MINING THE BIOMEDICAL LITERATURE FOR DETAILED INFORMATION ABOUT BIOSPECIMENS

Paul Fearn, MBA

PhD Candidate

University of Washingon

Biomedical and Health Informatics

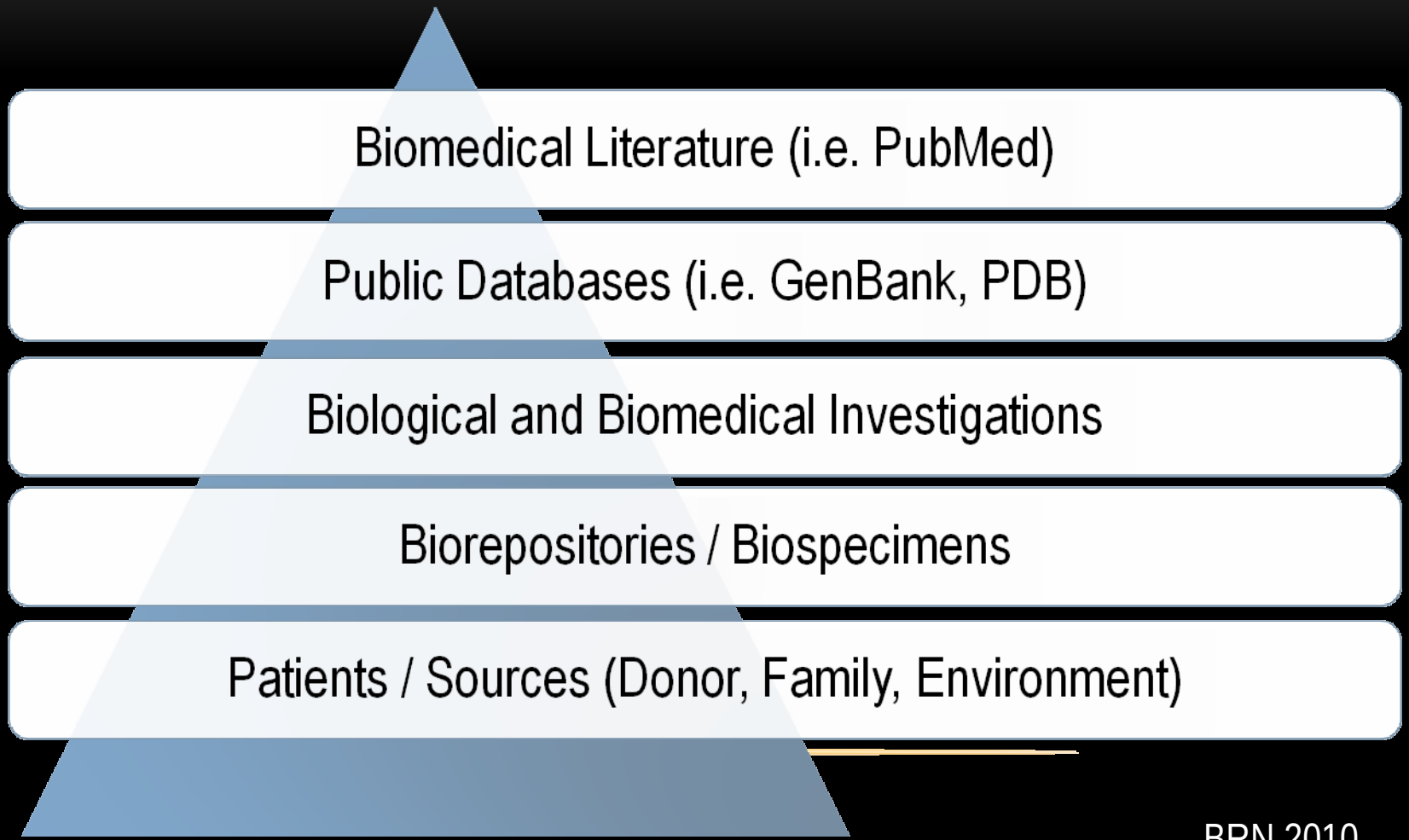
BRN Symposium, February 22-23, 2012

Support from NLM training (NIH NLM #T15 LM07442) and ITHS (NIH NCRR RR 025014) grants

PROBLEMS AND MOTIVATION

- Problems
  - Biospecimen information may be “lost in translation” from patient care to results of experiments in biomedical literature
  - Post-hoc statistical correction for batch effects?
  - Loss of statistical power, as well as accuracy and reproducibility of biospecimen experiments
  - High costs of data management ~ Herbek G, Grizzle W @ BRN 2012
- Motivations
  - Improve accuracy and reproducibility of research using biospecimens
  - Accelerate turnaround time and quality of biomedical research
  - Reduce costs (and/or increase value) of information abstraction

