BMC Medical Informatics and Decision Making

BioMed Central

Research article

Classifying the precancers: A metadata approach

Jules J Berman*[1] and Donald E Henson[2]

Address: [1]Cancer Diagnosis Program, National Cancer Institute, NIH, Rockville, Maryland, USA and [2]Department of Pathology, The George Washington University Medical Center, Washington, D.C, USA

Email: Jules J Berman* - bermanj@mail.nih.gov; Donald E Henson - hensond@comcast.net

* Corresponding author

Published: 20 June 2003

BMC Medical Informatics and Decision Making 2003, **3**:8

Received: 31 January 2003
Accepted: 20 June 2003

This article is available from: http://www.biomedcentral.com/1472-6947/3/8

Abstract

Background: During carcinogenesis, precancers are the morphologically identifiable lesions that precede invasive cancers. In theory, the successful treatment of precancers would result in the eradication of most human cancers. Despite the importance of these lesions, there has been no effort to list and classify all of the precancers. The purpose of this study is to describe the first comprehensive taxonomy and classification of the precancers. As a novel approach to disease classification, terms and classes were annotated with metadata (data that describes the data) so that the classification could be used to link precancer terms to data elements in other biological databases.

Methods: Terms in the UMLS (Unified Medical Language System) related to precancers were extracted. Extracted terms were reviewed and additional terms added. Each precancer was assigned one of six general classes. The entire classification was assembled as an XML (eXtensible Mark-up Language) file. A Perl script converted the XML file into a browser-viewable HTML (HyperText Mark-up Language) file.

Results: The classification contained 4700 precancer terms, 568 distinct precancer concepts and six precancer classes: 1) Acquired microscopic precancers; 2) acquired large lesions with microscopic atypia; 3) Precursor lesions occurring with inherited hyperplastic syndromes that progress to cancer; 4) Acquired diffuse hyperplasias and diffuse metaplasias; 5) Currently unclassified entities; and 6) Superclass and modifiers.

Conclusion: This work represents the first attempt to create a comprehensive listing of the precancers, the first attempt to classify precancers by their biological properties and the first attempt to create a pathologic classification of precancers using standard metadata (XML). The classification is placed in the public domain, and comment is invited by the authors, who are prepared to curate and modify the classification.

Background

Premalignant lesions are arguably the most important disease entities of modern man. In theory, the identification and elimination of cancer precursors would lead to the near-eradication of cancer [1]. The importance of the precancers was recently emphasized by the American Association for Cancer Research Task Force on the Treatment and Prevention of Intraepithelial Neoplasia [2]. In this report, the Task Force recognized IEN [intraepithelial neoplasia] as a near-obligate precursor to invasive cancer and identified IEN as a treatable disease. "Reducing IEN burden, therefore, is an important and suitable goal for

medical (noninvasive) intervention to reduce invasive cancer risk and to reduce surgical morbidity. Achieving the prevention and regression of IEN confers and constitutes benefit to subjects and, in the opinion of this Task Force, demonstrates effectiveness of a new treatment agent."

In February, 2001, the NCI sponsored a workshop on precancer classification [3]. The task force concluded that "there has been a lack of uniform terminology for the precancerous and noninvasive lesions."and recommended that, "because of the consistent lack of a common diagnostic terminology, which is a major impediment to classification, agreement on the terminology and criteria for the precancerous lesions in all major sites should be sought."

Clinical Importance of a Precancer Classification

The best medical discoveries are generalizable. For example, if antibiotics were only effective on a single bacteria, Alexander Fleming's chance discovery would have had limited medical value. Bacteria have common biological properties (e.g. a cell wall, small size, etc.) that make them different from flowers, insects and people. Knowledge of the general properties of bacteria can inspire therapeutic strategies that extend its value to all the members of its class. Bacteriologists learn the names of all the different bacteria and group the different bacteria based on shared properties. Bacteriologists use their knowledge of bacterial classes to develop new antibiotics and new therapeutic strategies.

