gleason score (for prostate cancer patients), total and free PSA levels. Prostate biopsy was performed at Urological Unit by trans-rectal ultrasound biopsy of the prostate with a 10 to 16-core template. PSA levels were measured using the Immulite ®2000 system at Department of Laboratory Medicine, University Hospital of Padova. A bio-bank which contained patients' urines collected after digital rectal examination was created. All samples were stored at -80 ^{o}C before any further analysis.

3.2 Experimental set-up

3.3 Samples pre-processing by Dialysis

Urines were firstly spiked with an internal standard (IS) 14 amino acids synthetic peptide (NH2–MLTELEKALNSIID–COOH) (Primm srl, Milan, Italy), reaching a final concentration of 12.58 pmoli/μL. With except of the Reproducibility Study (see below), for all the other studies the amount of dialyzed urine was of 200 μL. Dialysis was performed by a semi-permeable membrane Spectra/Por©7, MWCO 1 kDa (Spectrum laboratories, CA, USA) maintained at 4^{o}C in gently agitation for 16 hrs in 500 ml of ultra-pure water. During dialysis water buffer was changed one times, after 2 hours, prior to leaving overnight dialysis. After the dialysis process, all the dialyzed samples were stored at -80 ^{o}C for no more than 1 week until the MALDI-TOF/MS analysis. A pooled urine sample was analyzed before and after dialysis for evaluate salt content by Gas analyzer, Rapidlab865 (Bayer S.P.A., Milano, Italy).

3.3.1 Reproducibility study

In the reproducibility study, both the intra- and the inter-run **MALDI-TOF/MS and sample pre-processing** reproducibilities were evaluated. For both analyses we used the pooled urine. For the intra-run experiment, a total of 1 ml of pooled urine was dialyzed and, starting from a volume of around 1.1 ml obtained after the dialysis, the sample were subdivided in 26 aliquots of 40 μL. For a total of 26 days, a single aliquot was daily thawed and analyzed by MALDI-TOF/MS for the instrumental variability assessment.

Differently, for the inter-run experiments, the pooled urine was firstly subdivided in 14 aliquots of 200 μL and then independently dialyzed. For each aliquot a post-dialysis volume ranging from 200 to 220 μL of urine was obtained. Aliquots were immediately frozen until the further MALDI-TOF/MS analysis. Aliquots were analyzed by MALDI-TOF/MS in a total of 4 different analytical sessions, each one including 4 aliquots, in order to estimate the MALDI-TOF/MS instrumental and sample pre-processing variabilities.

3.3.2 sLOD estimation

Three aliquots of the dialyzed pooled urine, collected for the Reproducibility Study, was thawed and serially diluted by ultra-pure water up to 1/256. Each dilution was analyzed by MALDI-TOF/MS during a single experimental session.

3.3.3 Measurement error structure of MALDI-TOF/MS features estimation

Urine samples were centrifuged at 16.000 g for 15 minutes to eliminate cell debris and, after dialysis, stored at -80 ^{o}C for less that 1 month. Samples collected for the measurement error analysis were analyzed in three analytical sessions. At each analytical session, a varying number of samples (from 10 to 20) were thawed and immediately analyzed by MALDI-TOF/MS.

3.3.4 PCa patient's biomarker study

For each analytical session, a list of samples was randomly extracted from the bio-bank, without knowing patient's name or disease status. Specimens were thawed, centrifuged at 16.000 g for 15 minutes to eliminate cell debris and, after dialysis, analyzed at MALDI-TOF/MS. A total of 11 analytical sessions (which included a varying number of samples, ranging from 14 to 27), were evaluated.

3.4 Profiling workflow

3.4.1 MALDI-TOF/MS urine analysis

Dialyzed samples were directly analyzed by MALDI-TOF/MS, without further preprocessing steps. All the instrumental analyses were identically performed for all the three studies.
For each sample, ten μL of dialyzed urine was mixed with 10 μL of saturated HCCA (α-cyano-4-hydroxycinnamic acid), prepared in 50% 0.1% TFA and 50% ACN (1:1 v/v). Following, 1 μL of this mixture was spotted four times on a ground steel MALDI-TOF/MS target. The crystallization was performed at constant humidity and temperature ranges during all the experimental sessions. MALDI-TOF/MS measurements were performed using an Ultraflex II MALDI-TOF instrument (Bruker Daltonics, Bremen, Germany), operating in reflectron positive ion mode. Ions were formed by a pulsed UV laser (λ=337nm) beam. The instrumental conditions were: IS1 = 25kV; IS2 = 21.65kV; reflectron potential = 26.3kV; delay time = 0 nsec. External mass calibration (Peptide Calibration Standard, Bruker Daltonics, Bremen, Germany) was based on monoisotopic values of $[M+H]^+$ of Angiotensin II, Angiotensin I, Substance P, Bombesin, ACTH clip (1-17), ACTH clip (18-39), Somatostatin 28 at m/z 1046.5420, 1296.6853, 1347.7361, 1619.8230, 2093.0868, 2465.1990 and 3147.4714, respectively. For each analyzed sample, one spectrum was collected averaging 2000 laser shots obtained from the respective four replicate spots. Reagents were freshly made each running session. All chemicals and solvents were purchased from Sigma-Aldrich (Sigma Aldrich SRL, Milan, Italy) and Bruker Daltonics (Bruker Daltonics, Bremen, Germany).

3.4.2 Spectra pre-processing

Once spectra were acquired, instrumental data were processed by the instrumentation software. After spectra baseline correction, peak detection was carried out using Flex Analysis, version 3.3 (Bruker Daltonics, Bremen, Germany), using the SNAP algorithm and a S/N set to 3. SNAP algorithm is based on *local maxima* for exact peaks m/z identification while peaks are defined as "detected" if their S/N ratios are greater than the chosen S/N threshold. The utility of Flex Analysis is that it allows different groups of peaks to be compared visually, with some peaks directly selected with it. To enhance peak detection we chose to manual revise low abundant peaks in each spectrum in order to enhance accuracy of peak detection without incurring in peak overlooking. Therefore, the detected peaks were aligned before the further analyses. Peak alignment was performed using MatLab, version 2010a (The MathWorks, Inc., Natick, MA, USA). In particular, overlapping all the obtained spectra, the peaks falling within a sliding window of ±0.3 Da were considered identical (features). So, the real m/z values of any single peak included in a 0.6 Da window were averaged and this average m/z value was assigned to all those peaks. After a brief visual revision to identify possible misalignment, the dataset was exported in a single text file. Moreover, a text file contained the corresponding m/z peaks-list was eventually generated and carefully revised for miss-alignment.

3.4.3 Optimized signals detection for MALDI-TOF/MS profiling

The identified peaks-list were used to determine whether, with a feed-back procedure, it was possible to re-evaluate the acquired spectra and extracts features signals more accurately, increasing the overall MALDI/TOF-MS protein profiling reliability. Therefore, the previously detected signals were discarded and the following method was used to calculate the new features signals.

1. Calibrated spectra were firstly exported in text files by Flex analysis;

2. by using a Matlab in house routine, spectra were loaded in memory, baseline subtracted and smoothed for high frequency by the Savitzky and Golay function (*msbackadj* and *mssgolay* functions respectively (from the The MathWorks Bioinformatic Tool);

3. Peaks intensities were finally retrieved by evaluating the local maximum at the m/z specified position from the m/z peaks-list, in a sliding windows of ±0.4 Da.

3.4.4 Features' signals normalization

Data normalization was based on mean, median, internal standard (IS), relative intensities (RI), total ion current (TIC) and linear rescaling normalization. Formulas used for to normalize data are reported above in equations from 1.2 to 1.7. All these normalization procedures are **local normalizations** and so the analyses were made separately for each spectrum. A dedicate R in house function was written to perform all the normalization starting from feature's signals.

3.5 MALDI-TOF/Ms urine profiling signal detection limit estimation

Signal detection limit (sLOD) was calculated by the commonly used definition (91). The signal background intensities obtained at the highest dilution were used to estimate the mean and the standard deviation. Finally, sLOD were calculated by adding to the means three standard deviations. Different approaches were assessed to correctly estimate a function to interpolate data. Spline smoothing and polynomial fitting were both tested in comparison to a lowess smoothing line (100 iterations and a 0.4 span window) to identify the best solution. Analyses were performed by R using the *smooth.Pspline* and the *rlm(y poly(5,x))* functions.

3.6 Reproducibility study

All the analyses were repeated by considering the automated Flex Analysis peak detection and the optimized signal detection method.

3.6.1 sLOD adjustment

sLOD was estimated both for the intra- and inter-run studies. The following steps were made:

1. After spectra pre-processing, features were sorted from the lowest to the highest m/z.

2. For each feature, starting from the lowest m/z, intensities were scanned for values below sLOD.

3. These values below sLOD were substituted with the corresponding feature's estimated sLOD divided by 2 (esLOD/2), calculated by the equation derived from the signal detection limit estimation experiments.

3.6.2 Intra- and inter-run variability assay

For all the 26 intra-run or 14 inter-run technical replicates, the variability has been estimated by means of coefficient of variation ($SD/mean$). Features CVs and features means were calculated by considering each feature singularly, and the results were graphically evaluated by scatter plot. The overlapping smoothing lines were calculated with the following parameters: 100 iterations and a 0.4 span window. Pooled estimated CVs were calculated by medians and interquartile ranges, a more robust evaluation with respect to the means and standard deviations. Box Plots were drawn to analyze the replicates comparability by using all the features' values obtained for each single replicate. Analyses were made by using R in house routines.

3.6.3 Representativeness effect evaluation by features exclusion

On considering each single feature, across spectra representativeness may be defined as the total number of spectra minus the number of spectra which contain feature's signals below sLOD. On the other hands, the feature's percentage of undetected signals is equal to one minus feature's representativeness.

For these analyses, features representativeness and percentage of undetected signals were firstly calculated. A set of undetected threshold were chosen, ranging from 100% to 0%, with steps of 10 %. Therefore, features with percentage of undetected signals below the threshold were maintained while those above the threshold were excluded from the analyses, gradually excluding less representative features. The analyses were iterated to cover all the undetected threshold steps as following:

1. features were excluded if their percentage of undetected signals are above the threshold;
2. signals of the remaining features were sLOD adjusted;
3. the features' median overall CVs were calculated;
4. threshold was lowered of one step.

CVs bootstrapped standard errors were calculated by resampling data and with a total of 1000 interactions, as described below.

3.6.4 Standard error estimation by Bootstrapping

Sampling independently from an unknown distribution F, the boostrap estimation of the standard error, denoted by $\sigma_B(\widehat{\rho})$ was described by Efron as following:

1. Let $\widehat{F}$ be the empirical probability distribution,
2. and let $X_1^*, ...X_n^*$ be a random sample from $\widehat{F}$, i.e. n indepedendent draws each with distribution $\widehat{F}$
3. and let $\widehat{\rho}^* = \widehat{\rho}(X_1^*, ...X_n^*)$

therefore the boostrap estimate is $\sigma_B(\widehat{\rho}) = \{var_*(\widehat{\rho}^*)\}^{0.5}$ where $var_*(\widehat{\rho}^*)$ indicates the variance of $\widehat{\rho}^*$ under the probabiity mechanism, with $\widehat{F}$ fixed at its observed value.
In other words, the bootstrap estimate $\sigma_B(\widehat{\rho})$ is simply the standard deviation of the quantity of interest $X_1^*, ...X_n^*$, if the unknow distribution F is taken equal to the observed distribution $\widehat{F}$ (93).

3.7 External dataset and measurement error structure

The error structure was evaluated by the external dataset. Because this dataset contains information also on the samples creatinine level, we wanted to verify whether normalizing data by creatinine could decrease variability and increase repeatability. Therefore, MALDI-TOF/MS features were firstly evaluated before and after Creatinine normalization. Secondly, features were median normalized, sLOD adjusted and log-transformed. For the **ICC**, **with-in** and **between subjects** variation, the R package "ICC" and the function "ICCest" were used.

3.8 ICC estimation under measurement error and/or left censoring conditions

Under the hypothesis of multiplicative error, if X, W and U are log-normal distributed, the classical error model holds after logarithmic transformation of these variables (Equation B.3). As specified in Appendixes, for the measurement error theory, if x is the "true" long term average measure of the biomarker, w, which is the short term measure of the biomarker, vary for the measurement error which encompass biological variability and instrumental error. By considering the following model:

$$y = \alpha + \beta x + \epsilon \tag{3.1}$$

where x is the biomarker level and y is the healthy response on a continuous scale, under the **classical error model** ($w = x + \epsilon$), we can fit the equation:

$$y = \alpha + \beta^* w + \epsilon \tag{3.2}$$

where ICC (namely the Reliability ratio) can be calculated by the following formula:

$$ICC = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\epsilon^2} = \frac{\sigma_x^2}{\sigma_w^2} \tag{3.3}$$

where σ_x^2 represents between-person variation, and σ_ϵ^2 represents within-person variation.

Now, starting from the following hypothesis:

1. it is possible to assume that the classical error model hold
2. ϵ error structure is known $[N(0, \sigma_\epsilon^2)]$
3. the mean and standard deviation of the proxy variable w in equation 3.1 is known,

it is possible to simulate two variables w_1 and w_2 measured with errors. We simulated data based on our "real" condition. In order to mimic the exact conditions obtained from urine MALDI-TOF/MS analysis, we generated the variable x with mean and σ equal to the median of MALDI-TOF/MS features mean and the median of MALDI-TOF/MS features standard deviation, respectively. Therefore, once generate the variable x, it was firstly duplicated in x_1 and x_2. After, the two errors ϵ_1 and ϵ_2, which had mean 0 and a defined $\sigma_{\epsilon 1}$ and $\sigma_{\epsilon 2}$, were generated. So, the two variables measured with a know amount of error $W_1 = X + \epsilon_1$, and $W_2 = X + \epsilon_2$ were generated. W_1 and W_2 shared the same variable X, considered the *"true"* value, hypothetically obtained from two separate measurements of the same patient assuming *"error-free"* conditions.
Based on percentiles values of W_1 and W_2, a LOD threshold was chosen for both W_1 and W_2 and values below the threshold level were considered as "undetected". Values below LOD (see equation A.3) have been modified according to the commonly used methods described in literature. In particular, the following adjustments were evaluated:

1. Substitution of X < LOD by $E(X \mid X < LOD)$
2. Substitution of X < LOD by $E(X \mid X > LOD)$
3. Substitution of X < LOD by zero
4. Substitution of $X < LOD$ by $LOD/2$

Results were compared with respect to the full dataset (without the LOD threshold). **Monte Carlo simulations** have been performed by using a *in house* R routine. Variables generation was repeated N = 1000 times and, for each run, the ICC was calculated by considering W_1 and W_2 as repeated measures of the same group of patients. At the end of simulation, the mean ICC ($I\bar{C}C$) was calculated by averaging results from the 1000 replicates. Many simulations have been performed with increasing $\sigma_\epsilon = \sigma_{\epsilon 1} = \sigma_{\epsilon 2}$ values, because both errors are considered identical.

3.8.1 Monte Carlo standard error

As described in Appendixes, considering a generic experiment, if θ is the parameters under estimation, N is the number of simulations performed, the **Monte Carlo standard error** can be calculated by the following formula:

$$MC\ s.e.\ [\theta] = \sqrt{N^{-1} \cdot \sum_{i=1}^{n} (\hat{\theta}_i - \bar{\theta})^2} \tag{3.4}$$

3.9 PCa database urine evaluation by MALDI-TOF/MS

The optimized signal detection method was used to query MALDI-TOF/MS spectra database for features' intensities. Once obtained signals, they were firstly sLOD adjusted and after median normalized and log_2 transformed before any analysis. Age's correlations with outcome and MALDI-TOF/MS features were inspected by Spearman correlation, to identify a possible role of these variables as confounding factor.

3.9.1 Unsupervised clustering of MALDI-TOF/MS profiling data

Hierarchical clustering analyses were performed by considering all the detected features and all the patients. Divisive clustering dendrograms were drawn by using the *Ward method* and the *Euclidean distance*. The R functions used were **hclust** and **dist**. In addition, the analyses were repeated by including only subjects with negative biopsies and PCa patients' groups, so excluding the patients which had BPH, Inflammation, AAH, PIN and Atypical proliferation as histology of prostate biopsy.

3.9.2 *Naïve* logistic regression

Logistic analyses were made considering patients' outcome as a binary variable, which included References and PCa, and each single MALDI-TOF/MS feature as predictor. Analyses were performed also including for Age as confounding. Hosmer-Lemeshow goodness of fit test were evaluated only for logistic regression models which included Age as confounder.

3.10 PCa outcome prediction adjusted for measurement error

3.10.1 Regression calibration analyses

Intraclass correlation coefficients and within-subjects variances were calculated for features intensities on the external dataset, after median normalization and $\log_2$ transformation by the R function "ICCest" (Package "ICC"). For each feature, the point estimate of $\hat{\beta}^*$ was obtained by dividing the $\hat{\beta}$ coefficients, obtained after the *naïve* logistic regression, with the corresponding ICC ($\hat{\beta}^* = \hat{\beta}/ICC$). So, $\hat{\beta}^*$ confidence intervals were calculated as suggested by Rosner (94). Firstly, the $Var(\hat{\beta}^*)$, $Var(ICC)$ and k_0 were calculated by the following formulas:

$$Var(\hat{\beta}^*) = (1/ICC^2) \cdot Var(\hat{\beta}) + (\hat{\beta}^2/ICC^4) \cdot Var(ICC) \tag{3.5}$$

$$Var(ICC) = 2(1 - ICC)^2 \cdot [1 + (k_0 - 1) \cdot ICC]^2/[k_0(k_0 - 1)(n_1 - 1)] \tag{3.6}$$

$$k_0 = \left(\sum_{i=1}^{n_1} k_i \frac{\sum_{i=1}^{n_1} k_i^2}{\sum_{i=1}^{n_1} k_i} \right) /(n_1 - 1) \tag{3.7}$$

where k_i is the number of replicates for the i_{th} subject and n_i is the dataset sample size.

After $100 \cdot (1 - \alpha)$ CI for $\hat{\beta}^*$ is given by:

$$\hat{\beta}^* \pm z_{1-\alpha/2} \cdot \sqrt{Var(\hat{\beta}^*)} = \left(\hat{\beta}_1^*, \hat{\beta}_2^*\right) \tag{3.8}$$

All these calculation were made by an in house R function.

3.10.2 SIMEX logistic regression analyses

SIMEX evaluations were performed by using STATA 12.1, with the Stata programs Simex from J. W. Hardin, H. Schmiediche and R. J. Carroll as describe in the Stata Journal *"The simulation extrapolation method for fitting generalized linear models with additive measurement error"* (sj3-4, 2003), by using quadratic extrapolation and bootstrap standard error calculation (200 replicates).

4

RESULTS

4.1 Sample pre-treatment

4.1.1 Urine dialysis

Dialysis was important for sample desalting, avoiding loose of important low molecular weight peptides, which can be informative to predict Patients outcome. Spectra/Por©7 is a pre-cleaned dialysis membrane, which allows to retain more than 90% of peptides with MWCO higher than 1 kDa. In this work, dialysis efficiency was evaluated by measuring salt contents in pooled urine for Reproducibility study and results are reported in Table 4.1

Samples	Na^+ **(mM)**	K^+ **(mM)**	Cl^- (mM)	Ca^{2+} **(mM)**	PO_4^{2-} **(mM)**
Untreated urine pool	104	25	95	3.1	16
Dialyzed urine pool	< 0.01	0.01	4	0.44	0.12

Table 4.1: Na^+, K^+, Cl^-, Ca^{2+} and PO_4^{2-} concentration in the same urines pool, analysed before and after dialysis

4.1.2 Internal Standard effects on detected features

One aliquot of pooled urine, collected for the Reproducibility Study, was independently analyzed in the same way, with and without adding IS at concentration specified above. So, after dialysis, the two samples were spotted on MALDI target, analyzed and recorded. Spectra were pre-processed by Flex analysis for peaks detection and alignment. Firstly, the intensities of both samples were compared by a scatterplot (Figure 4.1). After, a 2^{nd} order polynomial fit and a lowess smoothing line were evaluated and overlapped to the plot. The estimated polynomial equation, reporting the relationship between the pooled urine features intensities with IS with respect to the pooled urine features intensities without IS was: $y = 2.73 - 0.93 \cdot X + 0.28 \cdot X^2$, where y indicate the pooled urine without IS. Evaluation were performed by using the $\log_{10}$ transformed values of the features' intensities, to compact data and reduce variability.

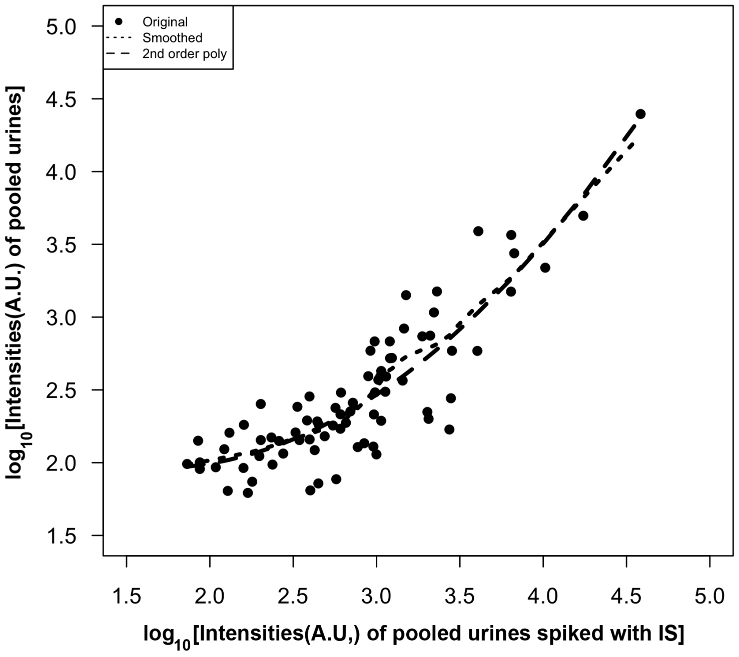

Figure 4.1: MALDI-TOF/MS analysis of two aliquots of the same dialyzed pooled urines, obtained with and without adding IS at a final concentration of 12.58 pmoli/μL

Figure 4.1 shows that IS spiking in urines, generated a overall increased intensities signal, which became marked for the low abundant features. The increasing trend in log scale become linear when signals were higher than 1000 a.u.

4.2 MALDI-TOF/MS urine profiling signal detection limit estimation

MALDI-TOF/MS analyses of the serial diluted pooled urines were performed during the same day, in the same conditions and in a single analytical session. After peaks peaking by the optimized signal detection method, we found that the detected MALDI-TOF/MS signals intensities decreased as long as the urine dilution increase, for all the identified features. According to the findings reported by Toghi Eshghi et al., the behavior followed approximately a sigmoid shape curve (data not shown) (91). Following, the signal limits of detection (sLODs) were calculated by the canonical definition of sLOD as described above in equation 1.8. The sLODs varied slightly between features of similar m/z, but varied widely across all the mass range. For example, between the feature with lower m/z and those with higher m/z, sLOD varied between 70.4 and 6.4 arbitrary units decreasing when m/z increase. Calculated sLOD mean were 20.4, standard deviation

20.13, median 13.5 and interquartile range (IQR) 7.40 to 27.42 (Figure 4.2).
Because urine peptide/protein contents is supposed to vary across samples, we needed to estimate sLOD also for peptides/proteins not detected by this specific experiment. So it seemed reasonable to estimate the expected sLOD (esLOD) for any hypothetical feature with respect to features m/z ($[E(sLOD \mid m/z)]$). Therefore, we evaluated a *lowess regression*, in comparison to a *smoothing Spline* and a 5^{th} order polynomial fitting. The better solution was the polynomial fitting, which resulting equation was:

$$E[sLOD] = 4.0{-}5.9{\cdot}10^{-01}{\cdot}m{+}3.6{\cdot}10^{-4}{\cdot}m^2{-}1.1{\cdot}10^{-07}{\cdot}m^3{+}1.70{\cdot}10^{-11}m^4{-}1.1{\cdot}10^{-15}{\cdot}m^5 \tag{4.1}$$

where m is the mass at which sLOD is estimated. In particular, the 5^{th} order polynomial gave the best fit obtainable. In the following paragraphs the terms *esLOD* and *sLOD* will be used interchangeably.