INFORMATION FRAMEWORK FOR BIOSPECIMEN RESEARCH

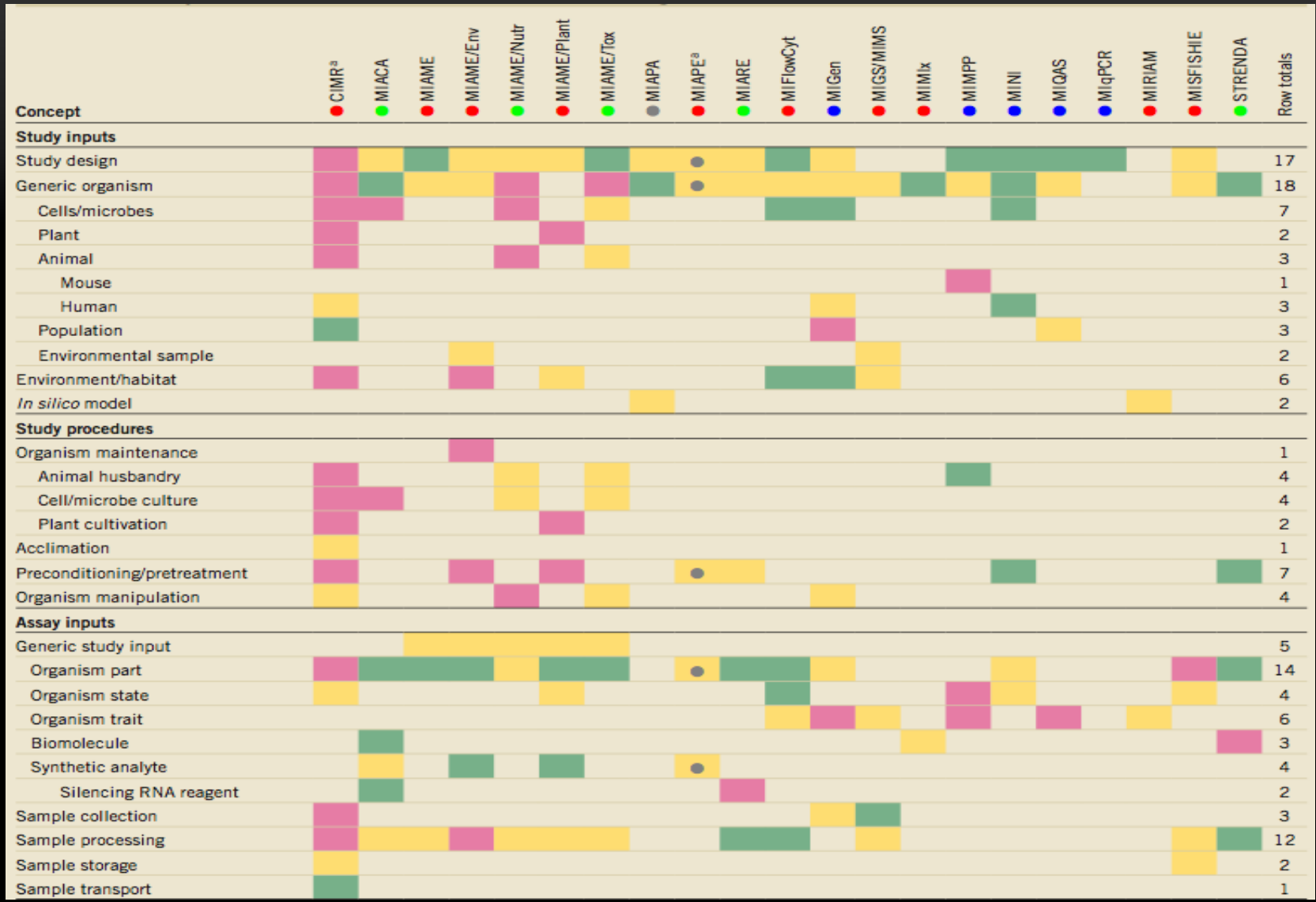


BRN 2010

RELATED WORK IN GUIDELINES AND STANDARDS

- Reporting checklists for biomedical and biological experiments
  - MIAME, MISFISHIE,… MIBBI
  - Biospecimen Reporting for Improved Study Quality (BRISQ)
- Controlled vocabularies, Terminologies, Ontologies
  - OBI, UMLS,…BioPortal
- Database or object models (e.g. FuGE, BioSample)
- Data formats (e.g. ISA-TAB, SPREC)
- Have these information guidelines and standards brought about measurable effects on information reported in literature?
- “Have to show the data to change behavior!” ~ Herbek, G

UNEVEN COVERAGE IN MIBBI CHECKLISTS



Taylor et al. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nature Biotechnology, 26(8), 889–896. doi:10.1038/nbt.1411

ZOOMING IN ON BIOSPECIMEN LIFECYCLE AND PRE-ANALYTIC VARIATION

FIGURE 1. The lifecycle of the biospecimen is illustrated. The preanalytical phase of the lifecycle of the biospecimen includes each stage from patient to distribution. Preanalytical variables are addressed in the BRISQ list.