Classification efforts typically begin by listing all the members of the classification domain (i.e. creating a taxonomy). Until now, there has been no effort to list the precancers and or to associate precancer terms with their synonyms. Any given precancer may have been studied by different researchers using different terms for the same lesion. Absent a definitive terminology, distinct lesions may have been studied under the same name. The absence of a comprehensive precancer terminology severely limits the clinical value of research that includes precancer specimens.

Until now, there has been no effort to group the precancers by shared clinical, morphologic or biomolecular features. If an agent were discovered that induced regression of a particular precancer, there would be no organized precancer classification prompting anyone to select biologically related lesions likely to respond to the same agent.

It is the opinion of the authors that precancers should have a biological classification. Database annotations using the precancer classification will provide a mechanism whereby each precancer, its' related precancerous lesions, and the cancers known to develop from these lesions, can be linked with relevant data contained in biological data sets (e.g. gene expression arrays and proteomics arrays, tissue microarrays, pathology data sets).

Informatics Aspects of Classification

Modern classifications serve as informatics devices capable of linking, integrating and retrieving information contained in diverse biological data sets. Creators of biomedical databases use terminologies to annotate individual data elements. Data annotation involves appending descriptive information to experimental data with the intention of creating an encapsulated data object that can be intelligently connected to related data in other databases. Annotations are a critical part of the data, because the annotations assist in the discovery of the biological relevance of the data element. Lesions annotated with terminology from the same classification can be linked, even when they occur in heterogeneous data sets.

In the last several years, a new type of data annotation model has emerged that will greatly enhance the research value of publicly available datasets. This model is the self-describing data collection. In this model, all of the data fields are tagged by metadata. Metadata is data that describes the data elements. XML (Extensible Mark-up Language) is the most popular format for metadata annotation [4]. The precancer classification is embodied in an XML file. The data elements of the classification are instances of the precancer domain and the properties of the domain instances. Examples of precancer data elements are values such as "DCIS" and "Actinic keratosis" and "000932" and "ENG". The metadata are the flanking tags such as <concept>, <synonym>, <cui>, and <language>. The Precancer XML file itself is annotated with sufficient information that anyone opening the file can fully understand the file contents.

Now and in the future, researchers will need to organize their data in a way that permits thoughtful analysis. Laboratories will be producing terabytes of data, in the form of images, tissue micoarrays, gene expression arrays and proteomic arrays. None of this data will have any value unless it is described in a standard manner that computers can understand. Classifications exist to organize the instances of a domain so that information assembled from databases can be generalized or related to defined taxonomic groups.

Definition of Precancer and Common Terms in Use

During carcinogenesis, morphologically identifiable lesions occur that precede the development of invasive cancer. These lesions are called precancers, premalignancies, preneoplastic lesions, incipient cancers,

intraepithelial neoplasias, and preinvasive cancers. The plethora of terms reflects the difficulty of choosing a "best" canonical class term for the precancerous lesions. Currently, the term "intraepithelial neoplasia" seems to enjoy wide usage among the community of pathologists, but this term has limitations:

1. Not all epithelial precancers are intraepithelial. Most of the mucosal dysplasias have a well-defined territory bounded by the junction between the epithelium and the underlying stroma. But not all premalignant epithelial lesions can be identified by the presence of atypical cell populations delimited by a basement membrane. Dysplastic lesions of the liver, kidney, thyroid and adrenal are not delimited by a basement membrane [1].

2. Not all precancers are epithelial [1]. Intratubular germ cell neoplasms of testis, myelodysplasias, and non-autonomous lymphomas are examples of non-epithelial precancers.

3. Not all intraepithelial neoplasms are precancers. Neoplasms that are intraepithelial but that are not precancers include: seborrheic keratoses, intraepidermal nevi, common warts and most so-called benign epithelial tumors.

Likewise, the term pre-invasive cancer raises an existential question. Use of the term "pre-invasive cancer" implies that precancers have attained the biological properties of a cancer. This assumption may not be true. Precancers may lack constitutive properties of cancer or may have certain attributes that are absent in cancers. At this point, there is insufficient knowledge to conclude that precancers are types of cancer. In this article, the authors use the term precancers because this term conveys only the defining features: occurrence prior to cancers, and existence as an identifiable lesion.