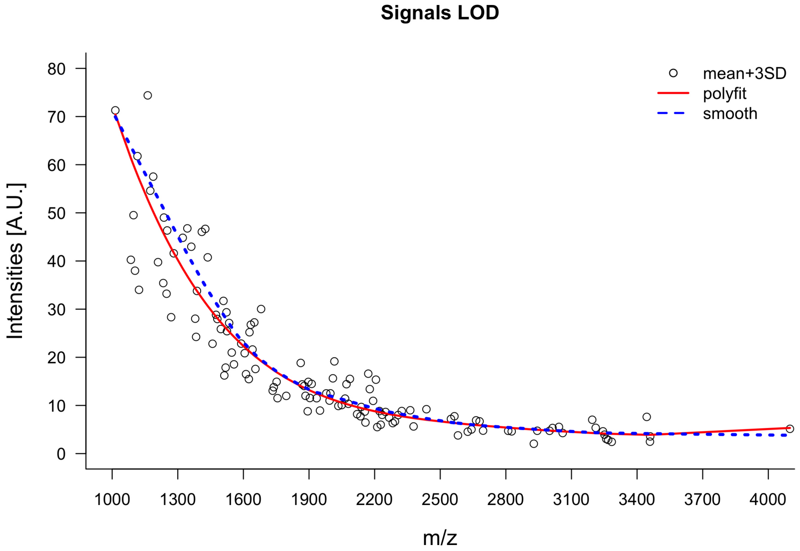

Figure 4.2: Estimated sLOD by diluting urine pool up to 1/256 in water, in triplicate. For each feature, represented by a single point, sLOD was estimated by the canonical way as described above in equation 1.8. Lowess smothing and 5^{th} order polynomial lines are overlapped in figure to show the expected sLOD derived from the two methods, with respect to each features m/z.

4.3 Reproducibility Study

Proteomic is a branch of mass spectrometry that is often considered poor *reproducible*, especially lacking across labs repeatability. Despite HUPO is developing SOP and protocols to increase across labs reproducibility, up to now proteomic has struggled to achieve a place as Clinical Chemistry instrumentation. The most important issues in comparing proteomic results are due to 1) different proteome coverage (e.g. sample are analyzed with different protocols, or difficulties in applying well defined SOP) and 2) technical challenging that mainly depends on sample complexity. However, before estimate the inter-labs reproducibility, instrumental reproducibility (which can be called also technical repeatability) should be assessed. In this study we wanted to verify whether MALDI-TOF/MS may be a reliable tool for protein profiling, especially for proteomic biomarkers discovery. Intra-run and inter-run variability can aid to assess analytical instrumental precision, while data normalization can help in increasing across spectra comparability.

4.4 Intra- and inter-run variability assay

Twenty six aliquots and fourteen aliquots of the same pooled urine were analyzed by MALDI-TOF/MS for the intra- and inter-run variability assay, respectively. After spectra acquisition and pre-processing by Flex Analysis, a total of 120 and 129 peaks were identified respectively for intra and inter-run studies, in a m/z ranging from 1000 to 4000.

Estimated sLOD (esLOD) for both the intra- and the inter-run assays were calculated with the corresponding detected features m/z values by using the equation 4.1. In both the intra- and inter-run experiments, a high percentage of features' intensities resulted below sLOD, being the median percentage 38.4 % (IQR: 0.0%-92.3%) and 42.8 %(IQR: 7.1%-85.7%), respectively. Moreover, the percentage of intensites below sLOD is highly inversely correlated with their median raw abundances (intra-run: ρ= - 0.913, $p < 0.001$ and inter-run: ρ= - 0.926, $p < 0.001$).

For each identified feature, **the trend in variability with respect to instrumental signal intensity** was evaluated by means of CV and by a scatterplot graphical representation. The plot, and the added smoothing line, show the behavior of CVs with respect to log_{10} of features signals, for both intra- and inter-run (Figure 4.3) before and after the sLOD ajustment.

After, mean, median, RI, TIC normalizations and linear rescaling (Equations 1.2 to 1.7) were performed to all spectra to assess which of them accounted better the **across spectra features variability**. Variabilities were estimated before (Raw data) and after applying normalization strategy. Results are summarized in Table 4.2.

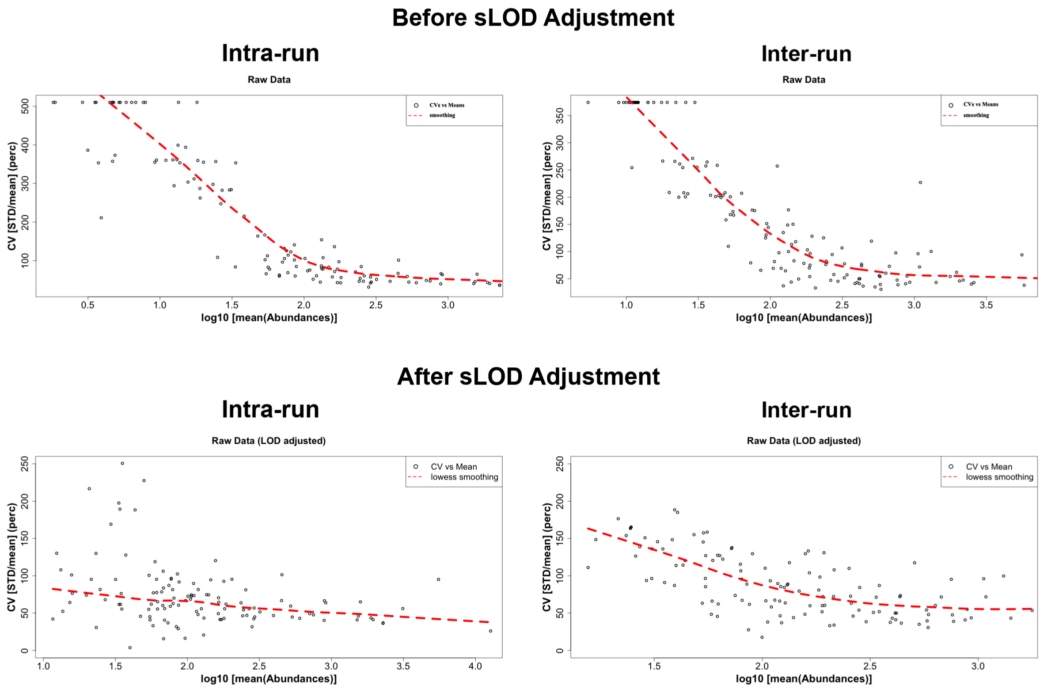

Figure 4.3: Trend in variability with respect to log_{10} instrumental signal intensity for the intra- and the inter-run assays. CVs (%) and means are calculated from the features' intensities, by considering all the acquired spectra.

	Normalization methods						
	Raw Data	**Mean**	**Median**	**TIC**	**Linear**	**RI**	**IS**
Intra-Run CVs (%)							
Median	104	103	103	107	108	107	132
(IQR)	(60-354)	(50-353)	(65-354)	(55-355)	(60-354)	(59-354)	(93-364)
Inter-run CVs (%)							
Median	118	109	113	110	114	113	212
(IQR)	(57-254)	(49-255)	(60-254)	(50-255)	(52-254)	(52-255)	(171-270)

Table 4.2: Before sLOD adjustment across spectra features variability. CVs were calculated for the intra- and inter-run assays without (Raw data) and after applying the Normalization Strategies, by considering all the detected features.

4.4.1 Optimized signals detection for MALDI-TOF/MS profiling

A new strategy was applied to re-evaluate features signals. Once obtained the new signals intensities matrix as described in Materials and Methods, the evaluations already done were repeated with the new data, to study whether the new approach, together with data normalization and sLOD adjustment, could ameliorate the trend in features variability, with respect to instrumental signal intensity, and the across spectra variability. Therefore, we applied the normalization strategy and results of the most important normalization methods, for both before and after sLOD adjustment, are reported in Figures 4.4, 4.5, 4.6, 4.7 and Table 4.3. In box plots, each single box depicts a single spot in MALDI plate, namely a single analyzed aliquot, for both in intra- and inter-run assay, before and after data normalization. sLOD adjustment was performed by the $LOD/2$ substitution method. In both intra- and inter-run studies, a low percentage of features were found below esLOD, being the median percentage 10 % (IQR: 1%-20%) and 3 % (IQR: 0%-8%), respectively. However, the percentage of intensities below sLOD is significantly correlated with their median abundances (intra-run: ρ = -0.783, p $<$ 0.001; inter-run: ρ = -0.870, p $<$ 0.001). Based on these results, **we chose the new optimized signal detection for all the further analyses**.

Optimized signal detection features' variability							
	Normalization methods						
	Raw Data	**Mean**	**Median**	**TIC**	**Linear**	**RI**	**IS**
Before sLOD adjustment							
Intra-Run CVs (%)							
Median (IQR)	40 (35-46)	27 (21-33)	26 (21-30)	32 (25-39)	35 (26-40)	35 (27-42)	115 (95-126)
Inter-run CVs (%)							
Median (IQR)	46 (39-59)	28 (22-38)	25 (20-35)	31 (25-41)	33 (25-44)	29 (23-41)	154 (144-168)
After sLOD adjustment							
Intra-Run CVs (%)							
Median (IQR)	31.8 (15-44)	28 (22-30)	20 (12-33)	-	37 (30-45)	36 (31-45)	101.1 (97-118)
Inter-run CVs (%)							
Median (IQR)	40 (34-49)	28 (23-34)	23 (15-34)	-	31 (25-44)	28 (23-34)	169 (152-179)

Table 4.3: Optimized signal detection across spectra features variability results. CVs were calculated for the intra- and inter-run assays without (Raw data) and after applying the Normalization Strategies. Before and after sLOD adjustment results were reported, except for the TIC normalization after sLOD adjustment, which was not applicable at sLOD adjusted data.

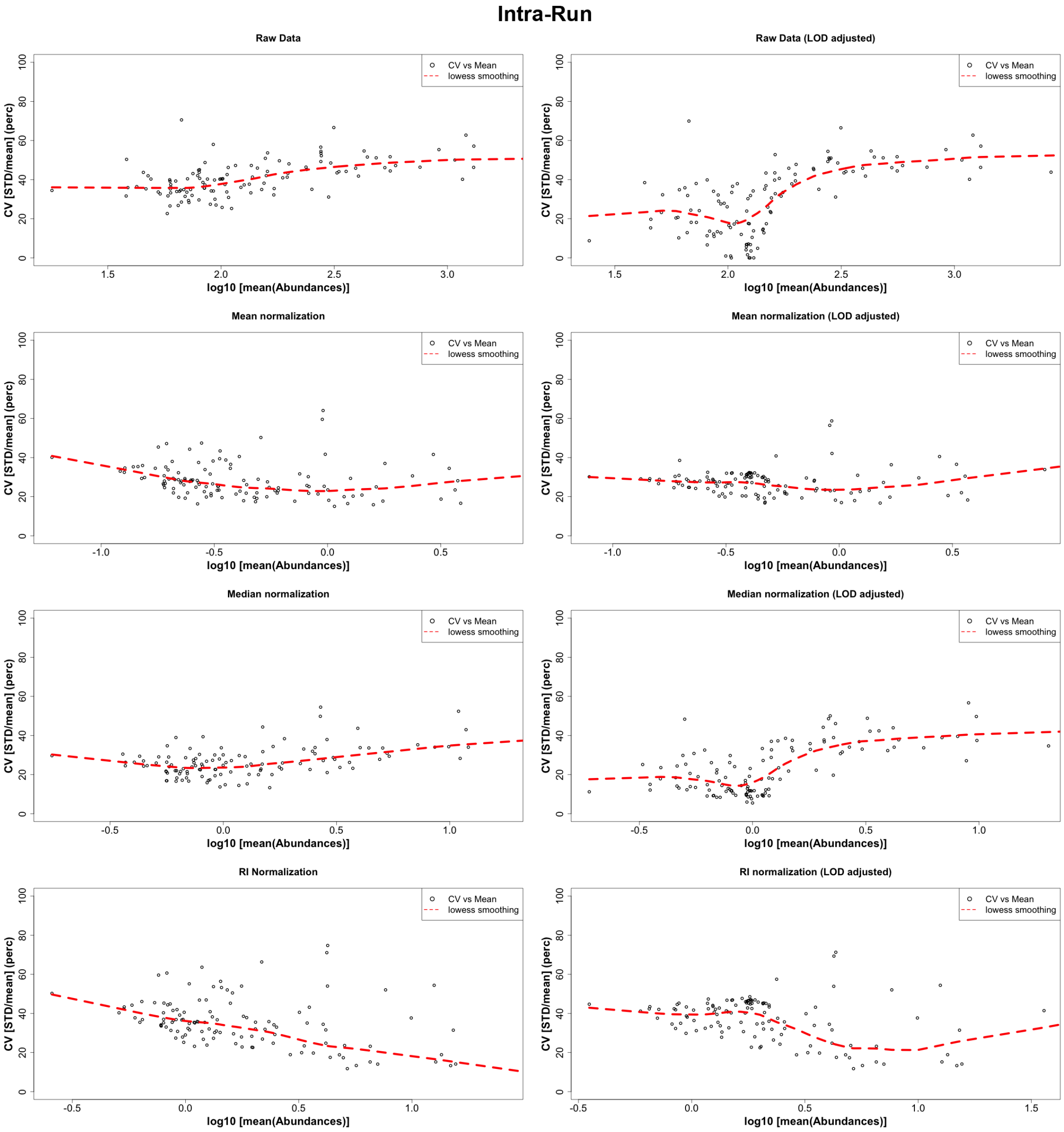

Figure 4.4: Intra-run across spectra features variability obtained by using the Optimized signals detection analysis and by applying the most meaningful normalization methods. CVs (%) and means are calculated by the features signals, considering all the acquired spectra. Before (left panels) and after (right panels) sLOD adjustment were reported.

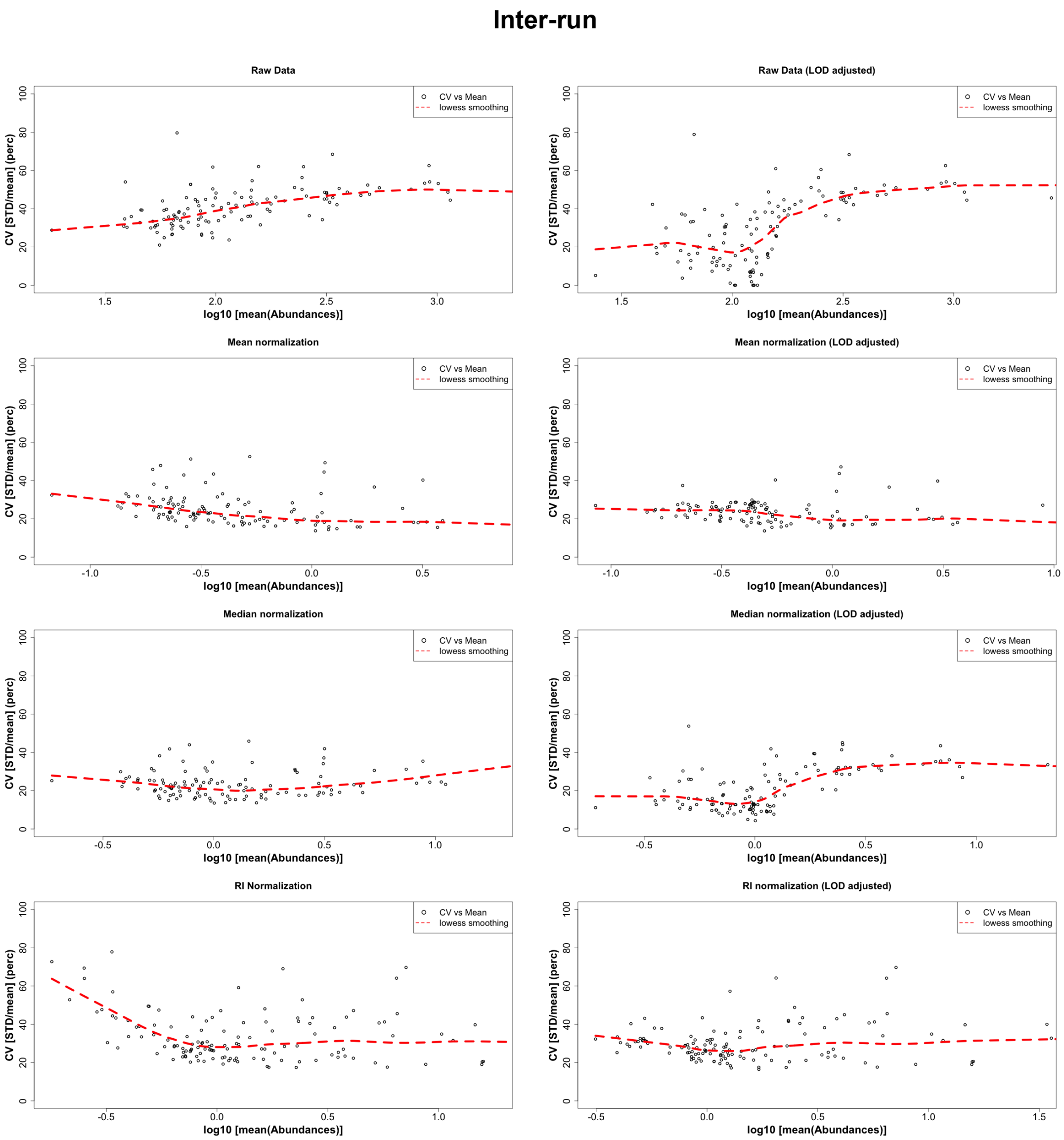

Figure 4.5: Inter-run across spectra features variability obtained by using the Optimized signals detection analysis and by applying the most meaningful normalization methods. CVs (%) and means are calculated by the features signals, considering all the acquired spectra. Before (left panels) and after (right panels) sLOD adjustment were reported.

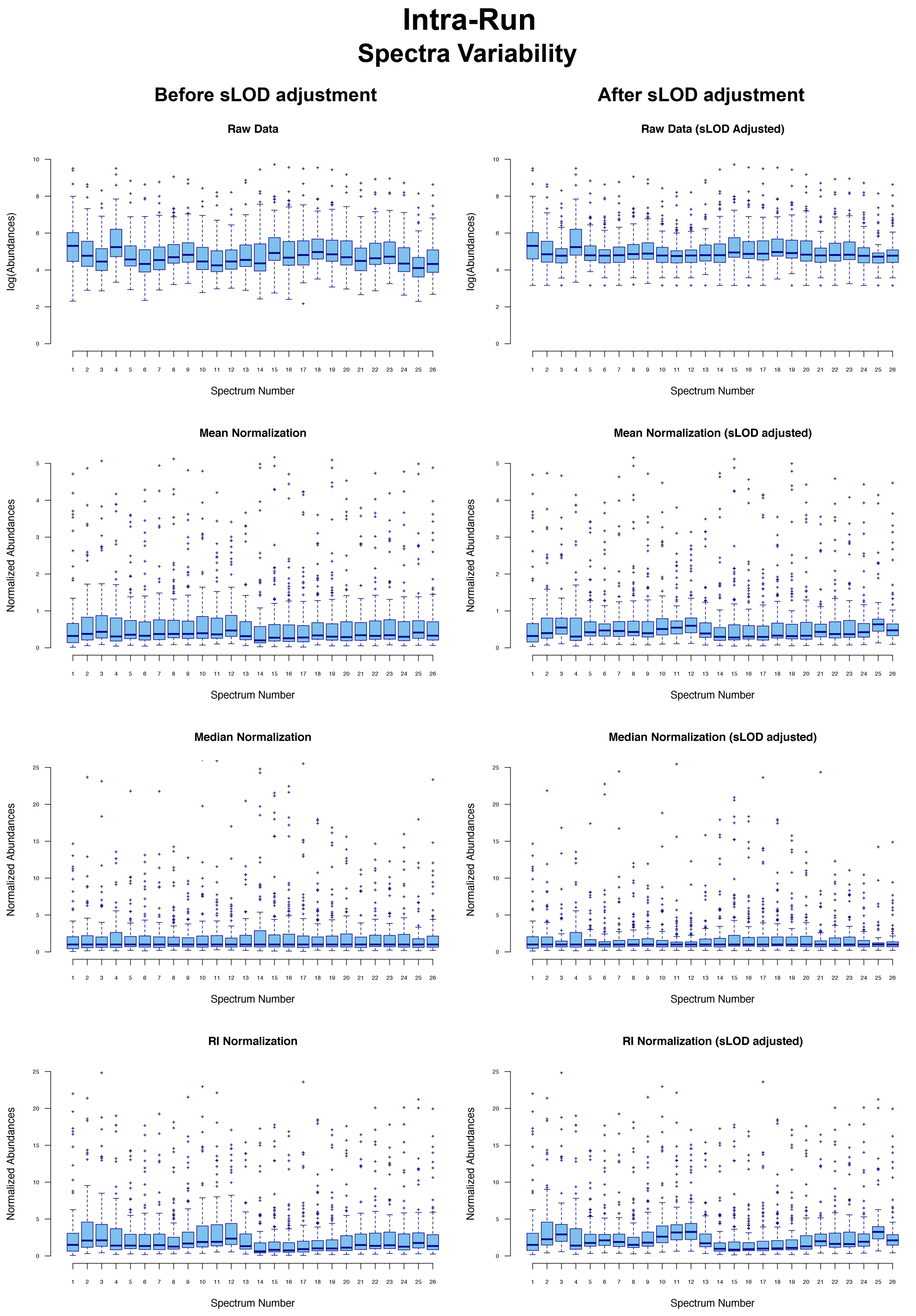

Figure 4.6: Intra-run between spots variation of median normalized features intensities, obtained by the Optimized signals detection method and by applying the most meaningful normalization methods. Each single box depicts a replicate, namely a single analyzed aliquot. Analysis was repeated before and after sLOD adjustment.

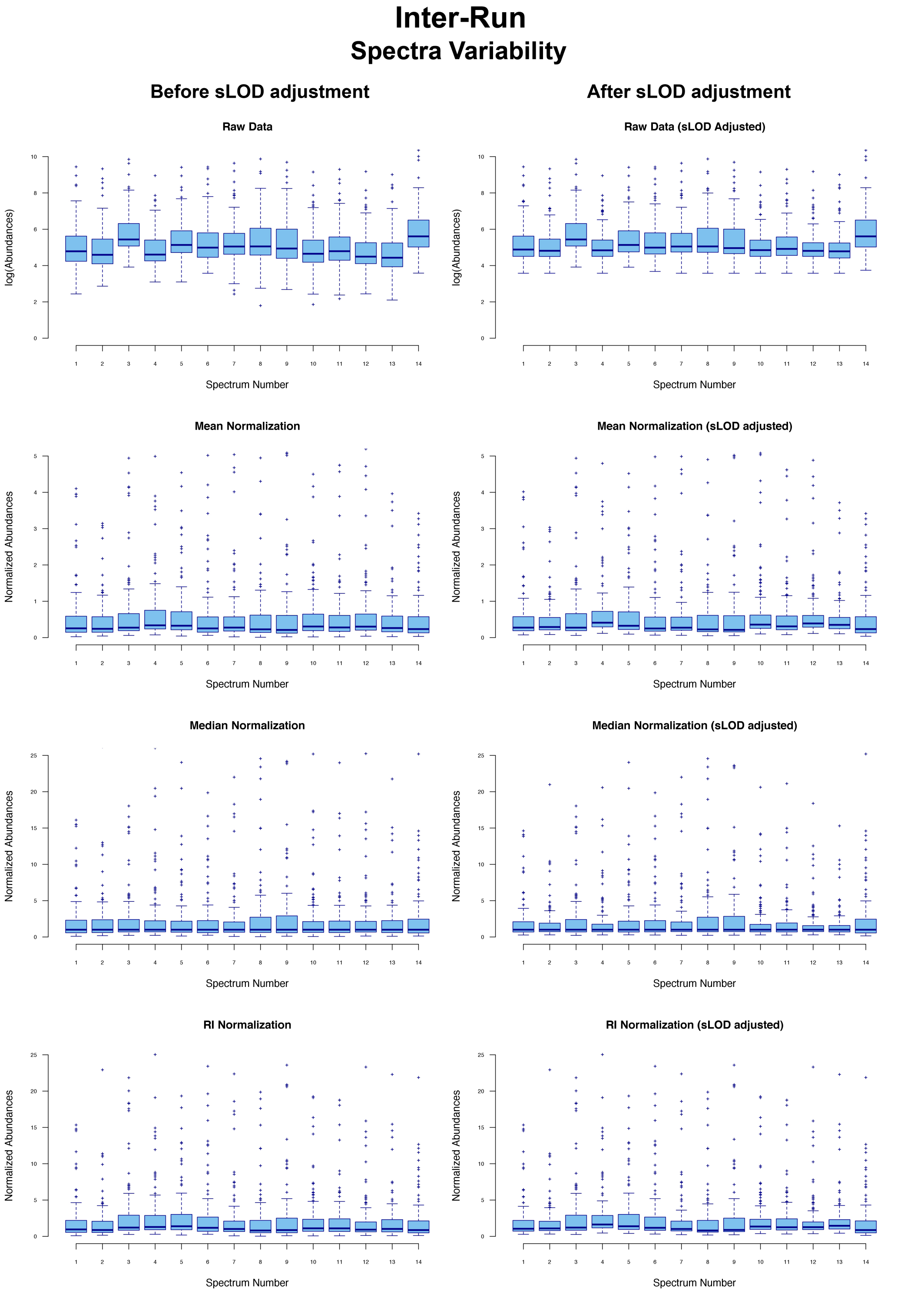

Figure 4.7: Inter-run between spots variation of median normalized features intensities, obtained by the Optimized signals detection method and by applying the most meaningful normalization methods. Each single box depicts a replicate, namely a single analyzed aliquot. Analysis was repeated before and after sLOD adjustment.