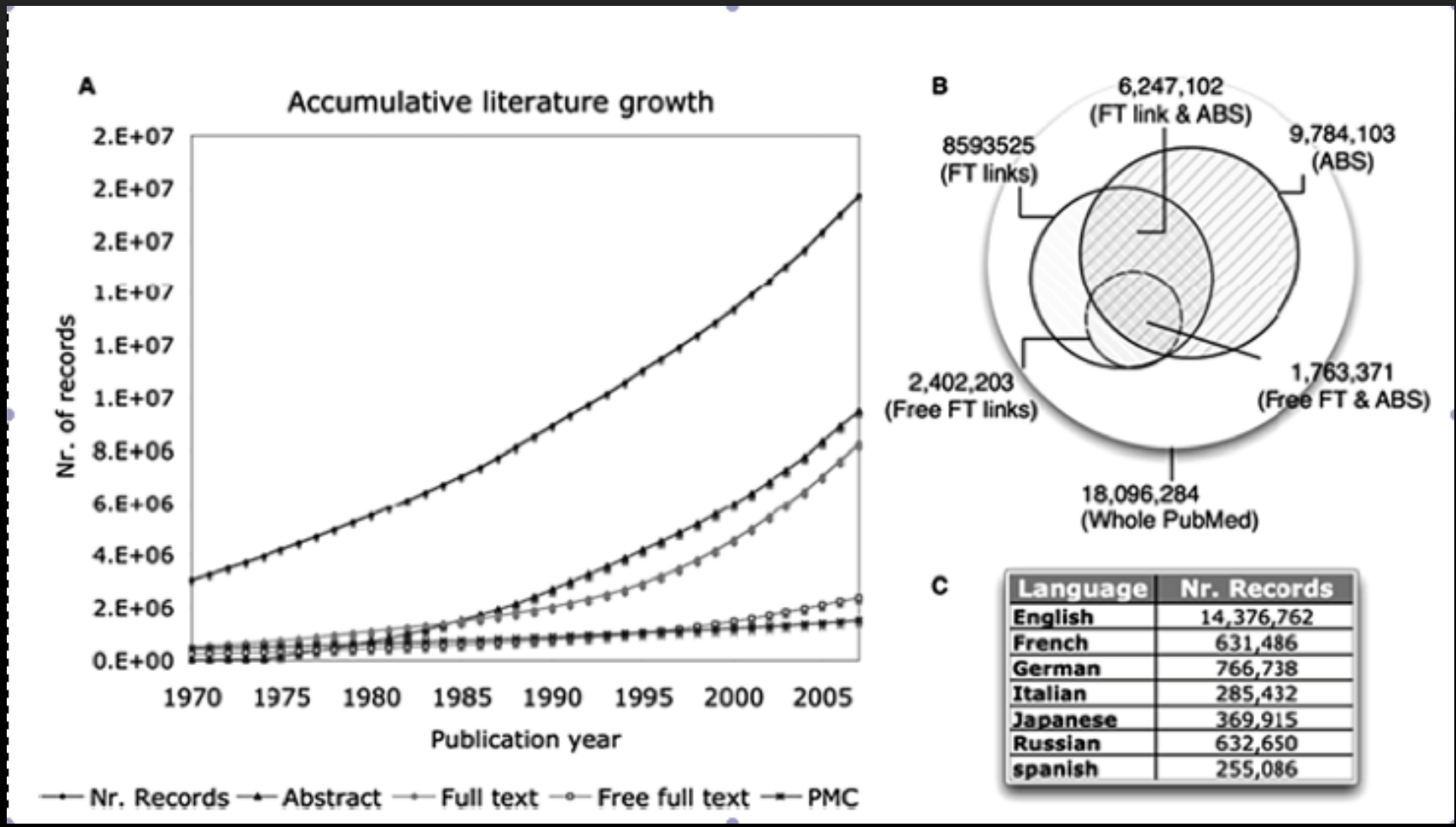
Moore HM et al. Biospecimen reporting for improved study quality (BRISQ). Cancer Cytopathol 2011;119(2):92–101.

