Published in final edited form as:
Clin Chem. 2014 July ; 60(7): 941–944. doi:10.1373/clinchem.2014.224840.

From Lost in Translation to Paradise Found: Enabling Protein Biomarker Method Transfer Using Mass Spectrometry

Russell P. Grant[1] and **Andrew N. Hoofnagle**[2]

[1]Laboratory Corporation of America, University of Washington, Seattle

[2]Departments of Laboratory Medicine and Medicine, University of Washington, Seattle

Keywords

liquid chromatography-tandem mass spectrometry; validation; protein biomarkers; calibration; selectivity; reproducibility

Recently, Begley and Ellis delivered a sobering message to laboratories around the world (1). The lack of meaningful progress in preclinical cancer research was highlighted by the irreproducibility of >70% of published studies. The authors also crystallized the importance of full disclosure and the validation of critical scientific discoveries for industry-wide improvement. Translation of novel biomarkers into clinical care for the evaluation of therapeutic safety and efficacy has been slow (2), partly attributable to the cost and complexity of immunoassay development. The potential for liquid chromatography-tandem mass spectrometry (LC-MS/MS) to streamline the translation of novel protein biomarkers is profound (3).

Most LC-MS/MS based protein assays incorporate denaturation and proteolytic digestion of proteins in the sample into peptides (traditionally called “bottom-up” proteomics). These preparative steps destroy potentially interfering proteins into peptides that can be resolved and ignored by LC-MS/MS (4). Inclusion of stable isotope-labeled internal standard proteins or peptides (which may be cleavable) in each sample enables correction for matrix effects, including sample-related digestion variability and/or ion suppression, both significant analytical benefits compared to immunoassays.

Downstream members of the scientific community are hopeful of translating important preliminary findings into clinical practice; however, success has been hampered by a lack of transparency and insufficient validation. Consequently, LC-MS/MS-based clinical protein analysis has predominantly focused on improved analytical measurement for well-established biomarkers (5). This is despite “fit-for-purpose” criteria for enablement (6, 7) and published recommendations for analytical validation (8), based primarily upon FDA guidance (9). While assays used in preclinical research are generally not held to the same standards as assays used in the immediate care of patients, which are governed by CLIA-88

Corresponding authors: Andrew Hoofnagle, MD PhD, Campus Box 357110, University of Washington, Seattle, WA 98195-7110. *Ph:* 206-598-6131. *Fax:* 206-598-6189. ahoof@u.washington.edu and Russell Grant, PhD, 1447 York Court, Burlington, NC 27215. *Ph:* 336-436-3605. *Fax:* 336-436-0639. grantr@labcorp.com.

and by extension many CLSI consensus documents, published fundamental discovery experiments and biomarker verification studies spawn costly research programs. To advance our efforts as a community, LC-MS/MS protein quantification data used to support published research findings should be from properly designed studies (10) and accompanied by a standard operating procedure that includes sufficient detail to facilitate assay reproduction in other laboratories.

The careful definition of the measurand is essential (11). Appropriate method validation should include experiments that evaluate and document key analytical performance characteristics. To this end, we present and discuss a minimal list of experiments (Table 1) that would allow downstream users of novel biomarkers to carefully evaluate their quality and potential reproducibility. While certain experiments are uniquely associated with bottom-up LC-MS/MS proteomics workflows, we believe that there are universally applicable concepts described within this document that should be applied to alternate technologies in biomarker translation. Given the lessons of the past (1), we cannot overemphasize the importance of this level of transparency and rigor in the publication of novel scientific discoveries.