When considering all the possible classes of precancers, it is worth noting that:

1. Not all precancers are neoplastic. A diffusely hyperplastic lesion with no known neoplastic properties, but with a frequent association with cancer arising from the hyperplastic tissue, would be considered a precancer. Examples include diffuse atypical endometrial hyperplasia, AIDS-associated lymphoid hyperplasia, helicobacter-associated gastric MALT hyperplasia, diffuse gastric intestinal metaplasia, etc.

2. Precancers need not progress to cancer and often have a high rate of regression [5,6]. The low-risk of progression to cancer suggests a strategy for treatment based on enhancing the intrinsic regression rate of precancers [6]. However, when a precancer progresses, cancer is the obligate outcome (i.e. precancers never progress into types of lesions other than cancer). This biological property allows us to infer that agents that induce precancers are carcinogens.

3. The different kinds of precancers may vary in every biologic feature except those specified in their definition (identifiable lesions that precede the development of cancer). Since precancers, by definition, are the morphologic lesions that precede cancers, one can expect precancers to occur in a somewhat younger population than the population of people who have cancers. Using the same line of reasoning, one can expect agents (chemical or biological) that induce precancers to also induce cancers.

The biological diversity of precancers

When the different precancers are listed, it becomes apparent that they fall into very different biological classes. Consider the following three lesions, all of which are usually considered to be precancers:

1. Squamous dysplasia of the uterine cervix. Squamous dysplasias are microscopic foci of atypical squamous cells. They are not tumors in the sense that they do not present as a growing mass. In the cervix, they are almost always associated with a viral etiology.

2. Tubular adenoma of colon. Tubular adenomas are benign tumors that can measure several centimeters in diameter. Nuclear atypia can be minimal or marked.

3. Barrett's esophagus. This is a glandular metaplasia occurring in the esophageal mucosa caused by local chronic inflammation. These lesions typically show no nuclear atypia. They are associated with an increased risk of adenocarcinoma of the esophagus.

What do these lesions have in common? They are identifiable lesions that can precede the development of cancer. Other than that, they would seem to have very few features in common. The diversity of biological types of precancers calls for the creation of a precancer classification.

Creating the Precancer Classification

A classification is a hierarchy of taxa (informative features that characterize an entity and distinguish it from other entities) and a set of generalizable features that apply to groups of taxa. For instance, if "chair" is classified under "furniture," then we can expect that all of the generalizations that we can form on the topic of furniture will apply to chairs. This is actually a remarkable concept as it allows us to apply general knowledge to specific items. Even if I know nothing about chairs, knowing that a chair is a type of furniture allows me to infer many things about chairs based on my general knowledge of furniture. If furniture

is something that belongs in a house, then a chair belongs in a house.

The task of classification usually begins by listing every member of a domain (in this case, every precancer) and then choosing groups that carry the greatest number of informative biological generalizations to every member of the group. The list of every member of a domain is called a taxonomy. A classification is a grouped taxonomy [7].

The process of classifying lesions is different from the process of identifying lesions. Identification involves assigning a name [from an existing classification] to a lesion. The distinction between classification and identification is of great importance, because classification schemes, unlike identification schemes, have properties that can be of immense importance in medical research [7].

1. A classification contains every instance in its domain, and every instance has a single and unique slot in the classification. This facilitates the design of experiments that include every instance of related lesions.

2. Good classifications contain classes carefully selected to have the maximally informative set of generalizable features common to all the class instances. Having a classification allows us to compare two different instances based on the inherited properities of their classes.

3. Classifications support annotation using the elements of the classification (classes and instances) as keys. Database annotation can be used to improve the classification system.

Classifications are also different from ontologies. Ontologies create logical rules between specified members of a group [8]. Ontologies are expected to have "competence", the ability to respond to queries that draw from the formal relationships among group members. A classification approach was chosen, because the two primary goals of this effort were to create a comprehensive listing of precancer terms and to provide a broad hierarchy for precancer concepts. Establishing a set of logical rules for the precancers is, at this time, not feasible.