4.4.2 Representativenes effect evaluation by features exclusion

Because a high percentage of features' intensities resulted below sLOD for both intra- and inter-run assays, it would be interesting to evaluate whether the overall intra- and inter-variability vary by gradually excluding features based on their representativeness. Less representative features have many signals detected below sLOD, because they are not shared across samples. In protein profiling studies it is a common practice to **discard less representative features**, especially when representativeness falls below 20 % or 10 %. According to statistical theory for left censored data, it is expected than more a feature will be detected below sLOD across spectra, and more it will contribute to increase the overall variability, because its signal is largely due to random noise rather than to the presence of a peptide/protein in sample (95). In order to try to evaluate how the percentage of intensities below sLOD influences variability, an iterated analysis was performed. Briefly, gradually excluding features from the analysis bases on their percentage of signals below sLOD (**undetected signals**), the overall CVs were estimated each time. The results of the estimation for both intra- and inter-run studies were reported on Figure 4.8.

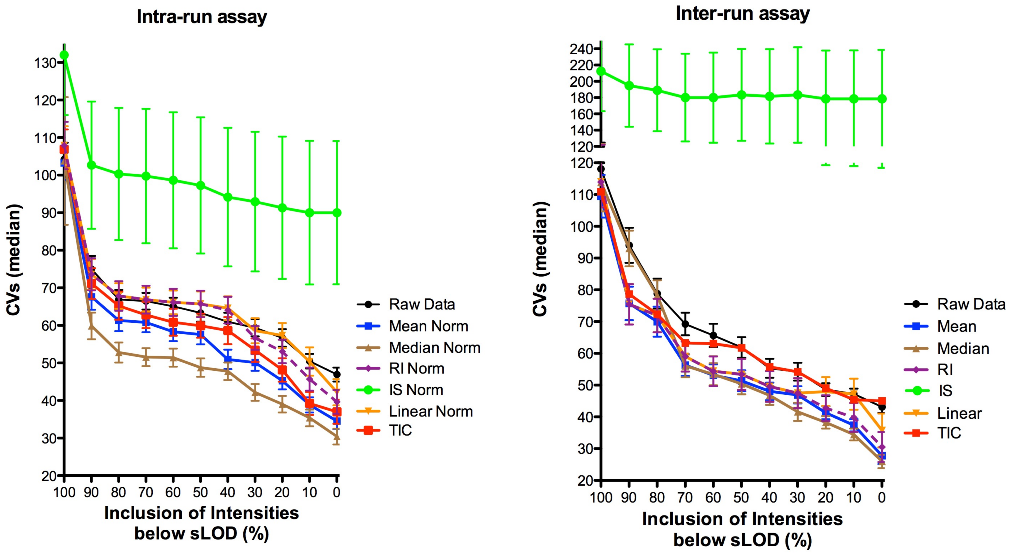

Figure 4.8: Across spectra variability and features representativeness by the Flex Analysis **peak detection method**. CVs were obtained from iterated analyses based on features' exclusion by their percentages of values below sLOD. Bars report standard error, estimated by bootstrapping.

Moreover, the analyses were repeated also for the **optimized signal detection method**, which performed quite different (Figure 4.9). Figures reported also the standard errors, estimated by the bootstrap statistic, performed as described in Materials and Methods.

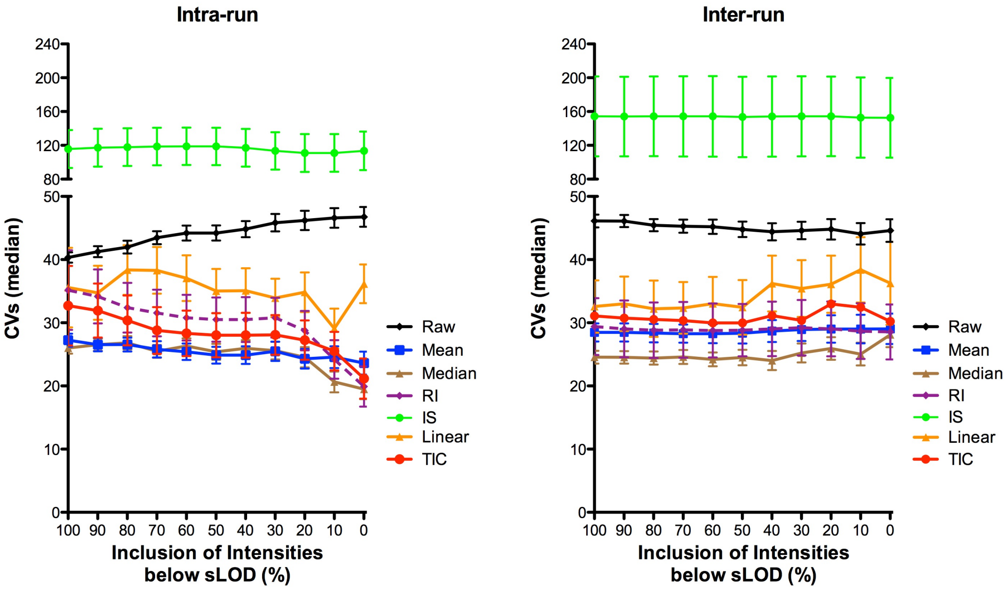

Figure 4.9: Across spectra variability and features representativeness by the **optimized signal detection method**. CVs were obtained from iterated analyses based on features' exclusion by their percentages of values below sLOD. Bars report standard error estimated by bootstrapping.

4.5 External dataset and measurement error structure

ICC is widely used in literature as reliability index, because it accounts for both the consistency of performances from test to retest, as well as change in average performance of the study participants as a group over time. Further, adjusting measurement error by regression calibration uses ICC to correct the estimated coefficient in regression modeling, especially when errors information are based on external datasets. In order to inspect the measurement error in MALDI-TOF/MS analysis of urine, we collected urine samples from male subjects, in healthy conditions, with mean age 41, ranging from 23 to 56. For each patient at least two samples were collected within 5 days, while for 4 subjects 3 replicates were collected in a time period of 1 week. Creatinine, measured in all the samples, could be an easy indicator of urine dilution. MALDI-TOF/MS analysis was performed on dialyzed urines after spiking IS at the specified concentration.

4.5.1 Creatinine's ICC

Samples creatinine values were normal distributed (Shapiro Wilk : $W = 0.98, p = 0.754$), with mean 10.52 $mmol/L$ and standard deviation 4.71 $mmol/L$. By using the repeated measures, ICC was 0.402 (95 % CI: 0.024-0.698), while the subjects within variance (σ_w^2) was 13.45.

4.5.2 Raw Intensities and creatinine normalized intensities ICC

After spectra pre-processing, 171 peaks were peaked and a datafile was created with the extracted data with the optimized signal detection method. Firstly, for each feature the raw signals were used to calculate the ICC value. ICCs were not normal distributed ($Shapiro\ Wilk: W = 0.976, p = 0.039$) with mean 0.357, SD 0.178, median 0.357 and IQR from 0.248 to 0.469. The with-in subjects variance was once again not normal distributed ($Shapiro\ Wilk: W = 0.131, p < 0.001$), with median $\sigma^2_w = 2266$, IQR from 606 to 27000. Therefore, for each patient, spectra features intensities were normalized by the individual urine creatinine measurement. By using creatinine normalized intensities, we calculated the ICC which was not normal distributed ($Shapiro\ Wilk: W = 0.976, p = 0.0039$) with mean 0.460, standard deviation 0.201, median 0.458 and IQR from 0.321 to 0.603. The within subjects variance was not normal distributed ($Shapiro\ Wilk: W = 0.125, p < 0.001$) with median 26.9, IQR from 6.5 to 325.5. Here we must consider that ICC is linear invariant, but creatinine have not the same value for all the subjects.

4.5.3 Median normalized features intensities ICC

Features intensities were median normalized (Equation 1.3). As specified above, median normalized intensities were obtained firstly by calculating each spectrum intensities median values and that value was used to normalize raw signals. So, because dividing by median values it is not merely a linear rescaling, ICCs differed from ICCs calculated by raw data. ICC was estimated for each normalized features. ICC were not normal distributed ($Shapiro\ Wilk: W = 0.961, p < 0.001$), with mean 0.4447, standard deviation 0.228, median 0. 445 and IQR from 0.266 to 0.624 (Figure 4.10). The within subjects variance was not normal distributed ($Shapiro\ Wilk: W = 0.125, p < 0.001$) with median 0.158 and IQR from 0.033 to 3.103.

When features intensities were median normalized (Equation 1.3) and re-normalized by creatinine levels the estimated ICC were not normal distributed ($Shapiro\ Wilk: W = 0.976, p = 0.0039$), with mean 0.230, standard deviation 0.132, median 0.222, interquartile range 0.124 - 0.312. Within subjects variance of normalized intensities were not normally distributed ($Shapiro\ Wilk: W = 0.142, p < 0.001$), with median 0.0152 and IQR from 0.0036 to 0.196.

Therefore, we tried the last approach. Features were firstly sLOD adjusted, median normalized and log-transformed (log_2). Estimated ICC were not normally distributed ($Shapiro\ Wilk: W = 0.977, p = 0.006$). Mean and SD were 0.473 and 0.222, while median was: 0.458 (IQR: 0.325 - 0.649). Within subjects variance was not normal distributed ($Shapiro\ Wilk: W = 0.527, p < 0.001$), with median 0.184 and IQR from 0.123 to 0.329.

Inspection of the error structure

Features' ICCs and with-in subject variances, calculated after sLOD adjustment, median normalization and log-transformation did not show a statistically significant correlation

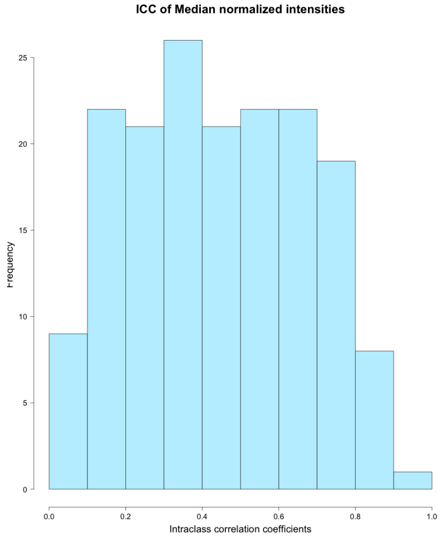

Figure 4.10: Histogram of the ICCs obtained after sLOD adjustment, median normalization and log_2 transformation of feature intensities for the external dataset.

with features m/z (Spearman ρ= -0.01, p = 0.922) and (ρ = -0.04, p = 0.626), respectively (Figure 4.11). Moreover, ICCs and with-in subjects variances were not correlated with median features signal intensities (Spearman ρ= 0.05, p = 0.522) and (ρ = 0.144, p = 0.06), respectively.

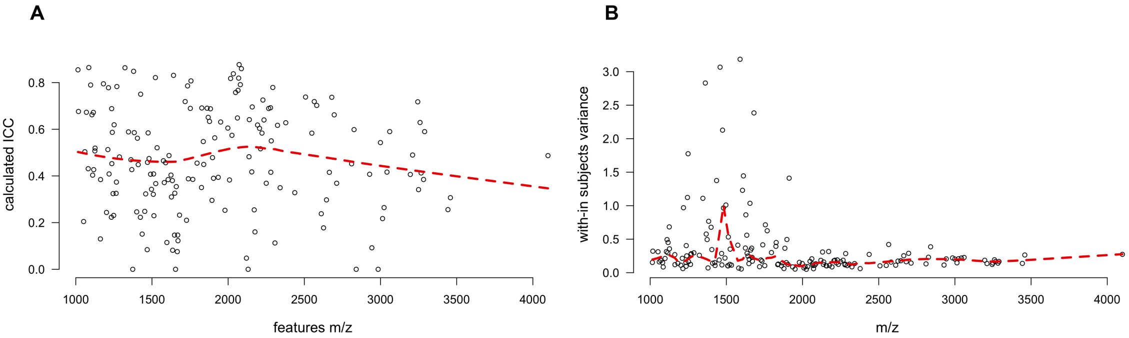

Figure 4.11: Inspection of the error structure of the MALDI-TOF/MS features intensities in the external dataset. Features intensities were sLOD adjusted, median normalized, $\log_2$-transformed. A) Feature's ICC and features m/z; B) Features with-in subject variances and features m/z. In both pictures, a lowess line (iteractions = 100 and smoother span = 0.4) were overlapped.

	ICC (median and IQR)	With-in subject variance (median and IQR)
Raw Data	0.36 (0.24-0.47)	2266 (606-27000)
Median Normalized	0.45 (0.26-0.62)	0.16 (0.033-3.103)
Median and Creatinine normalized	0.22 (0.12-0.31)	0.015 (0.003-0.196)
Median Normalized, sLOD adjusted and log_2-transformed	0.46 (0.32-0.65)	0.184 (0.12-0.33)

Table 4.4: Table reporting summary results of ICC, calculated after applying different normalization strategy to the features intensities of the external dataset.

4.5.4 Measurement error model assessment

Before applying the measurement error adjustment, some model assessments are needed. As suggested by Carroll et al., in an external dataset with replicate measurements, it is important to check whether the measured variable(s) and their error(s) have constant variance (homoscedasticity). In fact, under the classical error model U are symmetric and have constant variance, therefore, if W is replicable, its standard deviation should be uncorrelated with the individual mean. Moreover, because U has mean equal to zero, difference between replicates should have normal distribution with mean equal to zero (96).

Classical error model assessment

The following analyses have been performed on considering sLOD adjusted, median normalized features intensities, obtained by the optimized signal detection method on the external dataset identified features.

Constant variance

Because a standard way to remove non-constant variability is via a transformation (mainly log-transformation under the multiplicative error model), to assess the constant variance condition we generated two scatterplot of standard deviations *vs* the means of the features intensities, before and after log_{10} transformation (Figure 4.12).

Further, two Q-Q plots of the difference between replicates have been generated, representing data before and after log-transformation (Figure 4.13).

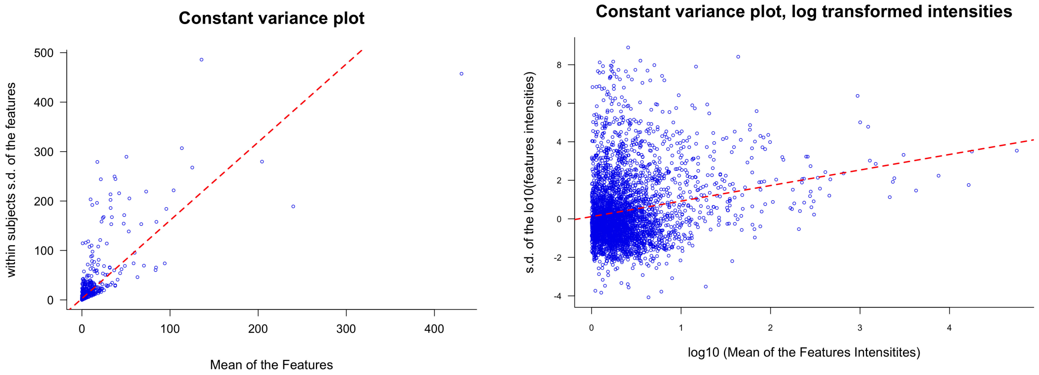

Figure 4.12: Scatterplots of variances with respect to the means of median normalized features intensities, before and after $\log_2$ tranformation.

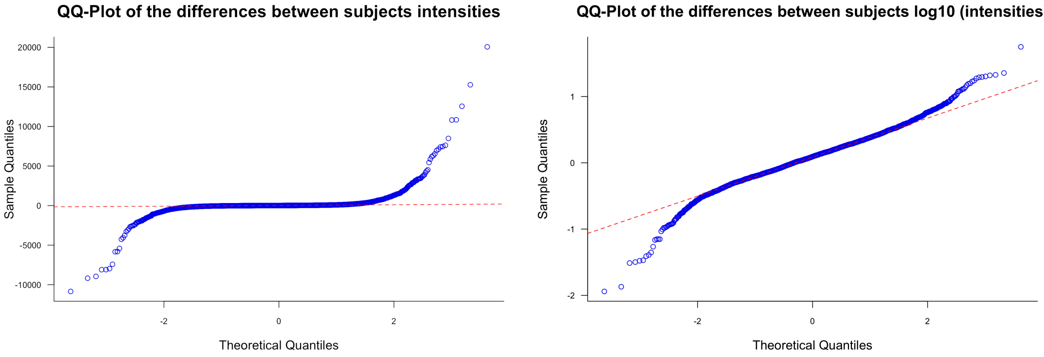

Figure 4.13: Q-Q plots of the difference between replicates of median normalized features intensities, before and after $\log_2$ tranformation.

Differences between measurement

The mean differences replications in features intensities was -0.23, with standard deviation of 16.84 and 95 % CI ranging from -0.796 to 0.334 while the median was 0.007 (IQR: -0.2610 to 0.2724). Similarly, the mean differences between subjects median and log-transformed intensities was -0.013, with standard deviation of 0.24 and 95% CI ranging from -0.018 to 0.002, while the median was 0.003 (IQR: -0.123 to 0.117).

4.6 ICC simulation under measurement error and/or left censoring conditions

Regression Calibration (RCAL) is a useful tool to account for biomarker measurement error, also when a "gold standard" instrument is not available to measure the "true" biomarker level. This approach can be applied to adjust for measurement error but, up to now, only a few evidences have been provided in case of variables that suffer both of

measurement error and detection limit problems.
We wanted to simulate data based on our "real" condition. So, on considering all the MALDI-TOF/MS identified features in the PCa dataset, they were sLOD adjusted, median normalized and $\log_2$ transformed. After, for each feature the mean and the corresponding standard deviation were calculated. The features means had mean 0.11 and median 0.10, while features standard deviations had median 0.823 and median 0.717. Therefore, the variable X was generated with mean 0.10 and $\sigma = 0.717$, while the resulting variables W_1 and W_2 have been generated as described in Materials and Methods. In this simulation, **the dataset containing W_1 and W_2 was defined as the "Full Dataset"**. To evaluate different left censoring conditions, we chose some appropriate W_1 and W_2 threshold levels, so that the amount of values below LOD were: 12.5 %, 25 % and 50 %. Values below LOD are treated as specified in Materials and Methods. In order to compare results, evaluations were also performed by using the full dataset. Different values for the error variance σ^2_ϵ were chosen, varying from 0.01 to 0.64. After N = 1000 repeated Monte Carlo simulations the mean ICC and the Monte Carlo ICC standard error were calculated. Results are summarized in Table 4.5 and in Figure 4.14.

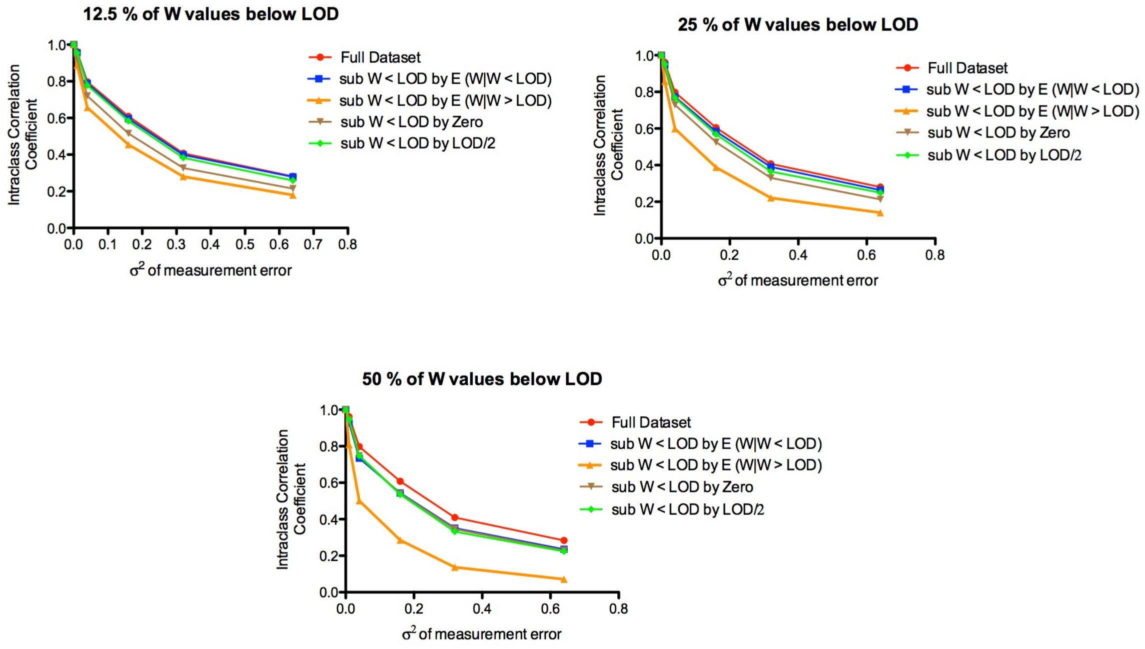

Figure 4.14: Estimated ICC after Monte Carlo simulation for a linear exposure-response association, when the esposure (W) suffers of both measurement error and detection limit problems. σ is the considered amount of measurement error included in simulation.