BIOSPECIMEN REPORTING FOR IMPROVED STUDY QUALITY

	Data Elements	Examples
☐	**Biospecimen type** *Solid tissue, whole blood, or another product derived from a human being*	Serum, Urine
☐	**Anatomical site** *Organ of origin or site of blood draw*	Liver, Antecubital area of the arm
☐	**Disease status of patients** *Controls or individuals with the disease of interest*	Diabetic, Healthy control
☐	**Clinical characteristics of patients** *Available medical information known or believed to be pertinent to the condition of the biospecimens*	Pre-menopausal breast cancer patients
☐	**Vital State of patients** *Alive or deceased patient when biospecimens were obtained*	Postmortem
☐	**Clinical diagnosis of patients** *Patient clinical diagnoses (determined by medical history, physical examination, and analyses of the biospecimen) pertinent to the study*	Breast cancer
☐	**Pathology diagnosis** *Patient pathology diagnoses (determined by macro and/or microscopic evaluation of the biospecimen at the time of diagnosis and/or prior to research use) pertinent to the study*	Her2-negative intraductal carcinoma
☐	**Collection mechanism** *How the biospecimens were obtained*	Fine needle aspiration, Pre-operative blood draw
☐	**Type of stabilization** *The initial process by which biospecimens were stabilized during collection*	Heparin, On ice
☐	**Type of long-term preservation** *The process by which the biospecimens were sustained after collection*	Formalin fixation, freezing
☐	**Constitution of preservative** *The make-up of any formulation used to maintain the biospecimens in a non-reactive state*	10% neutral-buffered formalin, 10 USP Heparin Units/mL
☐	**Storage temperature** *The temperature or range thereof at which the biospecimens were kept until distribution/analysis.*	-80 °C, 20 to 25 °C
☐	**Storage duration** *The time or range thereof between biospecimen acquisition and distribution or analysis.*	8 days, 5 to 7 years
☐	**Shipping temperature** *The temperature or range thereof at which biospecimens were kept during shipment or relocation.*	-170 °C to -190 °C
☐	**Composition assessment & selection** *Parameters used to choose biospecimens for the study*	Minimum 80% tumor nuclei & maximum 50% necrosis

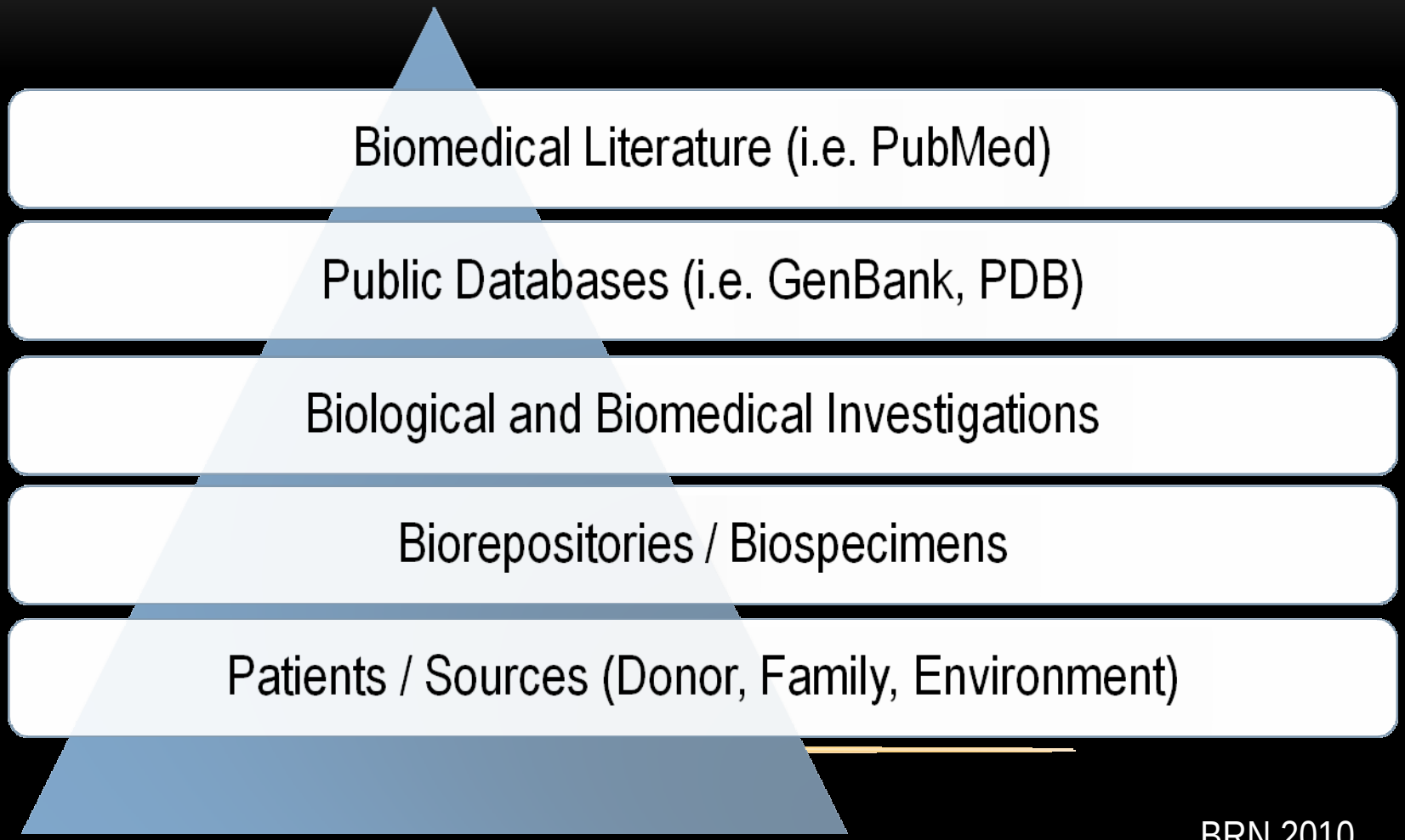
Moore, H. M., Compton, C. C., Alper, J., & Vaught, J. B. (2011). International Approaches to Advancing Biospecimen Science. *Cancer Epidemiology Biomarkers & Prevention*, *20*(5), 729–732. doi:10.1158/1055-9965.EPI-11-0021