What precancer classification is currently available?

ICD-O (International Classification of Diseases – Oncology) lists virtually no precancerous lesions [9]. The ICD-O is used by cancer registrars to annotate, in a uniform way, cases of treated malignancies. With a few exceptions (such as ductal carcinoma in situ of breast) the cancer registries do not collect information on precancers. The National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) and the Centers for Disease Control and Prevention use data collected by the cancer registries to compile national statistics on the incidence of cancer. The most comprehensive summary of the clinical and pathologic features of precancers is found in "Pathology of Incipient Neoplasia", edited by Henson and Alberes-Saavedra [10]. None of these sources provides a classification of the precancers.

Precancer Classification

The following six classes were used: 1) Acquired microscopic precancers; 2) acquired large lesions with microscopic atypia; 3) Precursor lesions occurring with inherited hyperplastic syndromes that progress to cancer; 4) Acquired diffuse hyperplasias and diffuse metaplasias; 5) Currently unclassified entities; and 6) Superclass and modifiers.

1. Acquired microscopic precancers

These are the lesions that most people think of when they hear the term precancer. All of the so-called intraepithelial neoplasias fall into this category. Most examples of the microscopic precancers occur commonly (actinic keratosis, cervical dysplasia). They tend to be multifocal. They tend to be non-inherited lesions, often with an identifiable causation (e.g. sunlight, human papillomavirus infection). They seldom occur in children. Exceptions are inherited diseases that heighten sensitivity to a causal agent, such as the early appearance of actinic keratoses in children with Xeroderma Pigmentosum. Morphologically, they tend to have a high degree of nuclear atypia. The microscopic epithelial precancers grow by a subtle replacement of the normal mucosa, without producing a mass, despite many replicative cycles of growth. They progress to invasive cancer while still relatively small. The term dysplasia is often applied to these lesions. Dysplasia, in the context of precancer, is somatically inherited nuclear atypia. Cytologists use the morphologic features of dysplasia to identify precancer cells. Class 1 precancers often have an identifiable non-dysplastic stage that precedes the appearance of nuclear atypia (e.g. squamous metaplasia of bronchus, Barrett's esophagus without atypia, junctional nevus, intestinal metaplasia of stomach)

2. Acquired large lesions with morphologic atypia

These lesions tend to have a uniform appearance throughout most of their long existence, even from the smallest size (i.e. they have a long, stable growth phase). They tend not to have precursor lesions from identifiable microscopic precancers (e.g. class 2 lesions do not seem to arise from class 1 lesions). Their chance of becoming malignant usually increases as the size of the lesion increases. When they become malignant, there is usually a morphologically apparent focus from within the large lesion that has crowding, irregular growth pattern and marked cellular atypia that is strikingly different from the surrounding

cells. This focus enlarges, shows frank invasion, and is the presumed origin of the cancer that develops from the precancer. These lesions tend not to regress spontaneously. They tend to be long-lived and do not progress to cancer without first growing to a large size. These lesions are often multiple but do not occur in large numbers (hundreds) unless there is a germline mutation. Prototypical acquired large precancers are colon adenoma and myelodysplasia

3. Precursor lesions occurring with inherited hyperplastic syndromes that often progress to cancer

These lesions tend to occur very rarely in the general population, but may occur with a high probability (sometimes 100%) in patients carrying the germline mutation. The prototypical lesions are the Ret-gene disorders. Mutations in the RET gene are associated with the disorders multiple endocrine neoplasia, type IIA (MEN2A), multiple endocrine neoplasia, type IIB (MEN2B), and hereditary medullary thyroid carcinoma.

Lesions in this general category tend to have a single gene mutation that may be the only lesion found in the precursor lesions. The precuror lesions tend to have the morphology of simple hyperplasias, without much nuclear atypia. Precursor lesions tend to be multiple, sometimes occurring in the hundreds, and bilateral in paired organs. These lesions tend to occur in a much younger population than the acquired precancers. The resulting cancers can also occur at a relatively young age.