Monte Carlo Simulation Results for ICC and ICC's standard errors					
Parameters	**Full Dataset**	**Substitution of W < LOD by** $E(W \mid W < LOD)$	**Substitution of W < LOD by** $E(W \mid W > LOD)$	**Substitution of** $W < LOD$ **by zero**	**Substitution of** $W < LOD$ **by** $LOD/2$
12.5 % of values are set below LOD					
$\sigma = 0.01$	0.961±0.006	0.955±0.008	0.887±0.022	0.924±0.013	0.954±0.007
$\sigma = 0.04$	0.797±0.027	0.788±0.029	0.657±0.042	0.720±0.038	0.778±0.032
$\sigma = 0.16$	0.609±0.048	0.596±0.048	0.455±0.063	0.516±0.060	0.584±0.052
$\sigma = 0.36$	0.407±0.062	0.398±0.065	0.280±0.073	0.327±0.070	0.383±0.064
$\sigma = 0.64$	0.281±0.069	0.279±0.068	0.179±0.076	0.215±0.075	0.259±0.076
25 % of values are set below LOD					
$\sigma = 0.01$	0.961±0.006	0.947±0.009	0.857±0.025	0.936±0.012	0.952±0.008
$\sigma = 0.04$	0.798±0.027	0.788±0.029	0.657±0.042	0.720±0.038	0.778±0.032
$\sigma = 0.16$	0.604±0.050	0.584±0.053	0.388±0.069	0.525±0.063	0.568±0.058
$\sigma = 0.36$	0.407±0.062	0.391±0.066	0.221±0.081	0.330±0.076	0.365±0.071
$\sigma = 0.64$	0.280±0.068	0.264±0.069	0.140±0.083	0.213±0.077	0.249±0.072
50 % of values are set below LOD					
$\sigma = 0.01$	0.961±0.006	0.928±0.015	0.808±0.035	0.945±0.012	0.945±0.011
$\sigma = 0.04$	0.798±0.027	0.735±0.042	0.501±0.079	0.748±0.043	0.747±0.045
$\sigma = 0.16$	0.608±0.048	0.543±0.058	0.285±0.095	0.540±0.066	0.535±0.067
$\sigma = 0.36$	0.409±0.063	0.351±0.071	0.137±0.100	0.346±0.079	0.332±0.081
$\sigma = 0.64$	0.283±0.068	0.235±0.076	0.071±0.090	0.229±0.082	0.224±0.081

Table 4.5: Table reporting the ICC Monte Carlo's mean and standard error for a linear exposure-response association, when the esposure (W) suffers of both measurement error and detection limit problems. σ is the considered amount of measurement error included in simulation. Monte Carlo ICC standard error has been calculated as specified above, by equation 3.4. **Full dataset** referred to the dataset without left censoring data.

Patients groups		N	Age (mean ± SD)
References		**106**	65.4 ± 6.7
	No alteration	60	65.3 ± 6.7
	BPH	15	64.7 ± 8.3
	Inflammation	14	66.2 ± 4.3
	AAH	2	61.5 ± 0.7
	PIN	4	69.0 ± 5.0
	Atypical proliferation	11	64.7 ± 7.7
PCa		**72**	66.8 ± 6.7

Table 4.6: Patients data summary, subdivided by their histological results of prostate biopsy. No alteration = no evidence of any tissutal alteration, BPH = Benign Prostatic Hyperplasia, AAH = Atypical adenomatous hyperplasia, PIN = high-grade prostatic intraepithelial neoplasia, PCa = prostate cancer.

4.7 PCa database urine evaluation by MALDI-TOF/MS

The registry collected were firstly validated for completeness. For eleven patients, the prostate biopsy results were missing, therefore data were recovered by revising medical records. In the bio-bank of urine samples, only 188 samples matched with the patients registry (9% incompleteness). Unfortunately, Gleason score data was missing for most of the records. After MALDI-TOF/MS analysis of the 188 specimens, for 9 samples the analysis was unsatisfactory (5%). For 2 samples, spectra showed polymers, which completely masked peptides/proteins signals while 7 samples displayed a blank spectrum, probably because urines were too much diluted. Differently, a total of 178 spectra gave satisfactory results. MALDI-TOF/MS analysis was performed on dialyzed urines after spiking IS at the specified concentration.

Because patients at accrual time referred generic Lower urinary tract symptoms, prostate biopsy results underlined how they are heterogeneous in histology, and it was possible to identify 7 different groups of subjects. However, according to Urologists, patients were reallocated into two groups, references and PCa. Characteristics of evaluated patients are reported in Table 4.6

Spectra were evaluated for peak peaking and 482 features were identified in a m/z ranging from 1000 to 4000 Da. So variable names were generated according to the features m/z values. After, raw signals were extracted from the raw data at the corresponding m/z values by the **optimized signal detection method**. Moreover, starting from the m/z values, sLOD were estimated by equation 4.1. Extracted abundances were firstly adjusted for sLOD, median normalized and after log_2-transformed.

Features name	Spearman's ρ	p-value
1322.5	-0.245	< 0.001
1373.7	0.194	< 0.01
1377.7	0.197	< 0.01
1474.8	0.242	< 0.001
1641.8	-0.207	< 0.01
1727.2	0.289	< 0.001
1950.0	-0.212	< 0.01
2087.1	-0.194	< 0.01
2088.1	-0.208	< 0.01
2272.3	-0.209	< 0.01

Table 4.7: Spearman's correlation between age and features LOD adjusted, Median normalized Intensities.

4.7.1 Unsupervised clustering of MALDI-TOF/MS profiling data

Divisive hierarchical clustering results, obtained considering all the identified MALDI-TOF/MS features and all the Patients' groups is reported in Figure 4.15. Repeated analysis performed considering only the groups PCa patients and subjects with no evidences of histological alterations (called **extreme analysis**) was reported in Figure 4.16.

4.7.2 Effects of Age as confounder

On considering all the MALDI-TOF/MS identified features in the PCa study their Spearman correlation coefficients ρ_s varies between -0.245 to 0.289, with mean -0.035, standard deviation 0.231, median -0.208 and interquartile range from -0.209 to 0.196. Age was found to be statistically significant correlated (p<0.01) with ten MALDI-TOF/MS features (Table 4.7).

As expected, age was correlated with diagnosis, being the Spearman's $\rho = 0.197$, p < 0.001.

4.7.3 Features' outcome prediction by *naïve* logistic regression

For each feature, a logistic regression was evaluated, by considering the patients' disease status as outcome. After this analysis, eight features at m/z 1404.7, 1556.0, 1688.0, 1707.1, 1755.7, 1782.1, 2594.3 and 2797.8 were found to be statistically associated with the outcome as reported in Table 4.8. Therefore, a logistic regression analyses were repeated by considering the patients' disease status as outcome, but including age as confounding factor.

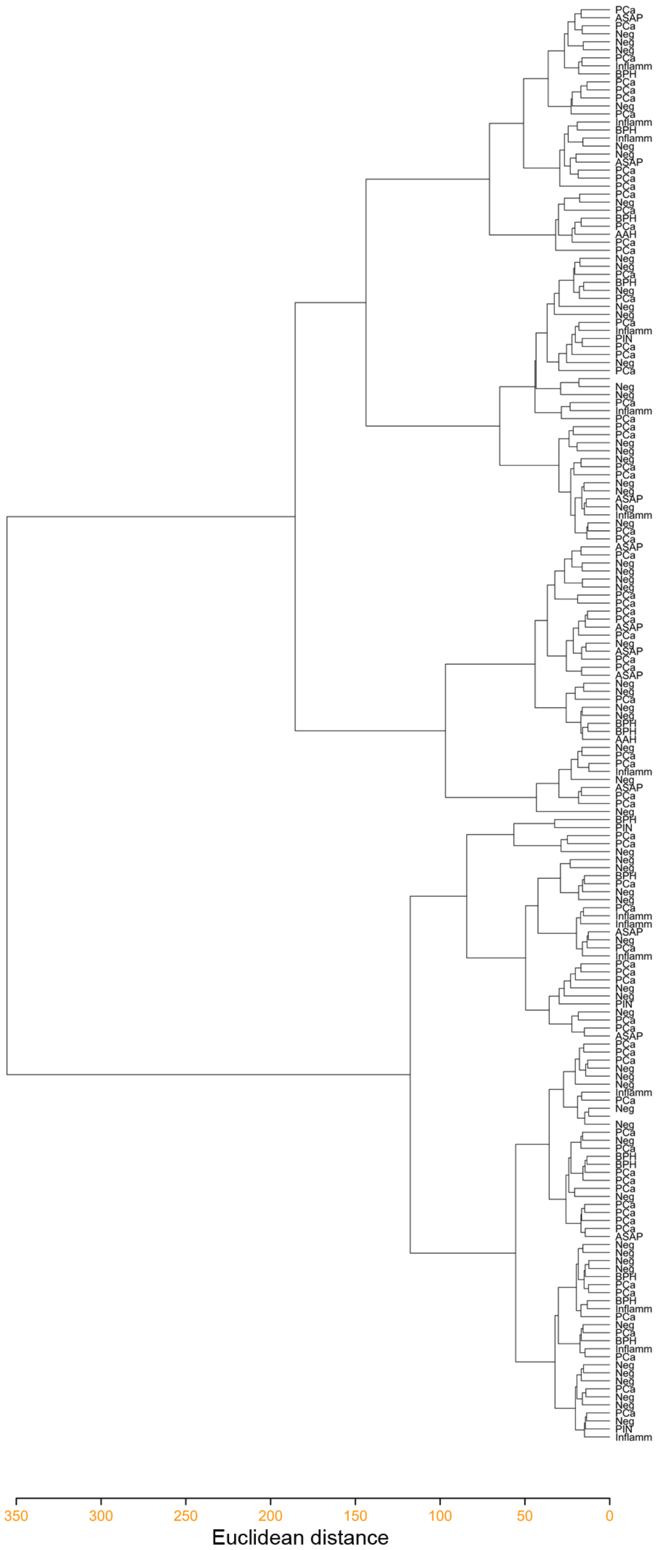

Figure 4.15: Dendrogram drew by considering all the identified features and all the patients' groups. Neg = no evidences of any histological alterations, BPH = benign prostatic hyperplasia, Inflamm = Inflammation, AAH = Atypical adenomatous hyperplasia, PIN = high-grade prostatic hyperplasia, ASAP = Atypical proliferation.

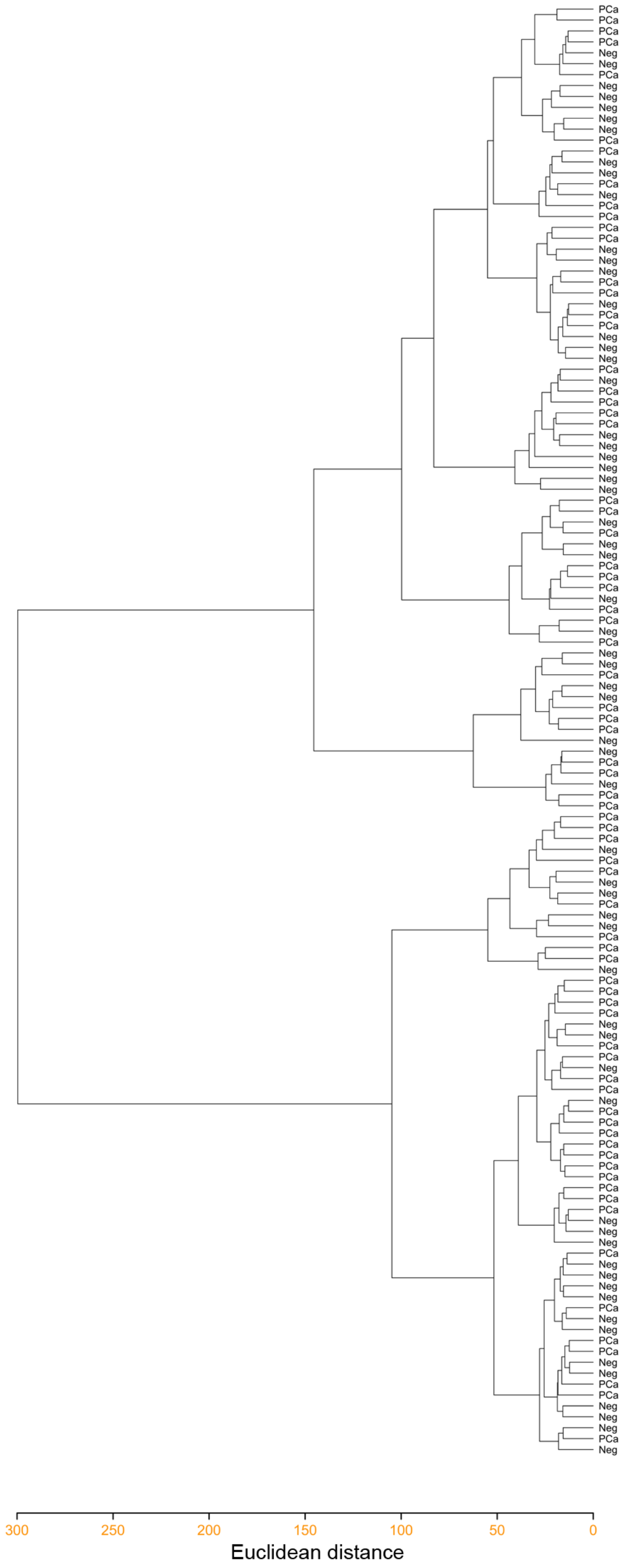

Figure 4.16: Dendrogram drew by considering all the identified features and the "extreme" patients' groups, namely PCa patients and subject with no evidence of any histological alteration.

***Naïve* Logistic regression**							
	Crude Analysis			**Age Adjusted Analysis**			
Features name	**$\hat{\beta}$ coefficient**	**95% CI**	**p-value**	**$\hat{\beta}$ coefficients**	**95% CI**	**p-value**	**Goodness of fit p-value**
1404.7	0.42	0.08 to 0.762	0.016	0.42	0.08 to 0.762	0.016	0.259
Age	-			0.03	-0.02 to 0.08	0.196	
1556.0	-0.76	-1.38 to -0.12	0.017	-0.76	-1.38 to -0.12	0.017	0.779
Age	-			0.03	-0.01 to 0.08	0.068	
1688.0	-0.71	-1.21 to -0.14	0.014	-0.79	-1.38 to -0.21	0.008	0.453
Age	-			0.04	-0.01 to 0.09	0.065	
1707.1	-0.92	-1.61 to -0.24	0.008	-0.93	-1.62 to -0.24	0.008	0.385
Age	-			0.04	-0.01 to 0.08	0.132	
1755.7	-0.25	-0.48 to -0.03	0.035	-0.26	-0.48 to -0.03	0.025	0.539
Age	-			0.03	-0.016 to 0.074	0.211	
1782.1	-0.76	-1.50 to -0.03	0.040	-0.79	-1.55 to -0.06	0.034	0.054
Age	-			0.04	-0.01 to 0.08	0.122	
2594.3	-0.54	-1.03 to -0.06	0.029	-0.53	-1.01 to -0.03	0.037	0.578
Age	-			0.03	-0.01 to 0.08	0.181	
2797.8	-0.51	-1.01 to -0.02	0.042	-0.52	-1.01 to -0.02	0.042	0.079
Age	-			0.03	-0.01 to 0.08	0.139	

Table 4.8: Logistic regression results of the statistically significan features by considering the Patients disease status as outcome and, where reported, age as confounding factor. Goodness of fit tests were calculated for the age adjusted model by the Hosmer-Lemeshow method, on considering 10 quantiles for groupping data.

RCAL results					
		Age Adjusted Analyses			
Features name	$\hat{\beta}$ coefficient	$\hat{\beta}^*$ coefficient	$\hat{\beta}^*$ 95% CI	p-value	ICC
1404.7	0.42	0.91	0.17 to 1.65	$<$ 0.001	0.459
1556.0	-0.76	-1.58	-2.89 to -0.26	$<$ 0.001	0.480
1688.0	-0.79	-1.64	-2.86 to -0.42	$<$ 0.001	0.480
1707.1	-0.93	-1.94	-3.37 to -0.51	$<$ 0.001	0.480
1755.7	-0.26	-0.56	-1.04 to -0.07	$<$ 0.001	0.455
1782.1	-0.79	-1.58	-3.12 to -0.04	0.02	0.480
2594.3	-0.53	-1.09	-2.12 to -0.06	$<$ 0.001	0.480
2797.8	-0.52	-1.08	-2.11 to -0.04	$<$ 0.001	0.480

Table 4.9: Regression calibration (RCAL) results of the statistically significant features. $\hat{\beta}$ = *naïve* logistic regression coefficient; $\hat{\beta}^*$ = measurement error adjusted coefficient by RCAL method. Wald test derived p-values for $\hat{\beta}^*$ were also reported. ICC = intraclass correlation coefficient, calculated from the external dataset and used for adjusting $\hat{\beta}$

4.8 PCa outcome prediction adjusted for measurement error

For MALDI-TOF/MS features found to be statistically significant associated with Patients' disease status by *naïve* logistic regression, the RCAL and SIMEX analyses were performed, to obtain the measurement error adjusted $\hat{\beta}^*$ coefficient. Both analyses were made by using the information on error structure obtained by the external dataset, calculated by considering sLOD adjustment, median normalized and $\log_2$ transformed features intensities. Because many features identified as statistically significant in the *naïve* logistic regression had not a corresponding error structure in the external dataset, for these ones median ICCs and median with-in subjects variances were chosen for the RCAL and the SIMEX analyses, respectively.

4.8.1 Regression calibration and SIMEX results

Results obtained after RCAL and SIMEX analyses were reported in Table 4.10 and Table 4.9.

SIMEX results				
		Age Adjusted Analysis		
Features name	$\hat{\beta}^*$ **coefficient**	**95% CI**	**p-value**	σ_U
1404.7	0.59	0.08 to 1.10	0.024	0.344
1556.0	-1.24	-2.18 to -0.31	0.010	0.184
1688.0	-1.25	-2.11 to -0.38	0.005	0.184
1707.1	-1.62	-2.85 to -0.41	0.009	0.184
1755.7	-0.35	-0.66 to -0.04	0.026	0.714
1782.1	-1.44	-2.84 to -0.04	0.043	0.184
2594.3	-0.77	-1.52 to -0.013	0.046	0.184
2797.8	-0.77	-0.157 to 0.03	0.059	0.184

Table 4.10: SIMEX Logistic regression results of the statistically significan features by considering the Patients disease status as outcome. $\hat{\beta}^*$ = measurement error and age adjusted $\hat{\beta}$ coefficients of features association with outcome, calculated by logistic regression. σ_U are the with-in subject variances calculated from the external dataset and considered in the SIMEX model.

4.9 Analytical goals and maximum allowable analytical variation

In Clinical Chemistry literature, it is widely accepted than for measurands the maximum allowable analytical variation (CV_A) should be less than or equal to half the average within subject biological variation (CV_I) (97). However, for particular analytes, the desirable precision could be less stringent, being:

$$CV_A \leq 0.75 \cdot CV_I \tag{4.2}$$

Under this condition ($\sigma_A/\sigma_I \leq 0.7$), using the following formula:

$$\sigma_w^2 = \sigma_A^2 + \sigma_I^2 \tag{4.3}$$

it is possible to derive that $\sigma_w/\sigma_I \leq 1.30$, which is equal to a 30% increase of σ_w. In order to estimate the features analytical variability we choose the **inter-run experiment**, as variance include the instrumental and the sample pre-processing (dialysis) variabilities. The variances calculated for the median normalized, sLOD adjusted features intensities, they are not normally distributed with median variance $\sigma_A^2 = 0.04$, while in the external dataset the median normalized feature intensities sLOD adjusted had a median value of $\sigma_w^2 = 0.184$. So, it is possible to calculate the $\sigma_I^2 = \sigma_w^2 - \sigma_A^2 = 0.149$ and derive the ratio $\sigma_w/\sigma_I = \sqrt{0.184}/\sqrt{0.149} = 1.13$, which is less than the desirable precision, so matching the minimal desirable analytical performances on considering features median variability.

5

DISCUSSION

Early prostate cancer is mostly a painless and indolent disease. However, when tumor causes the prostate gland to swell, symptoms can be confused with benign prostatic hyperplasia, as both of them share manifestations which are usually described in medicine as Lower urinary tract symptoms. Despite PCa screening is currently recommended by many National Public Health Systems, general practitioner are often encouraged to counsel and assist men who enquire about testing or ask to enter in the PCa screening program. In fact, screening benefits have some important limitation and the definite risks/advantages, associated with active monitoring programmes, are actually unknown. For example, men with undetected PCa can be falsely reassured by a low PSA concentration and/or negative DRE results; on the contrary, some trivial, low risk tumours which are detected during surveillance, can lead in a worthless patients over-treatment which can cause side-effects on sexual, urinary, and bowel function (98).

The American Urological Association (AUA) and European Association of Urology (EAU) have recently released new guidelines for PCa early detection (99, 100). Because the benefit/harm profile of PSA-based prostate cancer screening is highly age dependent, the AUA guidelines diversify screening recommendation on considering subjects' age and PCa family history. The four identified screening panels are 1) men less than 40 years of age; 2) high risk men (with positive family history or African-American race), aged from 40 to 54; 3) men with age from 55 to 69 years; 4) men with age 70 years or more, or men with less than a 10 to 15 year life expectancy. In brief, screening is not recommended for men of class 1, 2 and 4, because the greatest benefit of screening appears to be in men ages between 55 and 69 years (99). Differently, the EAU has a different recommendation for PCa early detection. Despite EAU guidelines do not recommend widespread mass screening for PCa, they do strongly recommend early detection in well-informed men and do not use a specific chronological age as a threshold for screening, but screening in men with a life expectancy >10 years is always recommended. Further, the baseline PSA determination starting at 40 - 45 years of age has been suggested upon which the subsequent screening interval can then be based (100). AUA and EAU guidelines recommendation are based on recent published results of two large trials, the European randomized study of screening for prostate cancer (ERSPC) and the US prostate, lung, colorectal, and ovarian (PLCO) cancer screening, which are based on DRE and PSA. On ERSPC, the evaluation of the association between total PSA in blood and biopsy outcome among unscreened men showed that PSA has moderate predictive value for positive prostate biopsy, as the positive predictive value is between 20% and 25% for a PSA value of 4 ng/ml, and increases to 50% for a value of 10 ng/ml (101). In this scenario, PSA are still unmet desirable specificity and sensitivity; however, there is also a need for new biomarkers that may enhance not only PCa detection but also prediction as to whether a biopsy-detect cancer may rather be left untreated, as such pose little if any threat to the quality or length of life of the individual tumour host.

Biomarker research is in focus at many laboratories and several biomarkers seems to be promising for PCa diagnosis. However, to date, only few of these biomarkers have shown to be really useful for PCa early detection or in predicting biopsy results in symptomatic men. Recently, a urine molecular assay, PCA3 has been approved from FDA because PCA3 score < 25 seems to be associated with a decreased likelihood of a positive biopsy, and some studies have demonstrated it effectiveness beyond PSA in re-biopsy (102, 103). However, PCA3 role in predicting prostate biopsy outcome is still debated. For example Auprich et al. showed that %fPSA is more informative than PCA3 in predicting PCa for the first biopsy, while PCA3 is better in predicting outcome of the re-biopsy (104). Further, the PCA3 advantages in diagnostic accuracy for re-biopsy seems to be dissipated at third and $\geq$ third repeat biopsy. Prostate health index, a new serum testing for PCa, was also shown to increase the tPSA and fPSA in detecting prostate biopsy results, but more studies are needed to demonstrate its superior clinical performances (27). In addition, a series of genetic alterations and alternative splicing variants of PSA (PSA-RPs) or candidate proteins (mainly matrix metalloproteinases, annexin 3, etc.) have been proposed as putative useful serum PCa markers (18, 19, 20, 25).

Besides serum/plasma, urine is another ideal body fluid for the detection of markers produced by PCa, especially for proteomic biomarkers, because it contains secreted and directly shed proteins from prostate. Many studies identified putative **urine genetic markers** (34, 35), while others evaluated **urine proteome** by MS protein profiling (39, 40, 41, 42, 43, 44, 45, 46, 47, 48). For these latter analyses, MALDI-TOF/MS and SELDI-TOF/MS have been mainly used because their high throughput and for the high number of detectable peptides/proteins for single run. However, some crucial topics like sample pre-cleaning, intra- and inter-labs reproducibility have been widely debated since protein profiling studies were introduced, debasing the importance of Researchers findings. In addition, studies derived from SELDI-TOF/MS and MALDI-TOF/MS are not directly comparable. For example, SELDI-TOF/MS, which is now dismissed form BioRad©, allowed a proprietary on-chip direct sample pre-cleaning, while MALDI-TOF/MS analysis requires an adequate choice of samples pre-treatment strategies. Moreover, due to the particular characteristics of SELDI-TOF/MS and MALDI-TOF/MS, biological fluid used for the analysis and sample pre-cleaning is intimately linked with instrumental reproducibility. Therefore, optimizations should be adequately evaluated and chosen well in advance, before starting the biomarker identification study (63). Other authors have shown that also data normalization is a crucial steps to increase comparability of spectra resulting from different measurements and, more importantly, to allow a correct statistical analysis of the identified features (83).