Imprecision, Repeatability, and Reproducibility

For the vast majority of clinical and preclinical assays, precise measurements facilitate longitudinal monitoring of disease, resolution of the disease continuum with confidence, and hypothesis verification. The imprecision of an assay can be assessed within batch (repeatability) or longitudinally and between laboratories (reproducibility). We propose the use of 2 pools, presumably with different concentrations, for the marker of interest, a “disease” pool comprising equal volumes of known disease samples (ideally n=20) and a “healthy” pool [n=400, which was derived from Ichihara, *et al.* (12)]. We have recently collaborated to enable a commercial source of pools from healthy controls (EDTA plasma and serum, men or women 20–50 years, Golden West Biologicals, Temecula, CA). An estimation of total variability per pool should be determined from 5 individual replicates of each pool assayed each day for 5 days. The mean intra-assay (CV_{intra}) and inter-assay (CV_{inter}) coefficients of variation for each pool should be calculated. CV_{intra} includes all 5 replicates (#1 – #5) per pool measured in a single day. CV_{inter} is determined for each replicate individually (CV for replicate #1 across 5 days, #2 across 5-days, *etc.*) across all 5 days. The total variability is determined using the mean CV_{intra} (across 5 days) and the mean CV_{inter} (across all 5 replicates) by the sum of squares: $CV_{total} = (CV_{intra}^2 + CV_{inter}^2)^{1/2}$ and is reported for each pool. The determination of mean inter-assay protein concentration ($Inter_{mc}$) for these pools (*i.e.*, the mean of 25 results for each pool) will be used for subsequent validation experiments (the concentration measured in each sample is calculated as the ratio of the endogenous peak area to the spiked internal standard peak area multiplied by the concentration of internal standard spiked into the samples and, if appropriate, results from multiple peptides are averaged for each sample). If internal standard peptide spiked after digestion is used in the calculation of the concentration, it is unlikely to be entirely accurate due to incomplete proteolytic digestion and non-linearity of the ratio when it deviates from 1.0, but it will provide a frame of reference for subsequent experiments. We

further propose that all subsequent experiments described in this paper be determined using triplicate samples, with CV reported at each concentration for each experiment.

Bias and Accuracy

Accuracy is often difficult to achieve for protein assays due to the lack of standard reference materials and assays, particularly for novel biomarkers. We propose the use of the inter-assay mean concentration determined for the healthy pool (or disease pool where biomarkers are normally absent) as a calibration material in preclinical experiments. The intrinsically normalizing size of the healthy pool offers a concentration anchor point ($Inter_{mc}$) for comparative accuracy purposes to improve repeatability and reproducibility concordance (13, 14).

The majority of preclinical research studies incorporate isotope-labeled internal standard peptides (IS) after digestion. However, the influence of proteolytic peptide formation/degradation relative to IS and its effect on assay bias must be determined (13). The disease and healthy pools are proteolyzed with IS addition pre-digestion (IS_{pre}) and protein concentrations are compared to $Inter_{mc}$ with IS addition post digestion (IS_{post}). Estimation of bias for protein determination due to peptide degradation during the proteolysis step is calculated as ($IS_{pre} - IS_{post}$)/ IS_{post}, expressed as a percentage. This experiment should be performed at least twice, but can be eliminated if internal standards are routinely added pre-digestion.

Linearity and Limit of Quantification

While the imprecision of a preclinical assay is important in distinguishing diseased from healthy individuals or one pathophysiologically important state from another, a narrow analytical dynamic range can make this difficult. To evaluate linearity, we propose a 5-point mixing scheme. The study includes the disease and healthy pools described above, together with 3:1, 1:1, and 1:3 admixtures of these pools prior to sample preparation. Admixture recoveries should be calculated against expected protein concentrations generated via linear extrapolation of expected disease and healthy pool $Inter_{mc}$ results (from the 5 replicate-5 day experiment above) and the ratio of admixtures (*e.g.*, expected 1:1 mixture concentration = mean of disease $Inter_{mc}$ and healthy $Inter_{mc}$). This experiment should be performed at least twice and highlights the analytical capability for disease differentiation at the individual analyte level, together with a preliminary determination of matrix effects.

Dilution studies of the healthy pool are used to estimate the lower limit of quantification when analyte is present (disease pool when analyte is absent). Healthy pool should be gravimetrically diluted (serial 2–5-fold dilutions) with analyte-free surrogate or alternate species matrix until analyte is no longer quantifiable. This experiment should be performed at least twice; recovery (accounting for dilution) and imprecision should be reported.