4. Acquired diffuse hyperplasias and diffuse metaplasias

With few exceptions, acquired small focal metaplasias and hyperplasias have a very low chance of progression to cancer, and have been excluded from the classification schema because they rarely result in cancer without first growing into diffuse lesions (the class 4 lesions) or acquiring nuclear atypia (class 1 lesions).

Diffuse metaplastic lesions commonly precede cancers. It is presumed that all bronchogenic squamous dysplasia arises from squamous metaplasia. The normal bronchus simply does not have any squamous cells. The squamous cells in bronchial squamous dysplasia must have originated from a metaplastic focus for directly from non-squamous bronchial cells that differentiated directly to a dysplastic squamous phenotype.

The prototypical lesions are the diffuse Barrett's esophagus, diffuse intestinal metaplasia of stomach, and diffuse endometrial hyperplasia. These lesions tend to have chronic identifiable causes (e.g. gastroesophageal reflux disease, post lye ingestion esophagus, chronic gastritis, long-term tamoxifen therapy), and tend not to regress so long as the causation persists. Small foci of dysplastic precancers (Class 1) may arise from the diffuse hyperplasias and metaplasias.

This class of precursor may include the so-called regressing cancers, such as helicobacter-associated maltomas and AIDS-associated Kaposi's sarcoma that can grow as multiple tumors, all of which can quickly regress when the causative agent is withdrawn (e.g. after antibiotic treatment for Helicobacter or after normal immune status is restored after withdrawal of cyclosporine in transplant recipients). This class may also include secondary aplastic anemia (e.g. benzene toxicity), where the marrow is repopulated by an emerging population of hyperplastic cells that carry a a heightened risk of progressing to acute leukemia.

5. Currently unclassified entities

Most precancers will fall into one of the first four described classes. However, classifications may contain a subset of cases that defy facile classification. For example, the platypus has challenged animal classifiers. Aristotle had no trouble recognizing that dolphins were mammals, but it took the scientific community two millennia to agree.

We have created an "unclassified" category of precancers for the current draft classification

6. Superclass and modifiers

A superclass is created to contain general precancer terms (e.g. precancer, dysplasia)

Methods

The National Library of Medicine's UMLS (Unified Medical Language System) is a set of tools that facilitate the use of medical terminologies and the semantic relationships between terms and vocabularies. The UMLS Metathesaurus is one of three knowledge sources within the UMLS and contains concepts and terms from about 100 different medical vocabularies. The primary UMLS metathesaurus file used in the construction of the precancer terminology is MRCON. The 2003 MRCON file is over 150 Mbytes in length and contains over two million different terms belonging to nearly a million different concepts. MRCON and the entire UMLS metathesaurus are available at no cost from the National Libary of Medicine at:

http://www.nlm.nih.gov/research/umls/

An example of some records from the MRCON file is shown:

C0004763|ENG|P|L0004763|VO|S1397347|Barretts Esophagus|0|

C0004763|ENG|P|L0004763|VO|S1397348|Esophagus, Barrett's|0|

C0004763|ENG|P|L0004763|VO|S1459012|barrett's esophagus|2|

C0004763|ENG|P|L0004763|VO|S1940341|Barretts esophagus|0|

C0004763|ENG|P|L0004763|VW|S0038968|Esophagus, Barrett|0|

C0004763|ENG|S|L0292386|PF|S0364034|Columnar-lined esophagus|3|

C0004763|ENG|S|L0292386|VCW|S0369397|ESOPHAGUS, COLUMNAR-LINED|0|

C0004763|ENG|S|L0292386|VO|S0842892|Columnar-lined oesophagus|3|

Notice that numerous variant terms for Barrett's esophagus all map to the same number in the first column, C0004763. This is the UMLS CUI (Concept Unique Identifier) for Barrett's esophagus.

The authors collected precancer terms from the UMLS. After review of the terms, the authors added supplemental terms from their own knowledge. Every additional term added by the authors matched a pre-existing UMLS concept. About 10% of the precancer synonyms were contributed by the authors.