BUT, RAPID GROWTH IN LITERATURE…



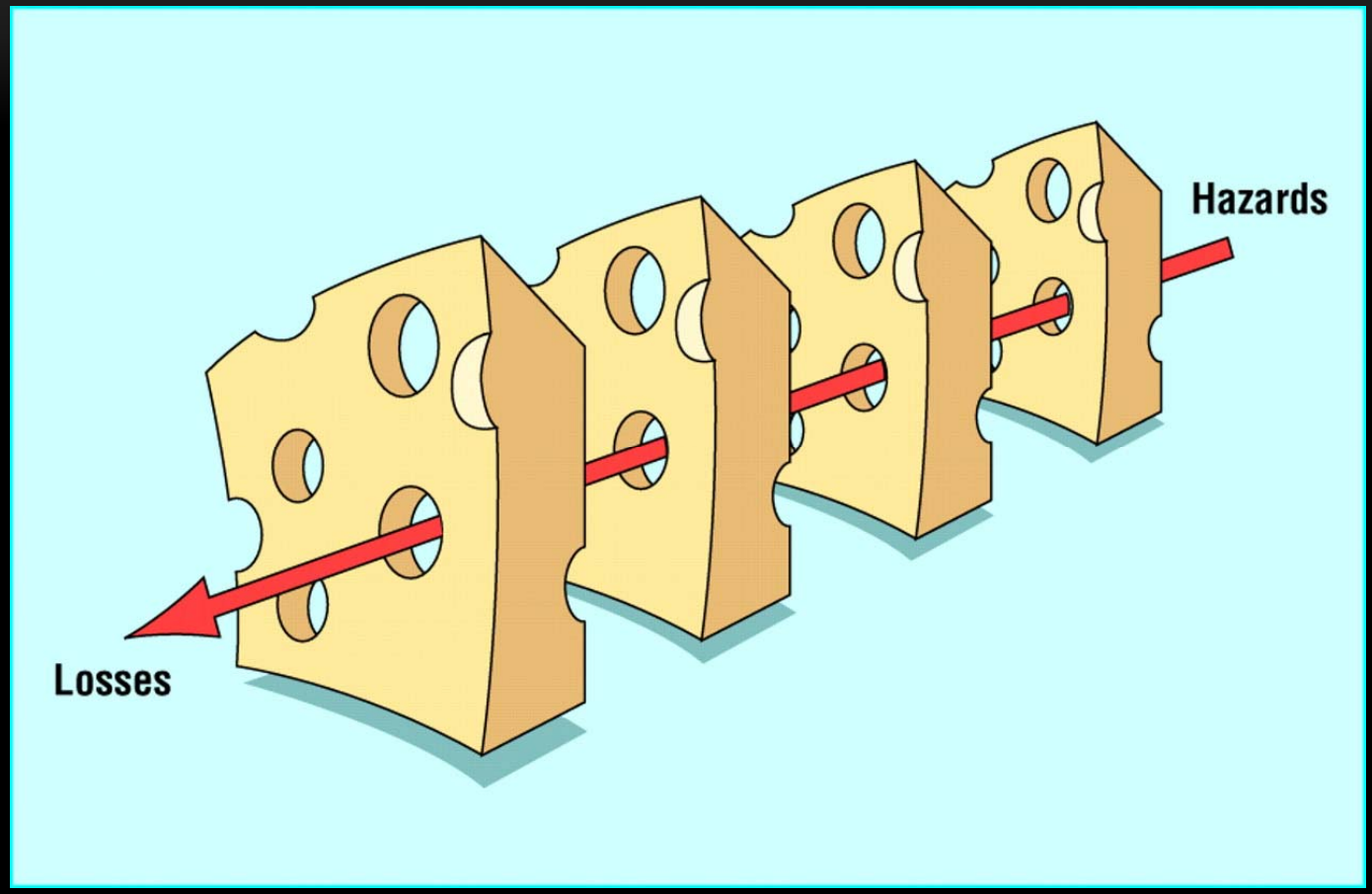
Krallinger M, Leitner F, Valencia A. Analysis of biological processes and diseases using text mining approaches. Methods Mol Biol 2010;593:341–382.