Matrix Effects and Selectivity

In addition to evaluating for matrix effects using linearity, we also propose to evaluate the effects of common clinical interferences. A test kit containing supraphysiological

interferences has recently been commercialized for this study (Assurance Interference Test Kit, Sun Diagnostics, New Gloucester, ME). Evaluation of bias is performed for lipemia (triglycerides of 3000 mg/dL or 33.9 mmol/L), hemolysis (hemoglobin of 500 mg/dL), icterus (bilirubin of 20 mg/dL or 342 μmol/L) and hyperproteinemia (total protein 12 g/dL). Influence of clinical interferents (determined as % bias) is performed by spiking interferents into the healthy pool, measuring the protein concentration, and comparing to the healthy $Inter_{mc}$, accounting for dilution in the expected concentration. When the spiked interferent contains the protein analyte, the concentration of analyte in the spiked interferent should be determined from a 1:1 admixture of the interferent and an analyte-free matrix. This concentration should be used to determine the contribution to the measured concentration of the interferent-spiked healthy pool and subtracted to evaluate for bias.

In routine clinical LC-MS/MS assays of small molecules, transition ratio monitoring (ratio of quantifying transition peak area to the qualifying transition peak area) is used to document selectivity of the approach, identifying samples with isotopic/isobaric interferences and thereby providing confidence in concentration assignment (15). This has been expanded to include proteins (5, 16), whereby alternative products of the same peptide generated by the mass spectrometer can be used to confirm the identity of the molecule being quantified. For each of the validation studies performed, we propose the disclosure of transition ratio monitoring results. In addition, transition ratio monitoring results must be disclosed for all samples assayed during preclinical studies.

Analyte Stability

In routine bioanalytical assay validation (9), assessment of stability requires purified analyte for generation of fresh calibration standards to assay samples both pre- and post-stressed-storage. We propose a relative bias approach (against $Inter_{mc}$) using both disease and healthy pools for stability evaluation. Frozen storage bias is assessed through the analysis of pool aliquots assayed 30 days after generation of $Inter_{mc.}$ Sample handling stability bias should be determined on pre-extracted samples following storage of aliquots at room temperature (20–24°C) for 4 h, refrigeration (4–8°C) for 24 h, and up to 2 freeze/thaw cycles. Post extraction stability should be determined for both pools following storage in the autosampler (>24 h, re-injecting aliquots if feasible), frozen (>72 h, if routine), and following extract freeze/thaw for 1 and 2 cycles. Since many preclinical studies rely on bio-banked materials, it should be noted that at least 3 freshly acquired samples should be evaluated for stability of one freeze-thaw cycle (assay fresh, freeze for >12 h, thaw for >2 h, re-assay and compare).

Transparency and Disclosure

For complete transparency, we propose that authors submit processed analytical data to a web-based repository, such as Panorama/Skyline (16), to enable detailed critical review of published results and the human influence in data reduction (10). A meticulous description of key reagents used in each assay should be included in the supplemental data section together with the standard operating procedure(s) used to perform the preclinical studies. The proposed framework will enable us as a community to fully evaluate the potential of

novel biomarkers published in the literature. If those biomarkers are truly discriminatory, we can improve patient care.

Acknowledgments

This work was supported in part by the National Cancer Institute/National Institutes of Health (U24 CA160034)