This study consists of 3 major parts. The **first part** regards: a) the evaluation of sample pre-treatment procedure, b) the estimation of MALDI-TOF/MS features signal detection limits and c) the assessment of MALDI-TOF/MS analyses reproducibility, performed on dialyzed urine samples. The **second part** was designed and conducted to estimate the error structure of the MALDI-TOF/MS identified features. The **third part** aimed to evaluate whether MALDI-TOF/MS identified features in urine by using the reflectron mode are associated with PCa, in patients referring lower urinary tract symptoms (LUTS) at consultants. Moreover, measurement error adjusted coefficient obtained by logistic regression were estimated.

5.1 Samples pre-preatment

It is widely known that MALDI-TOF/MS analyses requires only simple pre-treatment steps, because this methodology performance is scarcely influenced by salt and contaminants. However, it should be considered that the major effects of salts and contaminant are to increase across spectra variability. Because urine is generated from kidney after blood filter and concentration, reducing cations, anions and small molecules can lead to an increased reliability of MALDI-TOF/MS profiling analyses. Urine, when directly analyzed by MALDI-TOF/MS, showed a large number of peaks (data not shown), but, as expected, the noise component of the signal was elevated and shared across spectra. Dialysis is a simple procedure, applicable to any biological fluid, that was extensively used in the past to desalt and purify urine for many types of analyses. With respect to other sample processing methods, dialysis allows the removal of small, unwanted compounds from macromolecules in solution by selective and passive diffusion through a semi-permeable membrane. More interestingly, with a careful selection of membrane MWCO, during dialysis, peptides larger than membrane MWCO are not lost and remains in solution. Overall, this process is called buffer exchange. We assessed the performances of the dialysis system on urine (Table 4.1) and we verified that major cationic and anionic salt concentrations decreased. After dialysis, urine spectra had a decreased samples noise and an increased intensities of the detected signals.

In addition, we wanted to verify whether adding IS to samples modify profiling results. Clinical Chemistry mass spectrometry measurements commonly required IS addition to samples, especially when performing absolute quantitation. In fact, an appropriate internal standard will control for sample pre-treatment and ionization variability. The best IS for MS is an isotopically labeled version of the molecule under analysis. However, in MALDI-TOF/MS profiling, labeled IS is not suitable, because peptides/proteins are not *a priori* known. So, we compared detected features intensities by analyzing spiked and non-spiked urine samples. Obtained results (Figure 4.1) showed that spiking IS resulted in an increased features' signals, especially for the low abundant ionic species. This was not surprising, because Toghi Eshghi et al. demonstrated that spiking a peptide (carrier analyte) in a complex mixture, improved the signal of the contained analytes (targeted analytes), exerting a carrier effect. Therefore, we choose to spike IS in all the analyzed samples, not only to evaluated IS as a possible normalizing factor, but also for its carrier effect on low abundant ionic species.

5.2 MALDI-TOF/MS urine profiling signal detection limit estimation

MALDI-TOF/MS analyses give as results a mass spectrum, which consists on a file, made of two columns, containing more than one hundred thousand of points (typically 150k points). The process by which features are extracted from spectra is called **peak detection** or **peak peaking**. Many public peak detection algorithms have been recently published (see Introduction), but none of them have been really optimized for MALDI-TOF/MS reflectron mode spectrum analysis. Furthermore, all peak finding methods are known to be *error prone*, and their performances are difficultly estimabled, especially in protein profiling studies. However, peak peaking performances can be roughly evaluated by sensitivity and FDR. Sensitivity measures the amount of true signals which are correctly identified as such, while FDR is the percentage of false signals that are considered as "true peaks". Peak peaking algorithms are generally designed to obtain a *reasonalbe* compromise between specificity and FDR, most of times to favor FDR reduction. Therefore a small-moderate number of low abundant peptides are usually not detected by algorithms and few low abundant peaks are erroneously detected as signals. As result, the profiling features intensities matrix will contain many zeros, which correspond to "undetected" signals rather than "true" zero signals. In addition, trying to optimize low abundant peptides may lead to an enormous number of noisy signals detected as *true* signals, which highly increase the FDR. Based on these considerations, when many samples are being considered (e.g. like in proteomic profiling experiments), identified features signals will be highly variable, irrespectively of their real across spectra variability, due to erroneous handling of signals from peak peaking algorithms. Thus, this virtual variability not only decreases signals reliability but also heavily influences statistical data analysis results.

To handle this situation it would be desirable to modify the commonly used steps in protein profiling workflow (Figure 1.9). For example, a possible solution to deal with the excess of zeros is to consider the identified features as **left censored variables**. However, an instrumental derived LOD should be estimated *prior* any evaluation steps. By using a calibration function, an equation relating the instrument output signal to the analyte concentration, is possible to define a signal detection limit (sLOD) and, most importantly, the LOD, defined as the minimum concentration of substance that can be detected by the instrumentation with a predefined precision. However, in protein profiling studies, because the discovered peptides are not *a priori* known, it is not possible to estimate analytical LODs by the canonical way. In addition, although features signals are proportional to peptide concentration, MALDI-TOF/MS absolute peptide amount is not measurable, leading to across features variation of analytical LOD. Toghi Eshghi et al. described how to calculate LOD for MALDI-TOF/MS (91). In this work we tried to use features sLOD as a convenient surrogate of analytical LOD. sLOD is the **signal background noise**, estimated when analyte concentration is equal to zero. Despite protein profiling does not allow to evaluate samples with defined concentrations of analytes, it is possible to obtain similar results by serial diluting samples until obtaining the zero amount of any contained peptide. So, features' sLODs can be calculated by equation 1.8. This approach appears to be appropriate for MALDI-TOF/MS profiling as it is based on signal intensities and relative quantification of measurands rather than on peptides concentration and their absolute quantification.

5.3 Reproducibility study

In the reproducibility study we aimed to evaluate the intra-run and inter-run spectra variabilities, based on technical replicates of pooled urines. In particular, intra-run variability estimates MALDI-TOF/MS instrumental repeatability, while inter-run variability encompass not only instrumental but also sample pre-treatment variability, as dialysis can increase the overall variability.

5.3.1 Intra- and inter-run variability assay

As MALDI-TOF/MS analyses generated many variables after spectra pre-processing and peak peaking, intra- and inter-run variability need to be investigated for all these features. Firstly, we assessed the features CVs *vs* their abundances, to discover trends in variability with respect to instrumental signal intensities. Overall CVs were estimated by their median values, because it is a more robust estimation with respect to the mean. It was reported by Duncan et al. that features' variability decrease when their signals increases (86). Our results obtained after **automatic peak peaking** by Flex Analysis, confirmed findings previously reported and, as expected, the percentage of features found to be below sLOD was strictly inversely correlated with signals intensities. MS data can varied across replicates also for instrumental signal non linearity. Normalization can notably increase spectra comparability. Moreover, normalizations can act similarly to standardization. It is well known that for supervised and unsupervised machine learning algorithms it is important to standardized features intensities, because the most representative feature can lead algorithm optimization to consider that feature as more relevant, independently of its real importance for the accuracy of class prediction. Therefore we applied different normalization strategies: mean, median, internal standard (IS), relative intensities (RI), total ion current (TIC) and linear rescaling. All these methods are **local normalizations** because scaling parameters are calculated by considering single spectrum. RI and TIC are widely used in MALDI-TOF/MS. RI is based on relative quantification of ionic species relative to the most abundant peak, while TIC is performed by dividing abundances by the sum of all the different ions contributing to the spectrum. So, TIC normalization is similar to normalize for the total protein content. Differently, mean and median normalizations are two normalization strategy widely used for genetic data, that recently has beed suggested as a better choice than TIC, also for MALDI-TOF/MS (81, 82). Results showed a minimal reduction in the overall CVs, both for the intra- and the inter-run assay (Table 4.2). By adjusting features' values below sLOD with sLOD/2, CVs for low abundant features drastically decrease (Figure 4.3), although the overall median CVs didn't change significantly (data not shown). These results supported the hypothesis that low abundant features, namely peptides/proteins present at low concentration in samples, were detected by MALDI-TOF/MS with high uncertainty. TIC was not estimated after sLOD adjustment because this normalization require to calculate the sum of spectrum intensities, which markedly change after that below sLOD signals were substituted with sLOD/2.

5.3.2 Optimized signals detection for MALDI-TOF/MS profiling

Although algorithms designed for MALDI-TOF/MS linear model may be applied also to MALDI-TOF/MS reflectron mode, low abundant features are generally detected with very low sensitivity. Therefore, we wanted to verify whether applying a feed-back process in peak detection it was possible to decrease the overall resulting features variability. Basically, instead of focusing on peaks obtainable from a single spectrum, we consider the complete peak-list given by the set of all spectra. So, starting from the m/z peaks list identified by Flex Analysis, all the spectra were re-evaluated at each m/z position, in order to peak the new signals. Once obtained the new signals, all the evaluations already made were repeated.

Optimized signal detection method caused a dramatically decrease of overall features variability (Table 4.3), for both the intra- and the inter-run studies and sLOD adjustment caused a further slightly decrease of variability. Moreover, the behaviour of features CV *vs* Abundances became almost linear. Surprisingly, IS normalization gave extremely different results. In both the intra-run and the inter-run studies, IS normalization caused an increase of overall CVs. This unexpected effect underlines that normalizing abundances by a signal generated from an internal peptide spiked at a know amount, is not feasible for MALDI-TOF/MS profiling analysis. Some speculation can be done to explain IS results. For example, IS can act as carrier analyte, exerting an effect which is strictly dependent on the analyzed sample. Further, despite signals in MALDI-TOF/MS are proportional to peptides/proteins concentration, IS signals can be different across sample due to ionization or crystallization effects. However, these findings is not generalizable and should not be considered in case of other MALDI-TOF/MS experimental set-up like, e.g., absolute quantification.

As further step, we evaluated the across spectra variability of both studies by means of box plots, by comparing data **before and after sLOD adjustment**. Technical replicates evaluation provided information on the random and systematic variability that occurs in performing assays. So, after normalization replicates signals should be overlapping as much as possible. Raw data showed that signals medians of the identified features are not constant across replicates of the same sample (Figures 4.6 and 4.7).

However, median normalization gave good results, because spectra signals and their variability became almost overlapping. So, **median normalization appeared as the best choice to minimize both random and systematic across replicates variability**. sLOD adjustment acted by further decreasing internal spectra signals variability.

5.3.3 Representativeness effect evaluation by features exclusion

Many protein profiling studies, during statistical data analysis, exclude less representative features. This rational approach is based on the perception that features not shared across samples will poorly contribute to outcome prediction and cause models overfitting, leading to an erroneous choice of the less parsimonious algorithm. These are commonly referred as the **representativeness effect**. However, less representative features not only can contain valuable informations but also may be associated with patients' outcome. As explained above, a feature can be considered **undetected** when its signals fall below sLOD while **representativeness** represents the percentage of spectra in which that feature is "detected". We chose to analyze whether, with the real data obtained from the reproducibility study, the exclusion of less representativeness features can ameliorate the overall CV. Therefore, an iterated analysis was performed by gradually excluding features based on their representativeness. Moreover, standard error of the obtained CVs were calculated by means of bootstrapping techniques. The automatic peak detection by Flex Analysis method showed that the representativeness effect strongly influenced the overall median CV. In particular, for both the intra- and the inter-run studies, discarding that features with "undetected" percentage above 90% cause CV to halve, also by using data normalization (Figure 4.8). Moreover, discarding all the features which were resulting as undetected in at least one spectrum, CVs fallen below 40 %.

By using the optimized signal detection approach (with the except of Linear normalization), by varying features representativeness CVs remained almost constant (Figure 4.9). This result showed that, under a correct estimation of features signals, excluding less representativeness features seems to be not a necessary choice for MALDI-TOF/MS protein profiling studies.

5.4 External dataset and measurement error structure

Commonly used approaches to dealing with non differential measurement errors are based on **regression calibration** (RCAL) and **SIMEX**. The basis of RCAL is the replacement of X by the regression of X on (Z,W), which is called **calibration function**. It has been shown through theory as well as through a detailed simulation study that when the disease is rare, the relative risk is not large, and the measurement error is not large, adjusting β coefficients by the coefficient of bias (namely ICC), will remove most of the bias due to measurement error in the measured variable. To evaluate the calibration function, data sources can be internal or external. Moreover, replicated data, validation data or instrumental data can be collected to the purpose. In this work, PCa dataset contained no replicated urines collection. Some reasonable justification can be made. Firstly, DRE is much more invasive than blood venipuncture or urine collection and requires an Urological medical examination. As results, DRE could decrease patients compliance in prospective study as some patients may feel uncomfortable in repeating DRE. Secondly, as prostate manipulations by DRE can cause physiological alteration of the gland and as accrued patients rapidly undergone to prostate biopsy, as consequence further medical examinations, especially DRE, should be arranged far from biopsy. In this context it should be noticed that in order to estimate the **measurement error variance**, rather than measuring long term variations of the candidate biomarkers, repeated measurements of urine must be taken in a short period. Thirdly, a large part of patients found to be positive for PCa undergone to radical or partial prostatectomy, leading to exit these patients from the study.

To overcome the problems of a repeated collection of urines after DRE, we decided to collect an external dataset from male subjects apparently in healthy conditions, which have not previously referred LUTS, with age ranging from 23 to 56 years. Replicated urine samples were collected in a short time period, without performing DRE. Samples were evaluated also for Creatinine, which is a useful indicator of urines dilution. ICC and within-subject variances were calculated for samples Creatinine and for all the MALDI-TOF/MS identified features. Moreover, features ICC were calculated after: a) Creatinine normalization, b) median normalization, c) median and creatinine normalization and d) sLOD adjustment plus median normalization and log_2 transformation. Best results in terms of ICC (Table 4.4) were obtained for the creatinine normalized features and for features after sLOD adjustment, median normalization and log_2 transformation. Because ICC expresses also the reliability coefficient, results underline how normalizing by Creatinine did not decrease the within-subjects with respect to the between-subjects variance as well as it did not ameliorate MALDI-TOF/MS reliability. So, it is our opinion that creatinine normalization is not a mandatory choice for MALDI-TOF/MS profiling of urine samples.

Inspecting the error structure, both ICCs and with-in subject variances, calculated after features sLOD adjustment, median normalization and log_2 transformation, were neither correlated with features m/z nor with their abundances. Therefore, there was not a remarkable behavior of measurement errors in MALDI-TOF/MS features. Figure 4.10 showed that ICC distribution, being platykurtic, is not-normal. More importantly, median ICCs was similar to measured creatinine ICC. Creatinine determination was made by a high precision instrumentation, which meet the required Clinical Chemistry analytical quality specifications. By considering the instrumental error negligible, we can conclude that the creatinine reliability, which is far from good, may be mainly due to urine biological variability rather that *instrumental + biological variability.* Although urinary creatinine is not a "true" biomarker and so it is not supposed to behave as such, we can suggest that MALDI-TOF/MS profiling data contains a large quantity of measurement error, which may be intrinsic to urines and their physiological body production, rather than MALDI-TOF/MS instrumental error.

5.4.1 Measurement error model assessment

As described by Carroll et al., additive measurement error impose that errors U are symmetric and have constant variance (**measurement error check**). Moreover, if measurements W is replicated, the sample standard deviation of the W-values for an individual should be uncorrelated with the individual means (**constant variance check**) (96).

These topics were inspected by Figures 4.12 and 4.13 and by evaluating mean and standard deviation of the differences between replications. Figures reported results obtained by considering all the identified features of the external dataset. **Constant variance plot** of median normalized features showed a slope far from zero. In addition, Q-Q plot of the difference between replicates are not normal distributed. A standard way to remove non constant variance is via a data log transformation and so analyses were repeated after log_2 transformation. Although results did not perfectly meet the required specification, plots showed an important improvement for both issues. In addition, log_2 transformation of median normalized intensities had mean and median close to zero.

Obtained results need some specific considerations, which mainly regard log_2 data transformation. Additive error model implies that the "true" variable X is the sum of the surrogate variable W and the measurement error variable U (Equation B.1). By log-transforming data, this equation is no longer valid, unless one hypothesizes a **multiplicative error model** ($X = W \cdot U$). Under the assumption of a multiplicative error model, it is possible to derive the classical error model taking the logarithms of both sides, to get $log(W) = log(X) + log(U)$ (96). Log-normal distributions are particularly common in biomarkers measures. In fact, biomarkers values can not be negative, have generally low mean and large variances. Therefore, in this work we suppose that the **true measurement error model is multiplicative, becoming additive after variables log transformation**.

5.5 ICC estimation under measurement error and/or left censoring conditions by Monte Carlo simulations

Inspecting the measurement error structure in the collected external dataset allowed to underline that MALDI-TOF/MS features, derived from urine analysis, are affected by measurement errors. Median Normalization of features intensities was able to decrease the amount of error that was originally present in Raw data. Differently from serum or blood, urines contents largely depends on food and water intake but also they may vary for random fluctuation around a homeostatic set-point. For these reasons urinary biomarker studies should be conducted on urines collected in a 24hrs time interval, rather that in a single specimen (105). On the other hands, 24hrs urine collection not only causes a marked dilution of peptides and/or proteins released after post prostatic massage but also excreted urines stay at room temperature or at $+4\mathrm{C}^o$ for a relative long time, which may cause peptides/proteins degradation by proteolysis.

Another aspect is that MALDI-TOF/MS features, obtained from urine analysis, contained some uncertainty due to instrumental detection limit. As described above, an appropriate strategy for peak peaking can really mitigate LODs problems, but our results showed that sLOD adjustment can further ameliorate the overall features reproducibility when the optimized signal detection method was used instead of the automatic peak peaking by Flex Analysis (Table 4.3).

Up to now, measurement error or left censored data have been separately studies by statisticians. Measurement error has been deeply evaluated in error-prone predictors such as systolic blood pressure or in nutrient intake measuring, with the aims of obtain nearly **unbiased** estimated of exposure effects and valid inferences. The problems of measurements subjects to a limit of detection, usually defined as left censored measurements problems, has also been evaluated in numerous paper, and some specific strategies have been suggested to allow estimation of unbiased coefficients both in linear and in logistic regression (95, 106, 107). However, only a few studies tried to address both measurement error and limit of detection. Interestingly, Richardson and Ciampi published a paper in which they evaluated a *threshold model with error* and their findings showed that, when the standard deviation of the measurement error is lower than 0.4, the coefficient of biases is unbiased when substituting exposures below the threshold limit by LOD/2.

In this work we performed many Monte Carlo simulation analyses on normal distributed data, affected by a gradually increasing measurement error, suited to results obtained from MALDI-TOF/MS analysis of urines. We focused on ICC estimation. Starting from data measured with a defined amount of error, we derived three dataset with a different quantity of measurements set below LOD. We chose not only two LOD adjustment methods which are usable in case of non parametric distributions (the substitution of $W < LOD$ by zero and the substitution of $W < LOD$ by $LOD/2$) , but also two methods which can be applied only in parametric distribution, the **Richardson and Ciampi's method** (substitution of $W < LOD$ by $E(W \mid W < LOD)$) and the **Schisterman's method** (substitution of $W < LOD$ by $E(W \mid W > LOD)$). We didn't consider the "deletion method" (also suggested by some Authors) that is based on the exclusion of values below LOD, because this approach generates missing values, which are not biologically plausible. Different simulation situations were generated, increasing the percentage of values set below LOD. Obtained results are reported in Figure 4.14 and in Table 4.5. Results showed that **Richardson and Ciampi's method** performed better, while the worse is the **Schisterman's method**. Interestingly, **substitution of below LOD values by** $LOD/2$ worked quite well for measurement error values of $\sigma < 0.36$ and if values below LOD are less than 50 %. Finally, **Substitution of below LOD values by zero** leaded to a biased estimator, also under slightly measurement error condition. Obtained results are consistent with findings reported from many authors which analyzed LOD substitution by $LOD/2$. Cole et al. studied different LOD data handling methods, in different scenarios. They found that substituting values below LOD by $LOD/2$ gave similar results to the MLE method, which is reported to be unbiased by many Authors (106). Hewett et al. compared many LOD data handling methods and they showed that although MLE methods performed better, LOD/2 substitution gave biased results only when distributions were highly censored. Interestingly, in some scenarios LOD/2 method performances were similar to or less than that of the higher order methods (108). Therefore, ICC estimations wer e considered efficient also in presence of left censored data, especially if the percentage of values below LOD are less than 50%. After the optimized signal detection methods, the PCa database did not contain features with more that 50 % of values below sLOD.

5.6 PCa database urine evaluation by MALDI-TOF/MS

This cross-sectional study ended in 2011 with the accrual of 205 patients that referred Lower Urinary tract symptoms to consultants at Urological Unit of the University Hospital of Padova. Traditionally, LUTS are attributed to the enlarging prostate mainly for BPH, benign prostatic enlargement or benign prostatic obstruction. However, chronic prostatitis, previous inflammation or other prostate gland alterations can be present, like atypical cellular proliferation as well as other pre-malignant lesions like high-grade prostatic intraepithelial neoplasia or malignant lesions like adenocarcinoma or glandular cancer.

The collected registry had missing prostate biopsy results for eleven patients, which were recovered from revising medical records. The bio-bank collected for the study had 9% of incompleteness, while MALDI-TOF/MS analyses revealed that 5% of samples were not evaluable by this instrumentation. Although this percentage is high, samples which could not be evaluated did not belong to a single groups of patients. More interestingly, specimens which could not be evaluated contained polymers or were overdiluted urines.

Prostate biopsy results underlined that heterogeneous patients were accrued, leading to 7 stratification groups. However, because this study aimed to predict the PCa outcome, in order to discriminate high risk patients in subjects referring Lower Urinary tract symptoms, a further patients reclassification was made, originating two groups based on the presence or absence of malignant prostate cancer tissues lesions.

The MALDI-TOF/MS analyses of all the samples revealed that, in a relative small window of mass ranges, a high quantity of peptides/proteins can be detected. This result is consistent with findings previously reported from other authors with identical or similar MS instrumentation (40, 45, 48).

5.6.1 Unsupervised clustering of MALDI-TOF/MS profiling data

Hierarchical divisive clustering analysis is a method used in machine learning to inspect informations contained in features. In particular, hierarchical clustering generally allows to organize data into meaningful structures, which is depicted by branches and leafs of a tree called **dendrogram**.

Firstly, we chose to analyze all patients' groups with the purpose of evaluating whether MALDI-TOF/MS urine features could contains any prediction capabilities for PCa.

Despite four major groups were identified by cluster analysis (Figure 4.15), obtained clusters did not correspond to real patients' groups. Similarly, in the *extreme analysis*, made on considering only PCa patients and subjects with no evidences of any histological modification, cluster analysis failed to allocate patients in the correct groups and showed that six major group could be obtained using MALDI-TOF/MS features. Eventually, hierarchical clustering results, obtained by considering PCa patients only, underlined that optimal subdivision of patients by MALDI-TOF/MS features did not fit with Gleason score (data not shown).

The explanation of these results is manifold. For example, profiling data could not contain any valuable information helpful to predict patients disease status; otherwise, data may contain a large quantity of **noise**, which is overlapped to true signals. Despite our results showed that the obtained MALDI-TOF/MS features have several limitation due to biological variability, the clustering findings could point to a more concerning reality about proteomic profiling of this type of tumor, which is the **prostate cancer clinical heterogeneity**. Perhaps, what is clinically referred to as "heterogeneity" really represents inability of proteomic profiling or other clinical attributes to untangle the key elements that would, if known, help in predicting which men, referring LUTS, will be diagnosed by PCa at prostate biopsy histology. Inter-tumor high degree of heterogeneity has been already reported as possible explanation in some studies that evaluated genetic expression profiles to distinguish lethal from indolent prostate cancer (109).

However, overall results from clustering analyses suggested that poorly classification performances could probably be achieved by applying the **bioinformatics supervised machine learning algorithms** for PCa prediction.

5.6.2 Effects of Age as confounder

In statistics, a confounding factor is a variable that, in statistical models, correlates with both the dependent and the independent variables. To assess the possible role of Age as confounding we evaluated: 1) the correlation of each single MALDI-TOF/MS features with Age; 2) the correlation of Age with the patients disease status. Results showed a slightly significant correlation with the outcome of the study and a slightly correlation with ten MALDI-TOF/MS features (Table 4.7). However, in patients referring LUTS is not clear whether a strong association of PCa with Age exist.