LOTS OF PLACES WHERE SLIPS OR ERRORS CAN OCCUR…



BRN 2010

JAMES REASON'S SWISS CHEESE MODEL



Clinic/OR

Pathology

Repository

Laboratory

Human error: models and management. (2000). Human error: models and management. BMJ

HIERARCHY OF INTERVENTIONS TO CHANGE PRACTICE, FROM STRONGEST (1) TO WEAKEST (4)

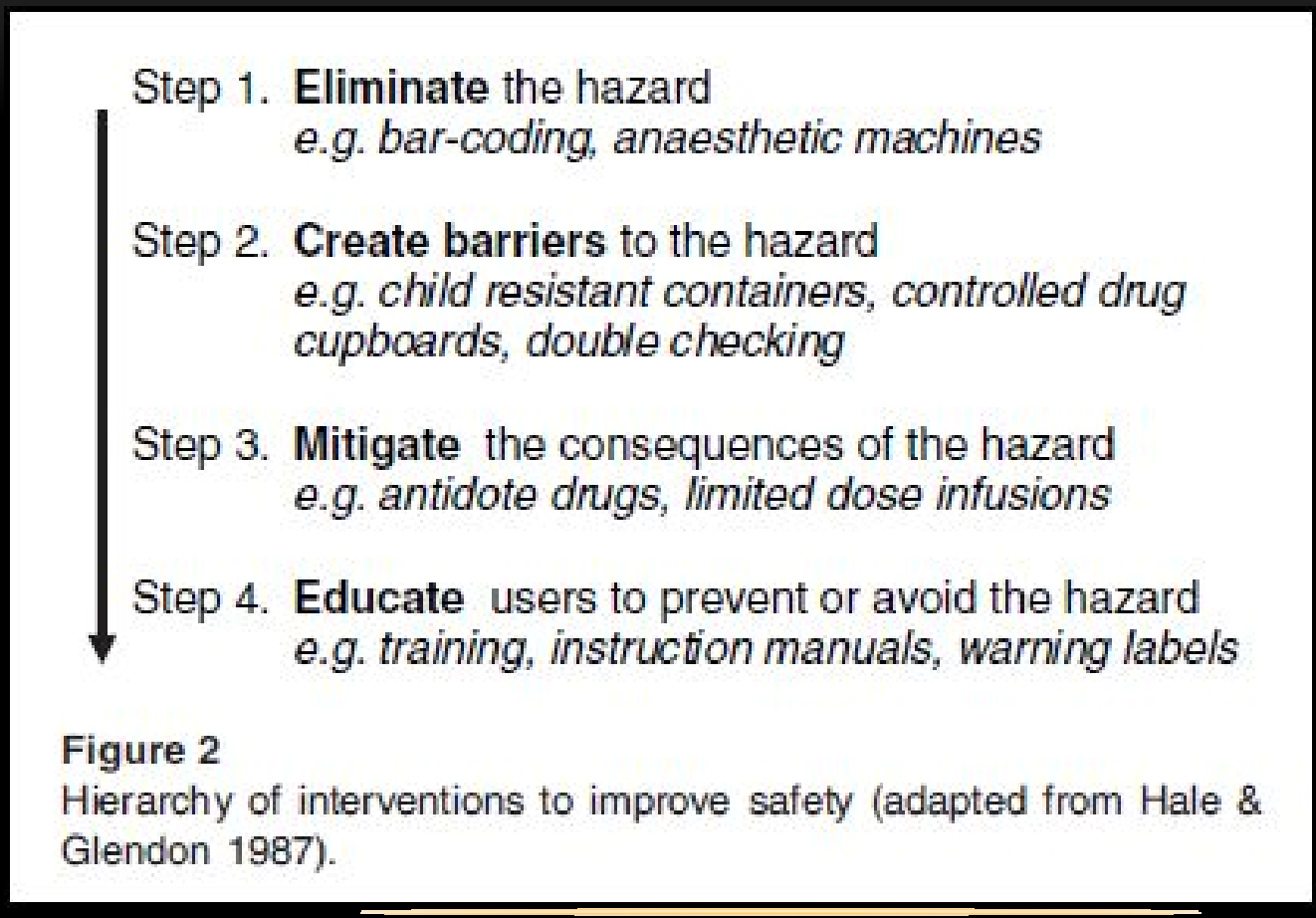


Figure 2
Hierarchy of interventions to improve safety (adapted from Hale & Glendon 1987).

Norris, 2009, Human factors and safe patient care. Journal of Nursing Managementt ology, 26(8), 889–896. doi:10.1038/nbt.1411

MORE “GARBAGE IN / GARBAGE OUT ” EVERY DAY…

- “Have to show the data to change behavior!” ~ Herbek, G

BIOSPECIMEN RESEARCH DATABASE

http://biospecimens.cancer.gov/brd/

1150 articles and growing ~ Bass, P

Features / annotations abstracted

Add level of evidence (# of patients, # of samples)?