Abbreviations

LC-MS/MS	liquid chromatography-tandem mass spectrometry

References

1. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012; 483:531–533. [PubMed: 22460880]
2. Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: The long and uncertain path to clinical utility. Nat Biotechnol. 2006; 24:971–983. [PubMed: 16900146]
3. Paulovich AG, Whiteaker JR, Hoofnagle AN, Wang P. The interface between biomarker discovery and clinical validation: The tar pit of the protein biomarker pipeline. Proteomics Clin Appl. 2008; 2:1386–1402. [PubMed: 20976028]
4. Hoofnagle AN, Wener MH. The fundamental flaws of immunoassays and potential solutions using tandem mass spectrometry. J Immunol Methods. 2009; 347:3–11. [PubMed: 19538965]
5. Kushnir MM, Rockwood AL, Roberts WL, Abraham D, Hoofnagle AN, Meikle AW. Measurement of thyroglobulin by liquid chromatography-tandem mass spectrometry in serum and plasma in the presence of antithyroglobulin autoantibodies. Clin Chem. 2013; 59:982–990. [PubMed: 23396140]
6. Carr SA, Abbatiello SE, Ackermann BL, Borchers C, Domon B, Deutsch EW, et al. Targeted peptide measurements in biology and medicine: Best practices for mass spectrometry-based assay development using a fit-for-purpose approach. Mol Cell Proteomics. 2014; 13:907–917. in press. [PubMed: 24443746]
7. Lee JW, Devanarayan V, Barrett YC, Weiner R, Allinson J, Fountain S, et al. Fit-for-purpose method development and validation for successful biomarker measurement. Pharm Res. 2006; 23:312–328. [PubMed: 16397743]
8. DeSilva B, Garofolo F, Rocci M, Martinez S, Dumont I, Landry F, et al. 2012 white paper on recent issues in bioanalysis and alignment of multiple guidelines. Bioanalysis. 2012; 4:2213–2226. [PubMed: 23046264]
9. U.S. Food and Drug Administration (FDA). [Accessed February 18 2014] Guidance for industry: Bioanalytical method validation. http://www.fda.gov/downloads/Drugs/Guidances/ucm070107.pdf
10. Oberg AL, Vitek O. Statistical design of quantitative mass spectrometry-based proteomic experiments. J Proteome Res. 2009; 8:2144–2156. [PubMed: 19222236]
11. Miller WG, Myers GL, Lou Gantzer M, Kahn SE, Schonbrunner ER, Thienpont LM, et al. Roadmap for harmonization of clinical laboratory measurement procedures. Clin Chem. 2011; 57:1108–1117. [PubMed: 21677092]
12. Ichihara K, Boyd JC, Intervals ICoR, Decision L. An appraisal of statistical procedures used in derivation of reference intervals. Clin Chem Lab Med. 2010; 48:1537–1551. [PubMed: 21062226]
13. Agger SA, Marney LC, Hoofnagle AN. Simultaneous quantification of apolipoprotein a-i and apolipoprotein b by liquid-chromatography-multiple- reaction-monitoring mass spectrometry. Clin Chem. 2010; 56:1804–1813. [PubMed: 20923952]
14. Cox HD, Lopes F, Woldemariam GA, Becker JO, Parkin MC, Thomas A, et al. Interlaboratory agreement of insulin-like growth factor 1 concentrations measured by mass spectrometry. Clin Chem. 2013
15. Kushnir MM, Rockwood AL, Nelson GJ, Yue B, Urry FM. Assessing analytical specificity in quantitative analysis using tandem mass spectrometry. Clin Biochem. 2005; 38:319–327. [PubMed: 15766733]

16. Abbatiello SE, Mani DR, Keshishian H, Carr SA. Automated detection of inaccurate and imprecise transitions in peptide quantification by multiple reaction monitoring mass spectrometry. Clin Chem. 2010; 56:291–305. [PubMed: 20022980]

Table 1

List of minimal experiments for assay validation of LC-MS/MS protein quantification.

Experiment	Description	Determination	Best Practice[a]
Reproducibility	Healthy and disease pools are analyzed 5 times on each of 5 days.	CV_{intra} and CV_{Inter}, CV_{total} as the sum of squares.	CV_{intra} and CV_{Inter}≤20%
Peptide Stability	Internal standard peptides are spiked before and after digestion to both pools.	Bias and CV of triplicate samples when IS added pre-digestion versus post-digestion.	Bias, CV ≤20%
Linearity	Healthy and disease pools are admixed 3:1, 1:1 and 1:3.	Bias and CV of triplicate admixed samples compared to extrapolated values from $Inter_{mc}$ determinations.	Bias, CV ≤20%
LLOQ[b]	Healthy pool is diluted with an analyte-free surrogate matrix or matrix from another species.	Bias and CV of triplicate diluted samples compared to expected values from $Inter_{mc}$ determinations incorporating dilution factor.	Bias, CV ≤25%
Interferences	Clinically relevant potential interferents are added to the healthy pool.	CV of triplicate spiked samples. Bias when accounting for dilution of spiking (5% – 50% dilution depending on interferent solution) compared to expected values from $Inter_{mc}$ determination.	Bias, CV ≤ 20%
Stability	Healthy and disease pools are stressed before and after sample preparation.	Bias and CV of triplicate samples compared to expected values from $Inter_{mc}$ determinations.	Bias, CV ≤ 20%

[a] Best practice acceptance criterion as defined by Ref. 7, acknowledged as a hybrid of Immunoassays and LC-MS/MS validation criteria derived from Ref. 8.

[b] The lower limit of quantification.