5.6.3 Feature's outcome prediction by *naïve* logistic regression

Measurement error literature usually refers to models relating outcome (Y) *and* the surrogate variable W (and its covariate) as *naïve* models. In fact, W is the **error prone** version of the variable under examination, X. Generally, prediction of a response is different from "parameters estimation" or "inference" and so, fitting a convenient model to Y as a function of W, became merely a matter of using models for prediction. There is no need then for measurement error to play a role in the problem. However, the aim of this work is not only to predict outcome, but also to estimate the effect strength. Therefore, coefficients resulting from *naïve* logistic regression of the significantly associated MALDI-TOF/MS urine profiling features with outcome, has been reported in Table 4.8. In comparison, Age adjusted results has also been reported, although $\hat{\beta}$ didn't change significantly. The choice of reporting also Age adjusted $\hat{\beta}$ coefficients derived from the observation that Age is slightly correlated with Patient's outcome.

A review of published paper in literature shows that only two studies evaluated post-prostatic massage urines for PCa detection by using a proteomic approach. Okamoto et al. analyzed post-prostatic massage urine specimens by SELDI-TOF/MS and found 72 peaks associated with PCa, in a mass range from 2600 to 38000 Da (46). Rehman et al. evaluated voided urine after prostatic massage by 2D gels and MALDI-TOF/Ms mass fingerprinting (47). Interestingly, these two research groups reported findings that did not overlap; in addition, our results do not overlap too. These apparent discrepancies may depend on the fact that Okamoto et al. and Rehman et al. used a different analytical methods. The use of different MS instrumentation generally leads to different mass spectra, and this may explain why Okamoto et al. found mainly features of higher molecular weight. Moreover, 2D gel is a completely different sample analysis technique. However, discrepancies in findings support the concept that inter-laboratory reproducibility in proteomics strictly depends on the application of the same protocols, mass range evaluated, and on similarities in patients studied.

Obtained logistic regression coefficients need some specific considerations for their interpretation. Firstly, median normalization is a factor rescaling. Because features were rescaled with respect to their median, > 1 features values were numbers which stayed over the median value; otherwise, < 1 values stayed below the median value. Moreover, median normalized features were further log_2 transformed. Base 2 logarithm has some convenient properties. When features values is equal to their median, normalized values is equal to 1 and $log_2 = 0$. When normalized values are equal to 2, it means that values are 2 times the median and $log_2 = 1$. In addition, recall that logarithms treat numbers and their reciprocals symmetrically: $log_2(1) = 0$, $log_2(2) = 1$, $log_2(\frac{1}{2}) = -1$, $log_2(4) = 2$, $log_2(\frac{1}{4}) = -2$, and so on. We can state that a feature median value is the median peptide and/or protein content, detectable by MALDI-TOF/MS, of a specific patient. Now, suppose that after median normalization and log_2 transformation a features value is 3: the correct interpretation will be that the specific feature intensity doubled 3 times with respect to its median value. Therefore, median normalized and log_2 transformed features values simply explain how many times intensities doubled with respect to the median intensities. In logistic regression, exponentiated coefficients can be explained as the Odds ratio associated with a doubling of the features values, with respect to their median values. It can also be referred as the Odds ratio associated with a doubling of peptide/protein contents, with respect to their median content.

Finally, it is also noteworthy to consider that only ten to 482 features were significantly associated with outcome. This results imply that many peptides/proteins detectable by MALDI-TOF/MS in urines are widely shared across subjects and patients. However, it is not a surprising fact that urines, like many other biological fluids (e.g. blood), share most of their molecules, peptides and proteins across subjects. Moreover, shared components are normally expected to vary inside a range, not only for between subject variability, but also for time dependent variations, leading to the well known concepts of reference intervals and biological variability, widely used in Clinical Chemistry for reporting results and by physicians to aid during decision making. On the other hands, it would be illogical that many peptides or proteins were correlated with patients outcome, also for urine obtained after DRE. For example, many serum biomarkers used in clinical practice for cancer detection and/or follow-up are normally present in non-cancer subjects, although markedly differences are usually detectable in cancer-patients. Therefore, what we are really looking for in this study is differences in MALDI-TOF/MS features in PCa patients, with respect to Reference subjects.

5.6.4 PCa outcome prediction adjusted for measurement error

In Monte Carlo simulations we showed that, under certain conditions, when a variable is subjected to left censoring problems, substituting values below LOD by LOD/2 do not lead to biases estimation of ICC. Therefore, we used the RCAL and the SIMEX approaches to evaluate the unbiased logistic regression coefficients, relating MALDI-TOF/MS features to patients outcome. For the RCAL either features' specific ICCs or median calculated ICCs were used to adjust the *naïve* coefficients ($\hat{\beta}$). In fact, 6 to 8 features found to be significantly associated with outcome in the PCa database were not detected in the external dataset used for the measurement errors estimation. RCAL and SIMEX methods showed some discrepancy in results. In particular, RCAL seemed to overestimate the association between MALDI-TOF/MS features and outcome, being SIMEX coefficients closer to that obtained by *naïve* logistic regression. However, it is not possible to evaluate whether SIMEX estimations are biased by LOD data and, in particular, if LOD/2 substitution can adjusted estimations inferences. Therefore, based on this study results it is not possible to determine which methods, RCAL or SIMEX, gave the better results. Some studied have compared RCAL and SIMEX as possible alternatives to analyze data measured with error. Beydoun et al. evaluated the fatty acid intake from a food frequency questionnaire to predict cognitive decline in verbal fluency by logistic regression analyses. They found that, in many cases, bias in *naïve* Odds ratios was towards the null but also that RCAL tended to correct for a larger amount of effect bias than SIMEX (110). Fung et al., performed computer simulations to evaluate logistic regression results based on RCAL and SIMEX, and they concluded: *"Until better measurement error adjustment methods become available, we recommend RCAL on the basis of our simulation results"* (111). Our results showed that all the $\hat{\beta}$ coefficient obtained by *naïve* logistic regression found to be associated with outcome, are biased toward the null, not only by RCAL approach but also by SIMEX approach. As expected, measurement error represented an important component of features signals and caused a significant attenuation of the association with outcome.

Some important considerations should be given about measurement error transportability between the external dataset and the PCa dataset. As described above, in this study measurement error structure can not be assessed directly, but instead was estimated by using a dataset collected from healthy subjects, not referring LUTS and with Age not perfectly matched (external dataset). Doubts can be arisen by considering the fact that many features, found to be significantly associated with PCa, were not detected on the external dataset. These features could be specific peptides/proteins released from prostate gland during prostatic massage. Moreover, logistic regression results showed that almost all features seemed to be "protective", being their logistic regression coefficients negative. Therefore, for seven to eight features, their increased levels were associated with benign prostatic conditions. A possible explanation may be that prostate cancers may surround prostatic duct, preventing or diminishing release of normal prostatic juice or tumor can directly obstruct prostatic duct. However, the features 1404.7 was positively associated with PCa, demonstrating that some potential cancer-specific marker can be found in urine after prostatic massage.

5.7 Analytical goals and maximum allowable analytical variation

In Clinical Chemistry, analytical variability should not be more that 0.75 times the with-in subject biological variability, to reach the minimal desirable performances, or minimal analytical goals. This allows that changes in biomarkers levels will be detected with enough analytical precision to find real variations of patients conditions. In this study, analytical variabilities were estimated by the inter-run assay, while with-in subjects variabilities were evaluated by the external dataset. By using the median values of the analytical variabilities and median values of the with-in subject variabilities, we calculated the **overall analytical performances** of MALDI-TOF/MS in urine analysis. For the evaluation we used the the median normalized, sLOD adjusted features intensities. Obtained results showed that, overall, MALDI-TOF/MS urine analysis met the minimal desirable analytical performances ($\sigma_w/\sigma_I = 1.13$). That's allows the usage of MALDI-TOF/MS urine features for monitoring patients during time.

6

CONCLUSIONS

The current study aimed to identify new PCa candidate biomarkers by mass spectrometry analysis of urines collected after prostatic massage. An high throughput mass spectrometry instrumentation, namely MALDI-TOF, was used to generate profiles of low molecular weight peptides/proteins contained in urines, and the identified features were used for predicting the patient's outcome. We started by focusing on some important aspects of MALDI-TOF/MS proteomic profiling workflow and statistical data analysis, in order to assess reproducibility of the obtained measurements and the validity of the resulting models, relating features and patients outcome. Proteomics, like other "-omics" sciences, generates massive quantities of data and requires appropriate data-mining strategies to discovery the most meaningful parts of these data. This process, which is like *"to look for a needle in a haystack"*, often need that biology, biochemistry, statistics and bioinformatics share knowledge and responsibilities.

Findings from the **reproducibility study** showed that the major contributing factor to MALDI-TOF/MS profiling variability is the peak finding process, which "transform" linear vectors of thousand of points in a limited number of variables, containing peaks signals. In particular, after Automatic Peak finding, many features heavily suffered of left censoring problems. Many values of these variables, were erroneously set to zero, irrespectively of their "true" signals. In this case, adjusting data for instrumental signals detection limits, a convenient surrogate of measures detection limits, highly decrease the features overall variability. However, on considering the optimized signals detection peak detection, the recalculated Raw CVs were around 40 %, while median normalization after sLOD data adjustment allowed a further CVs decrease up to 20 %. Moreover, median normalization, together with sLOD adjustment of data, ameliorated analyses, increasing across replicates comparability.

Evaluating an **external dataset**, based on urines repeatedly collected in reference subjects, we further confirmed that most of the measurable MALDI-TOF/MS features variability were due to the biological matrix (urine). The median of the MALDI-TOF/MS features ICC, which encompass both the between and the within subject variabilities, was far from good, but similar to urinary creatine ICC. Inspecting the error structure of the external dataset, we demonstrated that features measurement error is decoupled from their mass values, showing that the identified error structure had not a particular behaviour along mass detection range, by using the reflectron mode for the MALDI-TOF/MS analyses. Under pre-specified conditions and if left censored data were appropriately handled, Monte Carlo simulations showed that estimated ICCs were only slightly affected by data measured below the detection limit. In particular, the our findings showed that substitution of $x < LOD$ by $E(x|x < LOD)$ gave the better results, even so substitution of $x < LOD$ by $LOD/2$ allowed to obtain unbiased estimates when the measurement error and the percentage of values falling below LOD were sufficiently low.

Unsupervised clustering of MALDI-TOF/MS features, obtained by analyses of urines collected after DRE, suggested that data might contain too much noise to efficiently cluster patients accordingly to their prostate biopsy histological reports. Comparing results from **logistic regression *naïve* analyses, RCAL and SIMEX** showed that measurement error, overall, caused a bias "toward the null". However, SIMEX estimations seemed to correct for a smaller amount of bias than RCAL. From these analyses, eight features were found to be associated with PCa. In particular, the feature 1404.7 seemed to be an interesting PCa associated biomarker, generally detectable in urine collected after prostatic massage of patients referring LUTS, but found to be increased in PCa patients.

Finally, this is the first study which compare the MALDI-TOF/MS analytical performances with respect to the urines biological variations, showing that this instrumentation may achieve the Clinical Chemistry desirable analytical performances and that MALDI-TOF/MS can be used in monitoring urines of patients over time.

Appendix A

Left-censored data

Schisterman and Little in 2010 published a paper that can be considered a milestone for biomarker when measurements are subjected to LOD (112). When measured levels are low, biomarkers may be subjected to inadequate instrument sensitivity, resulting in a large percentage of measurements falling below the (experimental determined) LOD. This may occur, for example, in quantitation of immunoassays, that require antigen concentration sufficient for binding by antibodies, but also in mass spectrometry for both instrumental and peak detection sensitivity. Alternatively, assays may detect low biomarker levels but suffer from insufficient specificity, and measurement of exposure is hampered by background. So, numerical data are observable above and below the LOD even if, among values below the threshold, it may not be possible to clearly delineate between those that are “real” and those that are not. For example, this latter example is applicable to the proposed feed-back peak detection strategy. Intuitively, it is obvious to think that many times, biomarkers measurements can suffer of detection limit problem but, most of the times, these below LOD measurements are generally not handled. Thus, the impact on risk assessment can be strong, suggesting the need to address this important issue. From a statistical point of view, data with detection limit is also called **left censored data**. However, statistical modeling requires decisions regarding their handling (95). Firstly, a clear specification of below LOD values should be taken. For example, suppose to evaluate an exposure variable X, in which a lower threshold interferes with measurement of low exposure levels. In this case what we are really observing is the variable Z, which equals either X or "nondetects" (ND), according to the following:

$$\begin{aligned} &\text{for all } x > LOD, \quad z = x \\ &\text{for all } x \leq LOD, \quad z = \text{ND} \end{aligned} \tag{A.1}$$

In the alternative model, when the variable X is less than LOD, there is a quantitative random noise, ξ, rather than the qualitative response:

$$\begin{aligned} &\text{for all } x > LOD, \quad z = x \\ &\text{for all } x \leq LOD, \quad z = \xi \end{aligned} \tag{A.2}$$

However, when ξ is reasonable lower that LOD, equation A.2 can be a special case of equation A.1 (95). To understand this problem, we can formally suppose that $Pr(\xi > LOD) > 0$. This may be the case if detection limit is set as two, rather than three, standard deviation. Under this scenario, it is not possible to lead back equation A.2 to equation A.1.

When Equation A.1 is valid, a new equation can be defined as following:

$$y(x) = \begin{cases} I(x) & I(x) > LOD \\ a & I(x) \leq LOD \end{cases} \tag{A.3}$$

where a is a value, ranging between 0 and LOD itself, substituting intensities below LOD and $I(x)$ is the measured intensity. Up to now, many studies have developed or compared statistical methods for analyzing left censored data (106, 107, 108, 112, 113, 114, 115, 116, 117, 118, 119, 120). All these methods can be either based or not based on distributional assumption of X. Published or recommended methods for analyzing such datasets tend to fall into six categories or families (108):

- Substitution methods,
- Deletion method (omission)
- log probit regression (LPR) methods,
- maximum likelihood estimation (MLE) methods,
- non parametric (NP) methods,
- multiple imputation methods.

Most of these methods aims to estimate a in equation A.3 based on data that are above the detection limit.

A.1 Substitution methods

The three common substitution values for a (EquationA.3): LOD, $LOD/2$ and $LOD/\sqrt(2)$. Although the choice of the substitution fraction is largely arbitrary, $LOD/\sqrt(2)$ is usually recommended when data have been *log* transformed before LOD adjustment and researchers are estimating the geometric mean. All of the substitution methods are biased and, as the sample size increases, the bias asymptotically approaches a fixed value. El-Shaarawi and Esterby shown that the bias to be expected can be calculated when ranges of the mean, variance and proportion censored are available (121). However, their formulae have the limitation that cannot be used to determine the bias for small sample sizes.

A.2 Deletion method

It is based simply on discarding observations with $X_i \leq LOD_i$, as defined by Nie et al. (107).

A.3 Log probit regression methods

Log probit regression (LPR) method (also called regression on order statistics), the data are sorted and a linear relationship is assumed between the logarithm of occurrence values and the inverse cumulative normal distribution of the observations plotted position. This is a linear equation which is solved for each non-detect observation (122). A variation on this method is the robust LPR (LPR$_r$) method, which is described to be less susceptible to departures from the lognormal assumption.

A.4 Maximum likelihood estimation

Many authors consider MLE as the best approach from a methodological perspective for dealing with LOD data, in particular in small sample size scenarios (107, 108). The sample parameters are those estimates that maximize the likelihood function after the definition of a parametric distribution to best fit the data. If the underlying distribution is felt to depart significantly from the lognormal distribution assumption, it is recommended to use robust maximum likelihood estimation (MLE$_r$) (108).

A.5 Non parametric methods

Non-parametric methods are so named because they do not involve computing **parameters** such as the mean or standard deviation, of a given distribution. Instead they use the relative magnitude (ranks) of data. The standard non-parametric technique for censored data is the Kaplan-Meier (KM) method. This method is based on the cumulative distribution function. It was originally developed for estimating the mean of (right-) censored survival data (e.g. in medical research). A plot of the Kaplan-Meier which is estimated from the survival function, is a series of horizontal steps of declining magnitude which, when a large enough sample is taken, approaches the true survival function for that population. An important point in KM method is that the value of the survival function, between successive distinct sampled observations, is assumed to be constant. Its main advantage is the ability to estimate the mean in the presence of non-detects, without relying upon a distributional assumption.

A.6 Recomendations in literature

Most of the method for left censored data analysis are based on simulation studies. Hewett et al. compared substitution methods, LPR, MLE and NP methods by using three different simulations. In simulation 1 sample size ranged between 20 and 100 and the true percent censored ranged between 1% and 50%; in simulation 2 sample size ranged between 20 and 100, the true percent censored ranged between 50% and 80%; in simulation 3 sample size ranged between 5 and 19, the true percent censored ranged between 1% and 50%. They also compare a "pure" *log*-normal *vs* contamined *log*-normal distributions. They found that $LOD/2$ ad $LOD/\sqrt(2)$ substitutions methods performed quite well both in mean and in 95% CI of the mean when the percent censoring was $<$ 50%, albeit they demonstrating that LOD substitution was the worst of the compared methods (108). In fact, substitution methods performed worst with respect to the MLE- and LPR-based method. Moreover, LPR-based and KM methods tended to be in the middle to top half of the bias while MLE-based methods performed better in the multiple LOD and contaminated distribution scenarios.

Richardson's and Ciampi's paper was the first attempt to characterize the effect of random measurement error when there is a lower threshold for recorded values (118). They evaluated the direction and magnitude of bias in estimating exposure-response associations by logistic regression, under the assumption that the true exposure followed a *log*-normal or a gamma distribution. They calculated the bias coefficient (λ) for the following cases: 1) a threshold model (without measurement error) with x distributed according to the lognormal (0,1) distribution; 2) a threshold model (without measurement error) with x distributed according to the gamma (1,1) distribution; 3) a threshold model with measurement error with x distributed according to the lognormal (0,1) distribution; and 4) a threshold model with measurement error with x distributed according to the gamma (1,1) distribution. They firstly evaluated the situation without measurement error and shown that the estimated coefficient λ can be either less than 1 or greater than 1, depending on the distribution of x. In particular λ depends on which values are chosen to substitute values below LOD. The coefficient, λ, will be equal to 1 when values below LOD is substituted with the expected value of x conditional on x being below the threshold ($E(x \mid x \leq LOD)$). If the value assigned to below-LOD measurements is less than the expected value of x conditional on x being below the LOD, attenuation will occur in estimates of association; in contrast, if below threshold measurement are substituted with a larger value, inflation will occur in estimates of association. Secondly, they evaluate how λ varied in cases of increasing measurement error and by substituting x values below LOD with zero, with $LOD/2$ and with LOD. Interestingly they found that in log-normal distributions, $LOD/2$ substitution leaded to a minimal bias, especially when measurement error standard deviation ranged between 0 and 0.6. Differently, in gamma distributions λ are higher.

Nie et al. compared the substitution, deletion, MLE and a modified MLE based methods with x values below LOD substituted by the value of x conditional on x being below LOD ($E[x \mid x \leq LOD]$), evaluated by Richardson and Ciampi (RC) (118). They used using normally and not normally distributed distributions and showed that replacing the censored $X_i \leq LOD$ with LOD, $LOD/2$ ad $LOD/\sqrt{2}$ provide biased estimates, while deletion methods, which overestimates the standard error of the estimated coefficient, performed better that substitution method. However, in cases in which the variable distribution cannot be evaluated, methods that rely on distributional assumptions may yield highly biased estimates. Differently, when the parametric normal distribution assumption for the left-censored variable is correct, the MLE and the RC methods provide consistent estimates although the RC method gives slightly underestimated standard errors. Jain et al. studied the performance of MLE methods with or without multiple imputation when the percentage of observations below LOD are greater than 50 %. They suggested that MLE without imputations may be preferred in instances when N is small and the percentage of undetected is very large, and MLE with imputations may be preferred when N is relatively large and the percentage of undetected is relatively small.

Schisterman et al. evaluated the detection limit problems focusing on distribution-free methods for managing values below the LOD. They evaluated biases in linear regression coefficient β that result when exposure measurements are constrained by a lower threshold (95). In particular they evaluated the omission (namely deletion), imputation with a constant (e.g., $LOD/2$, $LOD/\sqrt{2}$), replacement of below-threshold data by the expectation for such data $[E(x \mid x < LOD)]$ and replacement of below-threshold data by the expectation for data above LOD $[E(x \mid x > LOD)]$ methods. As described above, using $E(x \mid x < LOD)$ for those data below LOD allows unbiased estimation of linear and, logistic regression parameters (108, 118); however, this approach requires assumptions regarding the underlying exposure distribution. In Schisterman study Authors shown that unbiased estimates for linear regression may also be obtained if data below the detection limit are replaced by $E(x \mid x > LOD)$ (95).

Appendix B

Measurement error

Measurement error has long been a concern in relating error-prone predictors such as systolic blood pressure (SBP) to the development of coronary heart disease. That SBP is measured with error is well known because it is has strong daily and seasonal variations (namely biological variations). However, in trying to measure SBP, the various sources of error also include simple machine recording error or machine inaccuracy (analytical errors) in SBP determination. Although the reasons for imprecise and inaccurate measurement are diverse, the inference problems they create, share in common the structure that statistical models must be fit to data formulated in terms of well-defined but unobservable variables X, using information on measurements W. Problems of this nature are called measurement error problems and the statistical models and methods for analyzing such data are called measurement error models (123). Non linear error-in-variables modeling began in earnest in the early 1980s, with the publication of a series of paper on diverse topics, from which the forefather is the work of Prentice on survival analysis(124). Following, many others paper has been published until now in this topic. However, before to analyze other measurement error aspects, a briefly etymological revision of some common definitions, like **biological variability**, errors in the **analytical phase** and **pre-analytical variations** is a pivotal aspects.

Biological variations, which can be defined as the variability of a laboratory quantity due to physiologic differences in the same subject over time, has actually many aspects to consider. For example, it is intuitive to think, e.g., that some quantities measurable in blood (like hormones and biomarkers), vary over periods of time in lifespan, when significant biological changes in individual occur (e.g., the neonatal period, puberty, the menopause, and old age). In addition, for some of the measurable quantities in blood, a predictable daily of weekly cyclical biological variation have been shown (a well-documented examples include serum cortisol). However, the variation of almost all quantities, including biomarkers, can be described as random variation over time, which is commonly termed “biological” variation (125).

The analytical phase can be defined as the instrumental sample analysis. The errors type that are important in this regard are imprecision and bias. Imprecision is random error and is defined as the closeness of agreement between independent results of measurements obtained under stipulated conditions. Bias is systematic error and is defined as the difference between the expectation of measurement results and the true value. Constant bias obviously does not affect serial results obtained on samples from an individual over time, although all of the values may be lower or higher than the true value. However, changes in bias may be a significant source of variation in serial results (125).

Pre-analytical variation occurs before the analytical phase of generation of an observed value. The sources of variation can be divided into two types, namely, factors that affect the individual before specimen collection occurs and factors inherent in the collection and handling of the specimens. The former are very important biological sources of true variation. The latter include important influences such as time of tourniquet application before specimen collection, temperature at which specimens are transported to the laboratory, anticoagulant used, time elapsed between specimen collection and examination, and time and speed of centrifugation (125). Some other aspects of the pre-analytical variation are described in detail above.

B.1 Error Models

The number ways a surrogate W and a predictor X can be related are countless. However, in practice it is often possible to reduce most problems to one of two simple error structures. For understanding the effects of measurement error and the statistical methods for analyzing data measured with error an understanding of the two simple error structures is generally sufficient. Models can be divided not only in Classical and Berkson models but also in functional and structural models. Functional models make no assumption on X, beyond what are made in absence of measurement error. Methods based on functional modeling can be divided into **approximately consistent** (remove most bias) and **fully consistent** (remove all bias) (123). In the measurement error theory the key point is that, as W is measured with error, regressing the outcome Y on W, the estimation coefficients will be biased because we use an inaccurate value of W to replace X (126).