After the terminology was assembled, the authors created a classification system and assigned each precancer term to one of the precancer classes. The entire classification was prepared as a metadata document using XML (eXtensible Markup Language) annotation.

Results

The XML document containing the precancer classification and all accompanying metadata is PRESUM.XML (425 kilobytes) [see Additional file: 3]. An example of the metadata annotation in the XML file is:

<concept><cui>0432487

</cui> <precancer_class>Acquired diffuse hyperplasias/metaplasias </precancer_class><synonym>

<term> Post-transplant lymphoproliferative disorder </term> </synonym>

<synonym> <term> PTLD – Post-transplant lymphoproliferative disorder </term> </synonym>

<synonym> <term> Post-transp lymphoprolif dis </term> </synonym>

<synonym> <term> PTLD-Post-trns lymphprolif dis </term> </synonym>

<synonym> <term> PTLD </term> </synonym>

<synonym> <term> PTLPD </term> </synonym>

<synonym> <term> PT-LPD </term> </synonym>

<synonym> <term> post transplantation lymphoproliferative disorders </term> </synonym>

<synonym> <term> post transplantation lymphoproliferative disorder </term> </synonym>

<synonym> <term> post transplantation lymphoproliferative disease </term> </synonym> </concept>

In the example, five metadata tags are employed: <concept>, <cui>, <precancer_class>, <term>, <synonym>, along with and their corresponding closure tags (marked by a slash character). Because XML is case-sensitive, lowercase letters were consistently employed to simplify implementation. The <concept> tag indicates that a new concept will follow. Since all of the precancer concepts derive from or correspond to existing UMLS concepts, it was convenient to assign each precancer concept with the UMLS Concept Unique Identifier and mark these with a <cui> tag. Each precancer concept is assigned one of the precancer classes. In this case, the term "PTLD, post-transplant lymphoproliferative disorder" is assigned to the precancer class of "Acquired diffuse hyperplasias/metaplasias." The term is flanked by <term> tags and the class designation is flanked by <precancer_class> tags. Because the term is a synonymous variant, it is nested in <synonym> tags. Term and synonym tags are used for each of the term variants of the single concept.

Raw XML files are made difficult to read by the large quantity of markup (XML tags) that annotate data elements. Typically, XML files are made readable with transformation scripts or with embedded presentation instructions (cascading style sheets) [11]. A Perl script (PRESUM.PL) was created to parse the XML file, counting the classified lesions and outputting a viewable HTML file (PRESUM2.HTM) and a summary statement as follows:

The total number of precancer terms => 4700

The total number of precancer concepts => 568

(a single concept has, on average, 8.3 near-synonymous terms)

Number of concepts falling in the six precancer classes:

Acquired microscopic precancers => 262

Acquired diffuse hyperplasias/metaplasias => 37

Acquired large lesions with cancerous potential => 110

Precancers in Inherited hyperplastic syndromes => 25

Precancer superclass and associated modifiers => 42

Unclassified => 92

Number of Non-English terms

Russian => 184

German => 320

Finnish => 59

Italian => 63

Portuguese => 188

French => 145

Spanish => 206

The viewable html file is provided as a supplemental file [see Additional file: 1] with this article.

The most current versions of the Precancer Classification XML and the transforming Perl script are avaliable from the following URL:

http://65.222.228.150/jjb/presum.tar.gz

Discussion

The different precancers have only their definition in common: they are the morphologically distinctive lesions that precede the development of cancer. Beyond that, precancers can differ from one another by almost every conceivable property. A microscopic cervical dysplasia seems to be fundamentally different from a RAEB (refractory anemia with excess blasts). However, both lesions are considered cancer precursors. A single tissue may have fundamentally different precancer entities, all preceding the same type of cancer. A bowel diffusely involved by ulcerative colitis, an aberrant crypt, and a colon adenoma are seemingly disparate lesions. But they all precede the development of colon carcinoma. In classifying the precancers, the authors tried to distinguish the precancers based on the intrinsic properties of lesions. The exercise of creating a comprehensive classification brought forth a variety of issues.