Find related articles for animal biospecimens?

CAN WE USE TEXT MINING TO ASSESS REPORTING, AND CREATE BARRIERS TO INADEQUATE REPORTING?

- Document classification
  - Classify existing articles according analyte, and technology platform and other BRD labels / categories
  - Classify new manuscripts according to BRD labels
- Information extraction
  - Named entity recognition of BRISQ variables
  - Describe relations between entities (more complex features)
- Complementary to synoptic reporting (Herbeck, G) and detailed prospective data capture (e.g. Moore, H with OpenClinica)

LABELS / ANNOTATIONS ABSTRACTED FROM ARTICLES AND ENTERED INTO BRD…

Label Field	Label Value
Biospecimen Type	Tissue, Blood... [e.g. dash et al 12414521]
Diagnosis	Neoplastic - Carcinoma
Biospecimen Location	Prostate
Preservation Type	OCT
Analyte	RNA
Technology Platform	DNA Microarray
Experimental Factor	Biospecimen Acquisition, Cold Ischemia Time
Experimental Factor	DNA Microarray Specific, Targeted nucleic acid
Summary of Findings	Using microarray, 0.6% of genes (61 genes, 41 of which were named) were upregulated in prostate tissues stored at room temperature for 1 h or longer compared to 0 h controls. Genes displaying elevated expression included several early response genes (early growth response 1, EGR-1; jun B proto oncogene, junB; jun D proto oncogene, jun D; activating transcription factor 3, ATF3). The degree of upregulation was variable and ranged between minute increases to nearly 2-fold upregulation (EGR-1, junB).

CAPTURE ANNOTATION AT THE SOURCE

biobanks defined the time of excision as the time of removal of the biospecimen from the operating table as documented in the operating room by a nurse. The time of cryopreservation was defined as the time that the biospecimen was stored at −70°C or in a liquid nitrogen vapor freezer.

Queries of the databases at MBTB and the BCCA-TTR were done by each biobank's informatics staff. These queries returned the excision time and cryopreservation time for each breast tumor biospecimen. Biospecimens for which either of these times were unknown were excluded from these analyses. Data were sorted into four groups: <30 min, 31 to 60 min, 61 to 120 min, and >120 min between excision and cryopreservation. These data were graphed using Microsoft Excel (Microsoft Corpo-

Barnes RO, Parisien M, Murphy LC, Watson PH. Influence of evolution in tumor biobanking on the interpretation of translational research. Cancer Epidemiol Biomarkers Prev 2008 Dec.;17(12):3344–3350.