B.1.1 Classical error model

As said above, in trying to measure SBP, there are many source of error, including time of day and season of the year. In such circumstance, it makes sense to hypothesize an **unbiased additive error model**, known also as **classical error model**. The classical error model in case W a surrogate measurement of X can be defined as following

$$W = X + U \tag{B.1}$$

where U has mean zero and is independent of X (127). Thus, the variability of the observed measurement will be greater than of the true dose ($\sigma^2_W = \sigma^2_X/\sigma^2_{WX}$). This model supposed that $var(U \mid X) = \sigma^2$ and is constant and that $U \mid X \sim Normal(0, \sigma^2_u)$. Generally speaking, applying a classical error models leads to modelling the conditional distribution of W given (Z and X), where Z are covariates. To check if the classical error model hold when multiple measurement are collected for each individual, it should be verified that the standard deviation of the W_{values} for an individual are uncorrelated with the individual means, and they are also uncorrelated with the covariates Z. For example, on considering SBP, if a "true" value of a long-term SBP is X, then the measured value at any given time (W) will differ from X because short-term SBP at that time differs from X, and because there is analytical and pre-analytical variability on the SBP measurement (e.g., a small sample of blood is taken or that sample is analyzed without a correct instrumental calibration). However, as underlined above, **the analytical variation should be affected by random error but not from systematic bias**.

B.1.2 Berkson error model

The Berkson error model is also known as **controlled variable model**. A comprehensive example can be the herbicide example (127). If a nominal measured amount W of herbicide is applied to a plant, the actual amount X observed by the plant will be different from W, because of potential error in application or plant absorption. In this case the error model can be specified as:

$$X = W + U \tag{B.2}$$

where U, the error, is assumed to be independent of W (127). In this case, the true measurement has more variability of the observed one. Berkson error model, including the **regression calibration models** is equal to model the conditional distribution of X given (Z, W).

B.1.3 Additive or Multiplicative error models

Classical error model describe an additive error model. However, for some data, especially for biomarker measurements, a multiplicative error model can be appropriate. If a log-normal distribution can be assumed for the distribution of W and U, it is possible to assume that also X is log-normal distributed and so:

$$W = X \cdot U, \quad log(U) \sim Normal\{0, \sigma_u^2\} \tag{B.3}$$

Under the assumption of a multiplicative error model, it is possible to derive the classical error model taking the logarithms of both sides, to get $log(W) = log(X) + log(U)$ (96).

B.2 Differential and Nondifferential error

Distinguish between differential and nondifferential measurement error is very important. A measurement error is said to be no differential if measured W contains no other information that what is available in X and covariates, with respect to the outcome. Therefore, Non-differential measurement error occurs in a broad sense, when one would not even bother with W is X were available. In this case W is said to be a **surrogate variable** of X. Non-differential measurement error typically hold in cohort studies, but is often a suspect assumption in case-control studies (127). Measurement error is differential otherwise. If measurement error is due solely to instrument or laboratory-analysis error, then it can often be argued that the error is non-differential.

B.3 Loss of power in measurement error

The increase in residual variance associated with surrogate measurements (including classical and Berkson) gives rise not only to a decrease in predictive power, but also contributes to reduced power for testing. Carroll et al. reported an example. If n = 20 is sufficient to obtain a 90 % power in absence of measurement error, when 1/20 of the variability in the observed predictor W is due to noise, a sample size of 45 is required (96).

B.4 Bias caused by Measurement error

It is commonly thought that the effect of measurement error is to bias estimates "toward the null". As Carroll said, "this lovely and appealing folklore is sometimes true but, unfortunately, often wrong" (127). However, in linear regression, the effect of measurement error is generally defined as *attenuation*. In particular, the effect of measurement error depends on the model under consideration and on the joint distribution of the measurement error and the other variables. Therefore, the effect can range from the simple attenuation to situation where real effects are hidden and even the sign of the estimated coefficients are reversed (96). In nonlinear regressions, measurement error effects are much the same qualitatively as in the normal linear model (123). Mainly, the use of a surrogate marker produce parameter bias, and Berkson measurements often produce much less severe than biases resulting from classical models. This fact forms the basis for the method know as **regression calibration** in which, an unbiased Berkson predictor is estimated by a preliminary calibration analysis, and then the usual (naïve) analysis is performed with $E(X \mid W)$ replacing X.

B.5 Source of data

In order to perform a measurement error analysis, one needs information about either W given (X, Z) (classical measurement error) or about X, given (Z, W) (regression calibration). In other words, one needs information about the error structure. The data source can be divided into two main categories:

1. *Internal* subset of the primary data,
2. *External* or independent studies.

Within each of these broad categories, there are three subtypes of data, validation, replication and instrumental data.

In *validation* data, X observed directly. Therefore validation studies are rare, especially if X is a long term exposure. They allow to understand whether the classical error model actually holds. Validation studies are similar to missing data problem.

In *replication* data, replicates of W are available. Replication data allow easy to estimate of σ_u^2 through the analysis of variance (ANOVA) calculation. For example, one would make replicate measurements of data if there were good reasons to believe that the replicated mean is a better estimate of X than a singe observation.

In the *instrumental variables*, one or mode additional variables T are measured. T has the following properties: a) are correlated with the true exposure X, b) are nondifferential, c) are independent of the measurement error U. Of course an internal validation data set would be ideal, because: 1) this allow the direct examination of the error structure, 2) this leads to much greater precision of estimation and inference. With external validation data, one must assume that the error structure also applies to the primary data (**transportability**). In many studies, the measurement error structure can't be assessed directly and researchers should collect an external data set. If a similar measurement protocol and a similar levels of training for technicians making the measurements can be assumed, it can be also reasonably assumed that distribution of the error in the recorded measures is independent of all these aspects. Thus, it seems reasonably to assume transportability of the error structure, namely that error distribution is the same across different population (127).

B.6 The exact predictor

In the measurement error theory, the definition of "exact" need to be carefully revised. In fact, an exact predictor, measured without error, can't exist. However, in measurement error literature the tem "**gold standard**" is often used for the operationally defined exact predictor. However, the best definition of gold standard should be "the best one could ever possibly hope to accomplish" (127).

B.7 Regression calibration model

Regression calibration is a conceptually straightforward approach to bias reduction and has been successfully applied to a broad range of regression models. Regression calibration is fully consistent in linear models and approximately consistent in nonlinear models. When the measurement error is non-differential, the regression model relating Y to the observed variable W and covariate Z is:

$$E[Y \mid Z, W] = E[E[Y \mid Z, X] \mid Z, W] \tag{B.4}$$

which is called calibration function. Therefore, the regression calibration acts substituting in the disease model X with $E[Y \mid Z, W]$ (eq. B.4). In this case, X "means" as the best estimate of X using the observed predictors (Z, W) in terms of minimizing mean square prediction error. Spiegelman et al. show that the best estimate of X can be an alloyed gold standard observation of that variable (128), while models can be adapted also when a gold standard doesn't exist. In particular, in replicated data, it is reasonable to consider the average of a large number of measurements as a gold standard, therefore as the "true" measure (X), that can be compared with the single measurement (W)(94). As stated above, although X is not directly measurable, it is possible to use the ANOVA with random effects to estimate the regression coefficient relating W (single measurement for the ith subject) to X (underlying mean of measurement for the ith subject).

$$\begin{aligned} &W_i = X_i + e_i \\ &X_i \sim N(\mu, \sigma_A^2), e_i \sim N(0, \sigma^2) \end{aligned}$$

In the previous equation σ_A^2 represents between-person variation, and σ^2 represent within-person variation.
After, we can calculate the following standard deviations, $sd(X)$ and $sd(W)$:

$$\begin{aligned} &sd(X) = \sigma_A \\ &sd(W) = \left(\sigma_{_A}^2 + \sigma^2\right)^{1/2} \end{aligned}$$

To implement the regression calibration method, we need to estimate γ that is the regression coefficient of x on X:

$$X = \alpha + \gamma W + e$$

Generally speaking, the relationship between a regression coefficient and a correlation coefficient is:

$$b(X\mathrm{on}W) = Corr(X, W)sd(X)/sd(W)$$

and:

$$\begin{aligned} &\rho_I = \sigma_A^2/(\sigma_A^2 + \sigma^2) = \mathrm{intraclasscorrelationcoefficient} = r_i \\ &Corr(X, W) = \mathrm{reliabilitycoefficient} = (\rho_{\mathrm{I}})^{1/2} \end{aligned}$$

Therefore, on combining Equations:

$$\begin{aligned} b(X\mathrm{on}W) &= (\rho_I)^{1/2}\sigma_A/(\sigma_{_A}^2 + \sigma^2)^{1/2} \\ &= (\rho_I)^{1/2}\left[\sigma_A/(\sigma_{_A}^2 + \sigma^2)\right]^{1/2} \\ &= (\rho_I)^{1/2}(\rho_I)^{1/2} = \rho_I \end{aligned}$$

Thus, it is possible to estimate the regression coefficient of X on W by the **intraclass correlation coefficient**, r_i (94). However, to obtain r_i we need to conduct a reproducibility study on a subsample of subject with at least two replicates per subject. The reproducibility study plays the same role as a validity study, which is used when a gold standard is available. It is also important to underline that in Case-control studies, regression calibration need a slight modification. If fact, it is important that regression calibration function will be estimated by using the controls only.

B.8 Simulation and extrapolation method

Simulation and extrapolation (SIMEX) is a simulation based method of estimating and reducing bias due to measurement error. SIMEX estimates are obtained by adding additional measurement error to the data in a resampling-like stage, establishing a trend of measurement error-induced bias versus the variance of the added measurement error, and extrapolating this trend back to the case of no measurement error. To describe the algorithm, we can suppose that we know the U variance (σ_u^2) and that, in addition to the original data used to calculate the naïve estimate $\hat{\beta}_{x,naive}$ there are M-1 additional data sets available each with successively larger measurement error variance $(1+\zeta_m)\sigma_u^2$, where $0 = \zeta_1 < \zeta_2 < ... < \zeta_M$. Therefore, the least square estimate of slope from the m_{th} data set, ignoring measurement error, $\hat{\beta}_{x,m}$, consistely estimates $\beta_x\sigma_x^2/[\sigma_x^2+(1+\zeta_m)\sigma_u^2]$. So it is possible to think of this problem as a nonlinear regression model, with dependent variable $\hat{\beta}_{x,m}$ and independent variable ζ_m. The technique was proposed by Cook and Stefansky, in 1994 (129). An integral component of SIMEX is a self-contained simulation study resulting in graphical displays that illustrate the effect of measurement error on parameter estimates and the need for bias correction. The **graphical display**, which is an integral component of SIMEX, is also useful when it is necessary to motivate or explain a measurement error model analysis.

Appendix C

Simulation studies

Simulation is a versatile problem-solving methodology that involves the abstraction of a real-life system into a symbolic model format and provides an alternative to a purely mathematical analytical solution. A simulation model must be developed in an accurate way and it must be also validated before considering its results. Monte Carlo methods and Bootstrapping methods are two examples of simulations.

C.1 Monte Carlo methods

The term “Monte Carlo” come from the famous casinos in Monte Carlo (MC). When study of MC methods became in earnest in the 1940’s and 1950’s, someone thought of the likeness of using random numbers to the randomness in the games at the casino, and named the method thereafter. The expression "Monte Carlo" is actually very general because the application fields is vast like Math, economics, physics. Differently, Monte Carlo simulation (which can be used interchangeably with the term **simulation**) is broadly used by statisticians to refer to computer experiments involving random sampling from probability distribution, for obtaining numerical solutions to problems which are too complicated to solve analytically. Therefore, instead of calculating exact quantities, simulation is used to produce stochastic approximation to the solution. As trivial example of a Monte Carlo simulation can be considered the calculation of the mean μ of a density $g(x)$. The analytical solution is:

$$\int xg(x)\,dx,$$

which may be difficult to evaluate. In the Monte Carlo approach, it is possible to sample k observations, $X_1, ..., X_k$, from $g(x)$ and form the Monte Carlo estimate:

$$\hat{\mu} = \frac{\sum_{i=1}^{k} X_i}{k}$$

The **law of large numbers** can be used to show that this estimate will converge strongly to the true value, provided the integral exists (130).
In a generic form, a typical MC simulation involves the following steps:

- Generate S independent data sets under the condition of interest
- Compute the numerical value of the estimator (θ) or test statistics for each data set (θ_1, θ_2, ... , θ_S)
- if S is large enough, summary statistics across θ_1, θ_2, ... , θ_S should be good approximations to the true sampling properties of the estimator/statistic under conditions of interest.

Therefore, by MC simulation it is possible to obtain the "true" value of the estimator of interest, without using real data in the statistical analysis.

C.2 Bootstrapping statistics

Bootstrapping statistics refers to a method for assigning measures of accuracy to sample estimates, and can be applied at almost any statistics. To explain the bootstrap technique, we can suppose to estimate a parameter θ from a school, made of 15 students. As Diaconis and Efron reported, suppose now that we had available data from other schools, each data made of 15 subjects. If we now calculate the parameter θ for each set of 15 students, we will collect more information on θ and increase the θ accuracy (131). Unfortunately, we don't really have there hypothetical other sets of students. The bootstrap algorithm overcomes this difficulty by constructing a sequence of "fake data sets" using only the data from the original 15 law schools. Interestingly, bootstrapping is a non-parametric methods and therefore its results is valid also with no prior assumptions about the distributions of the sampling population (131). This is the reason because the bootstrap technique is also called **non-parametric bootstrap technique**. We can define the non-parametric bootstrap by using the Efron's formulation: non-parametric bootstrap treats the original data set as a complete population and draws a new, simulated sample from it, picking each observation with equal probability (allowing repeated values) and then re-running the estimation. In fact, this is usually what people mean when they talk about "the bootstrap" without any modifier (93, 131).
An important variant of the non-parametric bootstrap is the **smoothed bootstrap**, where we re-sample the data points and then perturb each by a small amount of noise, generally Gaussian.

Bibliography

[1] J Ferlay, E Steliarova-Foucher, J Lortet-Tieulent, S Rosso, J W W Coebergh, H Comber, D Forman, and F Bray. Cancer incidence and mortality patterns in europe: estimates for 40 countries in 2012. *Eur J Cancer*, 49(6):1374–403, Apr 2013.

[2] Rebecca Siegel, Deepa Naishadham, and Ahmedin Jemal. Cancer statistics, 2013. *CA Cancer J Clin*, 63(1):11–30, Jan 2013.

[3] AIRTUM Working Group. [italian cancer figures, report 2010: Cancer prevalence in italy. patients living with cancer, long-term survivors and cured patients]. *Epidemiol Prev*, 34(5-6 Suppl 2):1–188, 2010.

[4] Daphne Hessels and Jack A Schalken. Urinary biomarkers for prostate cancer: a review. *Asian J Androl*, 15(3):333–9, May 2013.

[5] William G Nelson, Angelo M De Marzo, and William B Isaacs. Prostate cancer. *N Engl J Med*, 349(4):366–81, Jul 2003.

[6] H Grönberg. Prostate cancer epidemiology. *The Lancet*, 361(9360):859–64, Mar 2003.

[7] Irina Mordukhovich, Paul L Reiter, Danielle M Backes, Leila Family, Lauren E McCullough, Katie M O'Brien, Hilda Razzaghi, and Andrew F Olshan. A review of african american-white differences in risk factors for cancer: prostate cancer. *Cancer Causes Control*, 22(3):341–57, Mar 2011.

[8] Kathryn M Wilson, Edward L Giovannucci, and Lorelei A Mucci. Lifestyle and dietary factors in the prevention of lethal prostate cancer. *Asian J Androl*, 14(3):365–74, May 2012.

[9] Matteo Rota, Lorenza Scotti, Federica Turati, Irene Tramacere, Farhad Islami, Rino Bellocco, Eva Negri, Giovanni Corrao, Paolo Boffetta, Carlo La Vecchia, and Vincenzo Bagnardi. Alcohol consumption and prostate cancer risk: a meta-analysis of the dose-risk relation. *Eur J Cancer Prev*, 21(4):350–9, Jul 2012.

[10] Richard M Bambury and David J Gallagher. Prostate cancer: germline prediction for a commonly variable malignancy. *BJU Int*, 110(11 Pt C):E809–18, Dec 2012.

[11] C Hughes, A Murphy, C Martin, O Sheils, and J O'Leary. Molecular pathology of prostate cancer. *J Clin Pathol*, 58(7):673–84, Jul 2005.

[12] Ola Bratt. Hereditary prostate cancer: clinical aspects. *J Urol*, 168(3):906–13, Sep 2002.

[13] Mark L Gonzalgo and William B Isaacs. Molecular pathways to prostate cancer. *J Urol*, 170(6 Pt 1):2444–52, Dec 2003.

[14] Jielin Sun, A Karim Kader, Fang-Chi Hsu, Seong-Tae Kim, Yi Zhu, Aubrey R Turner, Tao Jin, Zheng Zhang, Jan Adolfsson, Fredrik Wiklund, S Lilly Zheng, William B Isaacs, Henrik Grönberg, and Jianfeng Xu. Inherited genetic markers discovered to date are able to identify a significant number of men at considerably elevated risk for prostate cancer. *Prostate*, 71(4):421–30, Mar 2011.

[15] Atish D Choudhury, Rosalind Eeles, Stephen J Freedland, William B Isaacs, Mark M Pomerantz, Jack A Schalken, Teuvo L J Tammela, and Tapio Visakorpi. The role of genetic markers in the management of prostate cancer. *Eur Urol*, 62(4):577–87, Oct 2012.

[16] Lucinda Hughes, Fang Zhu, Eric Ross, Laura Gross, Robert G Uzzo, David Y T Chen, Rosalia Viterbo, Timothy R Rebbeck, and Veda N Giri. Assessing the clinical role of genetic markers of early-onset prostate cancer among high-risk men enrolled in prostate cancer early detection. *Cancer Epidemiol Biomarkers Prev*, 21(1):53–60, Jan 2012.

[17] Leander Van Neste, James G Herman, Gaëtan Otto, Joseph W Bigley, Jonathan I Epstein, and Wim Van Criekinge. The epigenetic promise for prostate cancer diagnosis. *Prostate*, 72(11):1248–61, Aug 2012.

[18] Shannon R Payne, Jurgen Serth, Martin Schostak, Jorn Kamradt, Arne Strauss, Paul Thelen, Fabian Model, J Kevin Day, Volker Liebenberg, Andrew Morotti, Su Yamamura, Joe Lograsso, Andrew Sledziewski, and Axel Semjonow. Dna methylation biomarkers of prostate cancer: confirmation of candidates and evidence urine is the most sensitive body fluid for non-invasive detection. *Prostate*, 69(12):1257–69, Sep 2009.

[19] Mark A Reynolds. Molecular alterations in prostate cancer. *Cancer Letters*, 271(1):13–24, November 2008.

[20] Nathalie Heuzé-Vourc'h, Valérie Leblond, and Yves Courty. Complex alternative splicing of the hklk3 gene coding for the tumor marker psa (prostate-specific-antigen). *Eur J Biochem*, 270(4):706–14, Feb 2003.

[21] Axel Heidenreich, Joaquim Bellmunt, Michel Bolla, Steven Joniau, Malcolm Mason, Vsevolod Matveev, Nicolas Mottet, Hans-Peter Schmid, Theo van der Kwast, Thomas Wiegel, Filliberto Zattoni, and European Association of Urology. Eau

guidelines on prostate cancer. part 1: screening, diagnosis, and treatment of clinically localised disease. *Eur Urol*, 59(1):61–71, Jan 2011.

[22] Andrew M D Wolf, Richard C Wender, Ruth B Etzioni, Ian M Thompson, Anthony V D'Amico, Robert J Volk, Durado D Brooks, Chiranjeev Dash, Idris Guessous, Kimberly Andrews, Carol DeSantis, Robert A Smith, and American Cancer Society Prostate Cancer Advisory Committee. American cancer society guideline for the early detection of prostate cancer: update 2010. *CA Cancer J Clin*, 60(2):70–98, 2010.

[23] Richard Mayeux. Biomarkers: potential uses and limitations. *NeuroRx*, 1(2):182–8, Apr 2004.

[24] Peter H Gann. Risk factors for prostate cancer. *Rev Urol*, 4 Suppl 5:S3–S10, 2002.

[25] Walter Artibani. Landmarks in prostate cancer diagnosis: the biomarkers. *BJU Int*, 110 Suppl 1:8–13, Oct 2012.

[26] Carlo-Federico Zambon, Tommaso Prayer-Galetti, Daniela Basso, Andrea Padoan, Elisa Rossi, Silvia Secco, Michela Pelloso, Paola Fogar, Filippo Navaglia, Stefania Moz, Filiberto Zattoni, and Mario Plebani. Effectiveness of the combined evaluation of klk3 genetics and free-to-total prostate specific antigen ratio for prostate cancer diagnosis. *J Urol*, 188(4):1124–30, Oct 2012.

[27] Carsten Stephan, Sébastien Vincendeau, Alain Houlgatte, Henning Cammann, Klaus Jung, and Axel Semjonow. Multicenter evaluation of [-2]proprostate-specific antigen and the prostate health index for detecting prostate cancer. *Clin Chem*, 59(1):306–14, Jan 2013.

[28] E Floriano-Sánchez, N Cárdenas-Rodríguez, M Castro-Marín, P Alvarez-Grave, and E Lara-Padilla. Dd3(pca3) gene expression in cancer and prostatic hyperplasia. *Clin Invest Med*, 32(6):E258, 2009.

[29] Virginie Vlaeminck-Guillem, Marion Bandel, Martine Cottancin, Claire Rodriguez-Lafrasse, Jean-Marc Bohbot, and Patrice Sednaoui. Chronic prostatitis does not influence urinary pca3 score. *Prostate*, 72(5):549–54, Apr 2012.

[30] Julius Gudmundsson, Soren Besenbacher, Patrick Sulem, Daniel F Gudbjartsson, Isleifur Olafsson, Sturla Arinbjarnarson, Bjarni A Agnarsson, Kristrun R Benediktsdottir, Helgi J Isaksson, Jelena P Kostic, Sigurjon A Gudjonsson, Simon N Stacey, Arnaldur Gylfason, Asgeir Sigurdsson, Hilma Holm, Unnur S Bjornsdottir, Gudmundur I Eyjolfsson, Sebastian Navarrete, Fernando Fuertes, Maria D Garcia-Prats, Eduardo Polo, Ionel A Checherita, Mariana Jinga, Paula Badea, Katja K Aben, Jack A Schalken, Inge M van Oort, Fred C Sweep, Brian T Helfand, Michael Davis, Jenny L Donovan, Freddie C Hamdy, Kristleifur Kristjansson, Jeffrey R Gulcher, Gisli Masson, Augustine Kong, William J Catalona, Jose I Mayordomo,

Gudmundur Geirsson, Gudmundur V Einarsson, Rosa B Barkardottir, Eirikur Jonsson, Viorel Jinga, Dana Mates, Lambertus A Kiemeney, David E Neal, Unnur Thorsteinsdottir, Thorunn Rafnar, and Kari Stefansson. Genetic correction of psa values using sequence variants associated with psa levels. *Sci Transl Med*, 2(62):62ra92, Dec 2010.

[31] Hemang Parikh, Zhaoming Wang, Kerry A Pettigrew, Jinping Jia, Sarah Daugherty, Meredith Yeager, Kevin B Jacobs, Amy Hutchinson, Laura Burdett, Michael Cullen, Liqun Qi, Joseph Boland, Irene Collins, Thomas J Albert, Lars J Vatten, Kristian Hveem, Inger Njølstad, Géraldine Cancel-Tassin, Olivier Cussenot, Antoine Valeri, Jarmo Virtamo, Michael J Thun, Heather Spencer Feigelson, W Ryan Diver, Nilanjan Chatterjee, Gilles Thomas, Demetrius Albanes, Stephen J Chanock, David J Hunter, Robert Hoover, Richard B Hayes, Sonja I Berndt, Joshua Sampson, and Laufey Amundadottir. Fine mapping the KLK3 locus on chromosome 19q13.33 associated with prostate cancer susceptibility and PSA levels. *Human genetics*, 129(6):675–685, June 2011.

[32] David M Good, Visith Thongboonkerd, Jan Novak, Jean-Loup Bascands, Joost P Schanstra, Joshua J Coon, Anna Dominiczak, and Harald Mischak. Body fluid proteomics for biomarker discovery: lessons from the past hold the key to success in the future. *J Proteome Res*, 6(12):4549–55, Dec 2007.