Limitations of the Precancer Classification

The history of the classification of living organisms runs through thousands of years and numerous revisions. In a recent editorial by Thiele and Yates, the authors observed that taxonomy projects are often bypassed by funding agencies that prefer high profile experimental efforts [12]. Stephen Gould has commented that taxonomy is portrayed as the dullest of all fields, "But classifications are not passive ordering devices in a world objectively divided into obvious categories. Taxonomies are human decisions imposed upon nature – theories about the causes of nature's order. The chronicle of historical changes in classification provides our finest insight into conceptual revolutions in human thought. Objective nature does exist, but we can converse with her only through the structure of our taxonomic systems" [13].

Classifications are suggested by individuals, subject to modification by peers. Several issues, in particular, require community review.

Issues of inclusion

The authors chose to err on the side of inclusion when developing the taxonomy. If a lesion was considered a putative precancer (even when the evidence seemed doubtful), it was added to the taxonomy.

Issues of exclusion

How does the classification deal with conditions associated with cancer but for which no precancerous lesion is known? These conditions are often called cancer syndromes. Persons identified (possibly through genetic testing) with a cancer syndrome who have not yet developed cancers may be considered to be in a precancerous stage of their disease. In the absence of morphologically identifiable precancerous lesions, these conditions were excluded from the classification. Since these syndromes may have enormous relevance to our understanding of the carcinogenic process, they were collected as a separate listing. When the precancer classification undergoes community review, these syndromes may be added as a distinct class. It is available as a supplemental file with this publication [see Additional file: 2].

Issues of unclassifiability

Not all precancerous lesions fit into a biological group. In most cases, the unclassifiable lesions are concept "placeholders", such as "atypical squamous cells of undetermined significance."

Table 1: Class Examples

Acquired Small or Microscopic Precancers
HGSIL (High grade squamous intraepithelial lesion of uterine cervix)
AIN (Anal intraepithelial neoplasia)
Dysplasia of vocal cord
Aberrant crypts (of colon)
PIN (prostatic intraepithelial neoplasia)
Acquired Large Lesions with Nuclear Atypia
Tubular Adenoma
AILD (angioimmunoblastic lymphadenopathy with dysproteinemia)
Atypical meningioma
Gastric Polyp
Large plaque parapsoriasis
Myelodysplasia
Papillary transitional cell carcinoma in situ
Refractory Anemia with Excess Blasts
Schneiderian papilloma
Precursor lesions occurring with inherited hyperplastic syndromes that progress to cancer
Atypical mole syndrome
C cell adenomatosis
MEA
Acquired diffuse hyperplasias and diffuse metaplasias
AIDS
Atypical lymphoid hyperplasia
Paget's Disease of Bone
Post-transplant lymphoproliferative disease
Ulcerative colitis
Superclass and modifiers
ATYPIA
Atypical cell
Atypical hyperplasia
Atypical regeneration
Dysplasia
Dysplastic
In situ cancer
Inflammatory atypia
Mild dysplasia
Premalignant
Preneoplastic state

Issues of omission

Because the classification was completed by two authors, it is presumed that researchers in the field of precancers will wish to add lesions to the taxonomy. This may be particularly important for veterinary and comparative pathologists, as the current classification is heavily weighted toward human lesions.

Issues of incorrect classification

Classifications are hypotheses about the nature of their subject domain. A taxonomist needs to place every known instance (precancer, in this case) somewhere in the classification. Once this is done, the classification can be tested and re-organized.

Conclusions

This work represents the first attempt to create a comprehensive listing of the precancers, the first attempt to classify precancers by their biological properties and the first attempt to create a pathologic classification of the precancers that recognizes fundamental biologic and morphologic distinctions (taxons) among the precancers. A draft classification, placed into the public domain, is a first step toward a clinically useful classification of the precancers. The metadata format (XML) provides researchers with access to a comprehensive, organised listing that can be used to annotate and link precancer lesions contained in biomedical data sets. Public comment is welcomed.

Competing interests

None declared.