PREVIOUS WORK IN TEXT MINING

Pathology reports – caTIES and CAP protocols

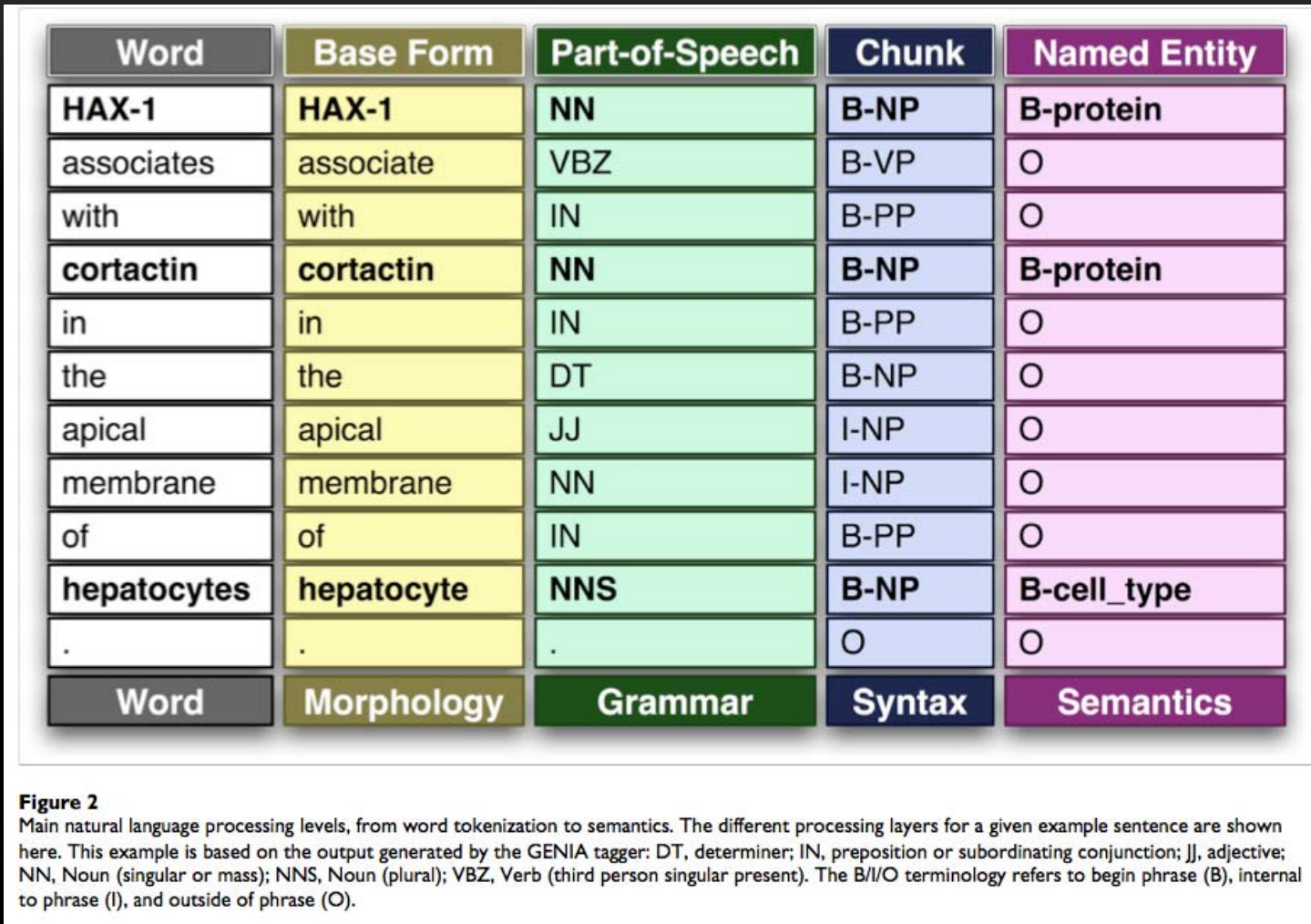
Extraction of experiment results

- Protein-protein interactions
- Genome annotation
- Relation extractions (facts or events)
  - Phenotypes
  - Species
  - Pathology

PubMed abstracts, PMC full-text, and supplementary online materials

Krallinger M, Valencia A, Hirschman L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol 2008;9 Suppl 2:S8.
Haeussler M, Gerner M, Bergman CM. Annotating genes and genomes with DNA sequences extracted from biomedical articles. Bioinformatics 2011 Apr.;27(7):980–986.

TEXT MINING FIRST STEPS FOR THIS DOMAIN



Word	Base Form	Part-of-Speech	Chunk	Named Entity
HAX-1	**HAX-1**	**NN**	**B-NP**	**B-protein**
associates	associate	VBZ	B-VP	O
with	with	IN	B-PP	O
cortactin	**cortactin**	**NN**	**B-NP**	**B-protein**
in	in	IN	B-PP	O
the	the	DT	B-NP	O
apical	apical	JJ	I-NP	O
membrane	membrane	NN	I-NP	O
of	of	IN	B-PP	O
hepatocytes	**hepatocyte**	**NNS**	**B-NP**	**B-cell_type**
.	.	.	O	O
Word	**Morphology**	**Grammar**	**Syntax**	**Semantics**

Figure 2
Main natural language processing levels, from word tokenization to semantics. The different processing layers for a given example sentence are shown here. This example is based on the output generated by the GENIA tagger: DT, determiner; IN, preposition or subordinating conjunction; JJ, adjective; NN, Noun (singular or mass); NNS, Noun (plural); VBZ, Verb (third person singular present). The B/I/O terminology refers to begin phrase (B), internal to phrase (I), and outside of phrase (O).

Krallinger, M., Valencia, A., & Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome biology*, *9 Suppl 2*, S8. doi:10.1186/gb-2008-9-s2-s8

EXPLORING THE VALUE PROPOSITION

- Market based research on text mining for biospecimen information
  - Journal publishers and editors
  - Investigators using human biospecimens
  - Funding of research with biospecimens
- More effective methods to improve practices
  - Education and training is a great first step, but may not sufficient to change individual practices
  - Accreditation (e.g. CAP) based on evaluation is key
  - Structured data capture is expensive and puts burden on pathology, clinical staff, etc; NLP can help

HOW DO WE GET THIS GOING?

- Build annotation of “corpus” and text mining into pipeline for acquiring, curating and evaluating biospecimen information from literature
- Build text mining support for data abstraction and data entry, to increase performance or reduce costs over time
- Explore potentially fundable text mining / BioNLP research for this new area

CONCLUSIONS

- Text mining can potentially be used to extract and evaluate biospecimen information from literature
- May be a tool to measure and improve practices (i.e. adherence to SOPs and reporting guidelines)
- Implementing text mining may require shifts in data abstraction/entry practices

ACKNOWLEDGEMENTS

- Peter Tarczy-Hornoch, MD
- Meliha Yetisgen-Yildiz, PhD
- John Castle, ScD
- Helen Moore, PhD
- PNW Prostate Cancer SPORE
- Annika Havnaer
- Kelly Engel, PhD