[33] Nagarjuna Nagaraj and Matthias Mann. Quantitative analysis of the intra- and inter-individual variability of the normal urinary proteome. *J Proteome Res*, 10(2):637–45, Feb 2011.

[34] Morgan Rouprêt, Vincent Hupertan, David R Yates, James W F Catto, Ishtiaq Rehman, Mark Meuth, Sylvie Ricci, Roger Lacave, Géraldine Cancel-Tassin, Alexandre de la Taille, François Rozet, Xavier Cathelineau, Guy Vallancien, Freddie C Hamdy, and Olivier Cussenot. Molecular detection of localized prostate cancer using quantitative methylation-specific pcr on urinary cells obtained following prostate massage. *Clin Cancer Res*, 13(6):1720–5, Mar 2007.

[35] F H Meid, C M Gygi, H J Leisinger, F T Bosman, and J Benhattar. The use of telomerase activity for the detection of prostatic cancer cells after prostatic massage. *J Urol*, 165(5):1802–5, May 2001.

[36] Eva Rodríguez-Suárez, Justyna Siwy, Petra Zürbig, and Harald Mischak. Urine as a source for clinical proteome analysis: From discovery to clinical application. *Biochim Biophys Acta*, Jul 2013.

[37] S P Gygi, Y Rochon, B R Franza, and R Aebersold. Correlation between protein and mrna abundance in yeast. *Mol Cell Biol*, 19(3):1720–30, Mar 1999.

[38] J Adachi, C Kumar, Y Zhang, J V Olsen, and M Mann. The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins. *Genome Biol*, 7(9):R80, 2006.

[39] K Benkali, P Marquet, Jp Rérolle, Y Le Meur, and Ln Gastinel. A new strategy for faster urinary biomarkers identification by Nano-LC-MALDI-TOF/TOF mass spectrometry. *BMC genomics*, 9:541, 2008.

[40] C D Calvano, A Aresta, M Iacovone, G E De Benedetto, C G Zambonin, M Battaglia, P Ditonno, M Rutigliano, and C Bettocchi. Optimization of analytical and pre-analytical conditions for MALDI-TOF-MS human urine protein profiles. *J Pharm Biomed Anal*, 51(4):907–914, March 2010.

[41] Georg Martin Fiedler, Sven Baumann, Alexander Leichtle, Anke Oltmann, Julia Kase, Joachim Thiery, and Uta Ceglarek. Standardized peptidome profiling of human urine by magnetic bead separation and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Clin Chem*, 53(3):421–8, Mar 2007.

[42] Qiang He, Lina Shao, Jiekai Yu, Shunxian Ji, Huiping Wang, Youying Mao, and Jianghua Chen. Urinary proteome analysis by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry with magnetic beads for identifying the pathologic presentation of clinical early iga nephropathy. *J Biomed Nanotechnol*, 8(1):133–9, Feb 2012.

[43] Annunziata Lapolla, Laura Molin, Annalisa Sechi, Chiara Cosma, Eugenio Ragazzi, Roberta Seraglia, and Pietro Traldi. A further investigation on a maldi-based method for evaluation of markers of renal damage. *J Mass Spectrom*, 44(12):1754–60, Dec 2009.

[44] Annunziata Lapolla, Roberta Seraglia, Laura Molin, Katherine Williams, Chiara Cosma, Rachele Reitano, Annalisa Sechi, Eugenio Ragazzi, and Pietro Traldi. Low molecular weight proteins in urines from healthy subjects as well as diabetic, nephropathic and diabetic-nephropathic patients: a maldi study. *J Mass Spectrom*, 44(3):419–25, Mar 2009.

[45] Amosy E M'Koma, David L Blum, Jeremy L Norris, Tatsuki Koyama, Dean Billheimer, Saundra Motley, Mayshan Ghiassi, Nika Ferdowsi, Indrani Bhowmick, Sam S Chang, Jay H Fowke, Richard M Caprioli, and Neil A Bhowmick. Detection of pre-neoplastic and neoplastic prostate disease by maldi profiling of urine. *Biochem Biophys Res Commun*, 353(3):829–34, Feb 2007.

[46] Akiko Okamoto, Hayato Yamamoto, Atsushi Imai, Shingo Hatakeyama, Ikuya Iwabuchi, Takahiro Yoneyama, Yasuhiro Hashimoto, Takuya Koie, Noritaka Kamimura, Kazuyuki Mori, Kanemitsu Yamaya, and Chikara Ohyama. Protein profiling of post-prostatic massage urine specimens by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry to discriminate between prostate cancer and benign lesions. *Oncol Rep*, 21(1):73–9, Jan 2009.

[47] I Rehman, A R Azzouzi, J W Catto, S Allen, S S Cross, K Feeley, M Meuth, and F C Hamdy. Proteomic analysis of voided urine after prostatic massage from

patients with prostate cancer: a pilot study. *Urology*, 64(6):1238–1243, December 2004.

[48] Stefan Schaub, John Wilkins, Tracey Weiler, Kevin Sangster, David Rush, and Peter Nickerson. Urine protein profiling with surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry. *Kidney Int*, 65(1):323–332, January 2004.

[49] Michael Karas and Ralf Krüger. Ion formation in maldi: the cluster ionization mechanism. *Chem Rev*, 103(2):427–40, Feb 2003.

[50] Hiroshi Umemura, Yoshio Kodera, and Fumio Nomura. Effects of humidity on the dried-droplet sample preparation for maldi-tof ms peptide profiling. *Clin Chim Acta*, 411(23-24):2109–11, Dec 2010.

[51] Michael Karas, Doris Bachmann, and Franz Hillenkamp. Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. *Anal Chem*, 57(14):2935–2939, 1985.

[52] R Krüger, A Pfenninger, I Fournier, M Gluckmann, and M Karas. Analyte incorporation and ionization in matrix-assisted laser desorption/ionization visualized by ph indicator molecular probes. *Anal Chem*, 73(24):5812–21, Dec 2001.

[53] Ralf Krüger and Michael Karas. Formation and fate of ion pairs during maldi analysis: anion adduct generation as an indicative tool to determine ionization processes. *J Am Soc Mass Spectrom*, 13(10):1218–26, Oct 2002.

[54] Juan Zhang and Renato Zenobi. Matrix-dependent cationization in maldi mass spectrometry. *J Mass Spectrom*, 39(7):808–16, Jul 2004.

[55] R Knochenmuss and R Zenobi. Maldi ionization: the role of in-plume processes. *Chem Rev*, 103(2):441–52, Feb 2003.

[56] Andreas Tholey and Elmar Heinzle. Ionic (liquid) matrices for matrix-assisted laser desorption/ionization mass spectrometry-applications and perspectives. *Anal Bioanal Chem*, 386(1):24–37, Sep 2006.

[57] Gary L Glish and Richard W Vachet. The basics of mass spectrometry in the twenty-first century. *Nat Rev Drug Discov*, 2(2):140–50, Feb 2003.

[58] Peicheng Du, Gustavo Stolovitzky, Peter Horvatovich, Rainer Bischoff, Jihyeon Lim, and Frank Suits. A noise model for mass spectrometry based proteomics. *Bioinformatics*, 24(8):1070–7, Apr 2008.

[59] Andrew N Krutchinsky and Brian T Chait. On the nature of the chemical noise in maldi mass spectra. *J Am Soc Mass Spectrom*, 13(2):129–34, Feb 2002.

[60] Marcus Bantscheff, Markus Schirle, Gavain Sweetman, Jens Rick, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem*, 389(4):1017–1031, August 2007.

[61] M Bantscheff, S Lemeer, M M Savitski, and B Kuster. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem*, 404(4):939–965, September 2012.

[62] A S Benk and C Roesli. Label-free quantification using MALDI mass spectrometry: considerations and perspectives. *Anal Bioanal Chem*, 404(4):1039–1056, September 2012.

[63] Jakob Albrethsen. Reproducibility in protein profiling by MALDI-TOF mass spectrometry. *Clinical chemistry*, 53(5):852–858, May 2007.

[64] Mirre E de Noo, Rob A E M Tollenaar, Aliye Ozalp, Peter J K Kuppen, Marco R Bladergroen, Paul H C Eilers, and André M Deelder. Reliability of human serum protein profiles generated with C8 magnetic beads assisted MALDI-TOF mass spectrometry. *Anal Chem*, 77(22):7232–7241, November 2005.

[65] Glen L Hortin. The MALDI-TOF mass spectrometric view of the plasma proteome and peptidome. *Clinical chemistry*, 52(7):1223–1237, July 2006.

[66] Muriel De Bock, Dominique de Seny, Marie-Alice Meuwis, Jean-Paul Chapelle, Edouard Louis, Michel Malaise, Marie-Paule Merville, and Marianne Fillet. Challenges for biomarker discovery in body fluids using seldi-tof-ms. *J Biomed Biotechnol*, 2010:906082, 2010.

[67] Aly Karsan, Bernhard J Eigl, Stephane Flibotte, Karen Gelmon, Philip Switzer, Patricia Hassell, Dorothy Harrison, Jennifer Law, Malcolm Hayes, Moira Stillwell, Zhen Xiao, Thomas P Conrads, and Timothy Veenstra. Analytical and preanalytical biases in serum proteomic pattern analysis for breast cancer diagnosis. *Clinical chemistry*, 51(8):1525–1528, August 2005.

[68] Rosamonde E Banks, Anthea J Stanley, David A Cairns, Jennifer H Barrett, Paul Clarke, Douglas Thompson, and Peter J Selby. Influences of blood sample processing on low-molecular-weight proteome identified by surface-enhanced laser desorption/ionization mass spectrometry. *Clinical chemistry*, 51(9):1637–1649, September 2005.

[69] Massimo Papale, Maria Carmela Pedicillo, Bradley J Thatcher, Salvatore Di Paolo, Lorenzo Lo Muzio, Pantaleo Bufo, Maria Teresa Rocchetti, Marta Centra, Elena Ranieri, and Loreto Gesualdo. Urine profiling by seldi-tof/ms: monitoring of the critical steps in sample collection, handling and analysis. *J Chromatogr B Analyt Technol Biomed Life Sci*, 856(1-2):205–13, Sep 2007.

[70] Stefano Barelli, David Crettaz, Lynne Thadikkaran, Olivier Rubin, and Jean-Daniel Tissot. Plasma/serum proteomics: pre-analytical issues. *Expert Rev Proteomics*, 4(3):363–70, Jun 2007.

[71] Jianhua Hu, Kevin R Coombes, Jeffrey S Morris, and Keith A Baggerly. The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Brief Funct Genomic Proteomic*, 3(4):322–31, Feb 2005.

[72] D Theodorescu and H Mischak. Mass spectrometry based proteomics in urine biomarker discovery. *World J Urol*, 25(5):435–443, October 2007.

[73] David W Greening and Richard J Simpson. A centrifugal ultrafiltration strategy for isolating the low-molecular weight (<or=25k) component of human plasma proteome. *J Proteomics*, 73(3):637–48, Jan 2010.

[74] V Thongboonkerd, K R McLeish, J M Arthur, and J B Klein. Proteomic analysis of normal human urinary proteins isolated by acetone precipitation or ultracentrifugation. *Kidney Int*, 62(4):1461–1469, October 2002.

[75] Tadashi Yamamoto, Robyn G Langham, Pierre Ronco, Mark A Knepper, and Visith Thongboonkerd. Towards standard protocols and guidelines for urine proteomics: a report on the human kidney and urine proteome project (hkupp) symposium and workshop, 6 october 2007, seoul, korea and 1 november 2007, san francisco, ca, usa. *Proteomics*, 8(11):2156–9, Jun 2008.

[76] S L Cohen and B T Chait. Influence of matrix solution conditions on the MALDI-MS analysis of peptides and proteins. *Anal Chem*, 68(1):31–37, January 1996.

[77] Ali Tiss, Celia Smith, Stephane Camuzeaux, Musarat Kabir, Simon Gayther, Usha Menon, Mike Waterfield, John Timms, Ian Jacobs, and Rainer Cramer. Serum peptide profiling using maldi mass spectrometry: avoiding the pitfalls of coated magnetic beads using well-established ziptip technology. *Proteomics*, 7 Suppl 1:77–89, Sep 2007.

[78] Jeffrey S Morris, Kevin R Coombes, John Koomen, Keith A Baggerly, and Ryuji Kobayashi. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9):1764–75, May 2005.

[79] Chao Yang, Zengyou He, and Weichuan Yu. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC bioinformatics*, 10:4, 2009.

[80] Shuo Chen, Ming Li, Don Hong, Dean Billheimer, Huiming Li, Baogang J Xu, and Yu Shyr. A novel comprehensive wave-form ms data processing method. *Bioinformatics*, 25(6):808–14, Mar 2009.

[81] Clement S Sun and Mia K Markey. Recent advances in computational analysis of mass spectrometry for proteomic profiling. *J Mass Spectrom*, 46(5):443–56, May 2011.

[82] Sanket P Borgaonkar, Harrison Hocker, Hyunjin Shin, and Mia K Markey. Comparison of normalization methods for the identification of biomarkers using maldi-tof and seldi-tof mass spectra. *OMICS*, 14(1):115–26, Feb 2010.

[83] Stephen J Callister, Richard C Barry, Joshua N Adkins, Ethan T Johnson, Wei-Jun Qian, Bobbie-Jo M Webb-Robertson, Richard D Smith, and Mary S Lipton. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J Proteome Res*, 5(2):277–86, Feb 2006.

[84] Judith M Fonville, Claire Carter, Olivier Cloarec, Jeremy K Nicholson, John C Lindon, Josephine Bunch, and Elaine Holmes. Robust data processing and normalization strategy for MALDI mass spectrometric imaging. *Anal Chem*, 84(3):1310–1319, February 2012.

[85] Wouter Meuleman, Judith Ymn Engwegen, Marie-Christine W Gast, Jos H Beijnen, Marcel Jt Reinders, and Lodewyk Fa Wessels. Comparison of normalisation methods for surface-enhanced laser desorption and ionisation (SELDI) time-of-flight (TOF) mass spectrometry data. *BMC bioinformatics*, 9:88, 2008.

[86] Mark W Duncan, Heinrich Roder, and Stephen W Hunsucker. Quantitative matrix-assisted laser desorption/ionization mass spectrometry. *Brief Funct Genomic Proteomic*, 7(5):355–70, Sep 2008.

[87] Sukhvinder S Bansal, John M Halket, Jane Fusova, Adrian Bomford, Robert J Simpson, Nisha Vasavda, Swee Lay Thein, and Robert C Hider. Quantification of hepcidin using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom*, 23(11):1531–42, Jun 2009.

[88] Annalisa Barla, Giuseppe Jurman, Samantha Riccadonna, Stefano Merler, Marco Chierici, and Cesare Furlanello. Machine learning methods for predictive proteomics. *Brief Bioinform*, 9(2):119–28, Mar 2008.

[89] David F Ransohoff. Rules of evidence for cancer molecular-marker discovery and validation. *Nature reviews. Cancer*, 4(4):309–314, April 2004.

[90] David F Ransohoff. Bias as a threat to the validity of cancer molecular-marker research. *Nature reviews. Cancer*, 5(2):142–149, February 2005.

[91] Shadi Toghi Eshghi, Xingde Li, and Hui Zhang. Targeted analyte detection by standard addition improves detection limits in matrix-assisted laser desorption/ionization mass spectrometry. *Anal Chem*, 84(18):7626–7632, September 2012.

[92] Maureen B Tracy, Haijian Chen, Dennis M Weaver, Dariya I Malyarenko, Maciek Sasinowski, Lisa H Cazares, Richard R Drake, O John Semmes, Eugene R Tracy, and William E Cooke. Precision enhancement of maldi-tof ms using high resolution peak detection and label-free alignment. *Proteomics*, 8(8):1530–8, Apr 2008.

[93] BRADLEY EFRON. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599, 1981.

[94] Bernard Rosner. *Fundamentals of Biostatistics*. BROOKS/COLE, seventh edition edition, December 2012.

[95] Enrique F Schisterman, Albert Vexler, Brian W Whitcomb, and Aiyi Liu. The limitations due to exposure detection limits for regression models. *Am J Epidemiol*, 163(4):374–383, February 2006.

[96] Raymond J Carroll, David Ruppert, Stefansky Leonard, and Ciprian Crainiceanu. *Measurement Error in Nonlinear models - A model perspective*. Chapman and Hall/CRC, second edition edition.

[97] C G Fraser and E K Harris. Generation and application of data on biological variation in clinical chemistry. *Crit Rev Clin Lab Sci*, 27(5):409–37, 1989.

[98] David E Neal, Jenny L Donovan, Richard M Martin, and Freddie C Hamdy. Screening for prostate cancer remains controversial. *Lancet*, 374(9700):1482–3, Oct 2009.

[99] H Ballentine Carter, Peter C Albertsen, Michael J Barry, Ruth Etzioni, Stephen J Freedland, Kirsten Lynn Greene, Lars Holmberg, Philip Kantoff, Badrinath R Konety, Mohammad Hassan Murad, David F Penson, and Anthony L Zietman. Early detection of prostate cancer: Aua guideline. *J Urol*, 190(2):419–26, Aug 2013.

[100] Axel Heidenreich, Per-Anders Abrahamsson, Walter Artibani, James Catto, Francesco Montorsi, Hein Van Poppel, Manfred Wirth, and Nicolas Mottet. Early detection of prostate cancer: European association of urology recommendation. *Eur Urol*, 64(3):347–54, Sep 2013.

[101] Chris H Bangma, Ron H van Schaik, Bert G Blijenberg, Monique J Roobol, Hans Lilja, and Ulf-Håkan Stenman. On the use of prostate-specific antigen for screening of prostate cancer in european randomised study for screening of prostate cancer. *Eur J Cancer*, 46(17):3109–19, Nov 2010.

[102] E David Crawford, Kyle O Rove, Edouard J Trabulsi, Junqi Qian, Krystyna P Drewnowska, Jed C Kaminetsky, Thomas K Huisman, Mark L Bilowus, Sheldon J Freedman, W Lloyd Glover, Jr, and David G Bostwick. Diagnostic performance of pca3 to detect prostate cancer in men with increased prostate specific antigen: a prospective study of 1,962 cases. *J Urol*, 188(5):1726–31, Nov 2012.

[103] Linda A Bradley, Glenn E Palomaki, Steven Gutman, David Samson, and Naomi Aronson. Comparative effectiveness review: prostate cancer antigen 3 testing for the diagnosis and management of prostate cancer. *J Urol*, 190(2):389–98, Aug 2013.

[104] Marco Auprich, Herbert Augustin, Lars Budäus, Luis Kluth, Sebastian Mannweiler, Shahrokh F Shariat, Margit Fisch, Markus Graefen, Karl Pummer, and Felix K-H Chun. A comparative performance analysis of total prostate-specific antigen, percentage free prostate-specific antigen, prostate-specific antigen velocity and urinary prostate cancer gene 3 in the first, second and third repeat prostate biopsy. *BJU Int*, 109(11):1627–35, Jun 2012.

[105] Anne Helene Garde, Ase Marie Hansen, Jesper Kristiansen, and Lisbeth Ehlert Knudsen. Comparison of uncertainties related to standardization of urine samples with volume and creatinine concentration. *Ann Occup Hyg*, 48(2):171–9, Mar 2004.

[106] S R Cole, H Chu, L Nie, and E F Schisterman. Estimating the odds ratio when exposure has a limit of detection. *Int J Epidemiol*, 38(6):1674–1680, December 2009.

[107] Lei Nie, Haitao Chu, Chenglong Liu, Stephen R Cole, Albert Vexler, and Enrique F Schisterman. Linear regression with an independent variable subject to a detection limit. *Epidemiology*, 21 Suppl 4(Supplement):S17–S24, July 2010.

[108] Paul Hewett and Gary H Ganser. A comparison of several methods for analyzing censored data. *Ann Occup Hyg*, 51(7):611–632, October 2007.

[109] Andrea Sboner, Francesca Demichelis, Stefano Calza, Yudi Pawitan, Sunita R Setlur, Yujin Hoshida, Sven Perner, Hans-Olov Adami, Katja Fall, Lorelei A Mucci, Philip W Kantoff, Meir Stampfer, Swen-Olof Andersson, Eberhard Varenhorst, Jan-Erik Johansson, Mark B Gerstein, Todd R Golub, Mark A Rubin, and Ove Andrén. Molecular sampling of prostate cancer: a dilemma for predicting disease progression. *BMC Med Genomics*, 3:8, 2010.

[110] May A Beydoun, Jay S Kaufman, Joseph Ibrahim, Jessie A Satia, and Gerardo Heiss. Measurement error adjustment in essential fatty acid intake from a food frequency questionnaire: alternative approaches and methods. *BMC Med Res Methodol*, 7:41, 2007.

[111] K Y Fung and D Krewski. Evaluation of regression calibration and simex methods in logistic regression when one of the predictors is subject to additive measurement error. *J Epidemiol Biostat*, 4(2):65–74, 1999.

[112] Enrique F Schisterman and Roderick J Little. Opening the black box of biomarker measurement error. *Epidemiology*, 21 Suppl 4(Supplement):S1–S3, July 2010.

[113] Paul H C Eilers, Esther Röder, Huub F J Savelkoul, and Roy Gerth van Wijk. Quantile regression for the statistical analysis of immunological data with many non-detects. *BMC immunology*, 13:37, 2012.

[114] Liya Fu and You-Gan Wang. Nonparametric rank regression for analyzing water quality concentration data with multiple detection limits. *Environ Sci Technol*, 45(4):1481–9, Feb 2011.

[115] K Krishnamoorthy, A Mallick, and T Mathew. Model-based imputation approach for data analysis in the presence of non-detects. *Ann Occup Hyg*, 53(3):249–263, April 2009.

[116] MinJae Lee, Lan Kong, and Lisa Weissfeld. Multiple imputation for left-censored biomarker data based on Gibbs sampling method. *Stat Med*, 31(17):1838–1848, July 2012.

[117] Neil J Perkins, Enrique F Schisterman, and Albert Vexler. Generalized ROC curve inference for a biomarker subject to a limit of detection and measurement error. *Stat Med*, 28(13):1841–1860, June 2009.

[118] D B Richardson and A Ciampi. Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *Am J Epidemiol*, 157(4):355–363, February 2003.

[119] Enrique F Schisterman and Albert Vexler. To pool or not to pool, from whether to when: applications of pooling to biospecimens subject to a limit of detection. *Paediatr Perinat Epidemiol*, 22(5):486–496, September 2008.

[120] Ram B Jain and Richard Y Wang. Limitations of maximum likelihood estimation procedures when a majority of the observations are below the limit of detection. *Anal Chem*, 80(12):4767–4772, June 2008.

[121] A H El-Shaarawi and S R Esterby. Replacement of censored observations by a constant: An evaluation. *Water Research*, 26(6):835–844, June 1992.

[122] Dennis R Helsel and Timothy A Cohn. Estimation of descriptive statistics for multiply censored water quality data. *Water Resour Res*, 24(12):1997–2004, July 2010.

[123] Jeffrey S Buzas, Leonard A Stefanski, and Tor D Tosteson. *Handbook of Epidemiology*. Measurement Error. Springer.

[124] R L Prentice. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69(2):331–342, 1982.

[125] Callum G Fraser. Inherent biological variation and reference values. *Clin Chem Lab Med*, 42(7):758–64, 2004.

[126] Somnath Sarkar and Yongming Qu. Quantifying the treatment effect explained by markers in the presence of measurement error. *Stat Med*, 26(9):1955–1963, April 2007.

[127] Raymond J Carroll. *Measurement Error in Epidemiologic Studies*. John Wiley & Sons, Ltd, Chichester, UK, July 2005.

[128] D Spiegelman, S Schneeweiss, and A McDermott. Measurement error correction for logistic regression models with an "alloyed gold standard". *Am J Epidemiol*, 145(2):184–196, January 1997.

[129] J R Cook and L A Stefanski. Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *J Am Stat Assoc*, 89(428):1314–1328, December 1994.

[130] Terence J O'Neill, Simon C Barry, and Borek Puza. *Monte Carlo Methods*. John Wiley & Sons, Ltd, Chichester, UK, September 2008.

[131] P Diaconis and B Efron. Computer-intensive methods in statistics. *Sci Am*, 1983.