Authors' contributions

Dr. Berman wrote the first drafts of the paper and the classification. Dr. Berman designed the XML metadata and wrote the Perl script that transforms the XML file to HTML. Dr. Henson organized the NIH workshop that developed the conceptual framework for the classification. Dr. Henson edited multiple drafts of the article and the classification.

Additional material

Additional File 1

Presum.htm is the classification file for the precancers, in viewable HTML format.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1472-6947-3-8-S1.htm]

Additional File 2

Omimprec.htm contains inherited conditions associated with cancer risks, in HTML format. OMIM (online Mendelian Inheritance in Man) numbers are provided for each lesion.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1472-6947-3-8-S2.htm]

Additional File 3

Presum.tar.gz is a tarballed-gzipped compressed file containing presum.xml, the XML data file of the precancer classification, and presum.pl, a Perl script that transforms the XML file into a viewable HTML file, presum2.htm

Click here for file

[http://www.biomedcentral.com/content/supplementary/1472-6947-3-8-S3.gz]

Acknowledgements

The opinions and assertions included in the article and accompanying files are the opinions of the authors and do not represent the policies of the NIH or any other Federal agency. This work was previously presented as an abstract to APIII, Advancing Pathology Informatics, Imaging and the Internet, October 2–4, 2002, in Pittsburgh, PA. The presentation abstract will appear in the Archives of Pathology and Laboratory Medicine in the fall, 2003.

References

1. Seidman JD and Berman JJ: **Premalignant non-epithelial lesions: a biological classification.** *Modern Pathology* 1993, **6:**544-554.
2. O'Shaughnessy JA, Kelloff GJ, Gordon GB, Dannenberg AJ, Hong WK, Fabian CJ, Sigman CC, Bertagnolli MM, Stratton SP and Lam S *et al.*: **Recommendations of the American Association for Cancer Research Task Force on the Treatment and Prevention of Intraepithelial Neoplasia. Treatment and Prevention of Intraepithelial Neoplasia: An Important Target for Accelerated New Agent Development.** *Clin Cancer Res* 2002, **8:**314-346.
3. Henson DE, AlboresSaavedra J, Berman JJ, Chung D, Czerniak B, Franklin WA, Hamilton SR, Hruban RH, Jaffe ES, Stanley E, Shackney SE, Sobin L, Srivastava S, Tavassoli F, Travis W and Wolfe HJ: **Meeting Summary: A Molecular Classification for Precancerous Lesions, Report of an Early Diagnosis Research Network Working Group.** [http://www3.cancer.gov/prevention/cbrg/molclass.html]. February 1–2, 2001
4. **W3C Architecture Domain. Extensible Markup Language (XML)** [http://www.w3c.org/XML/]
5. Foulds L: **Neoplastic Development.** *Academic Press, New York* 1969.
6. Berman JJ and Moore GW: **The role of cell death in the growth of preneoplastic lesions: a Monte Carlo simulation model.** *Cell Proliferation* 1992, **25:**549-557.
7. Mayr E: **The growth of biological thought: diversity, evolution and inheritance.** *Belknap Press, Cambridge* 1982.
8. Ahmed K, Ayers D, Birbeck M, Cousins J, Dodds D, Lubell J, Nic M, Rivers-Moore D, Watt A, Worden R and Wrightson A: **Professional XML Meta Data.** *Wrox, Birmingham* 2001.
9. Fritz A, Percy C, Jack A, Shanmugaratnam K, Sobin L, Parkin DM and Whelan S: **International Classification of Diseases for Oncology (ICD-O).** *World Health Organization, Lyons* Third2001.
10. Henson DE and Albores-Saavedra JA: **Pathology of Incipient Neoplasia.** *Oxford University Press, New York* 2001.
11. White C, Quin L and Burman L: **Mastering XML: Premium Edition.** *Sybex, San Francisco* 2001.
12. Thiele K and Yeates D: **Tension arises from duality at the heart of taxonomy: Names must both represent a volatile hypothesis and provide a key to lasting information.** *Nature* 2002, **419:**337. inclusive
13. Gould SJ: **Full house: The spread of excellence from Plato to Darwin.** *Harmony, New York* 1996.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1472-6947/3/8/prepub