# Consolidated Health Informatics

# Standards Adoption Recommendation

# Genes and Proteins

**Index**

# Summary

## Domain: Genes and Proteins

## Standards Adoption Recommendation:
Human Gene Nomenclature (HUGN) for genes. None for proteins.

**SCOPE**
To allow the federal health care sector to exchange information regarding the role of genes in biomedical research and healthcare, using a single unambiguous genetic nomenclature.

**RECOMMENDATION**
Human Gene Nomenclature (HUGN) sponsored by the Human Genome Organization (HUGO). No recommendation for Protein Nomenclature.

**OWNERSHIP**
HUGO is a non-profit body that is jointly funded by the UK Medical Research Council (40%) and the US National Institutes of Health, contract N01-LM-9-3533 (60%).

**APPROVALS AND ACCREDITATIONS**
-NA-

**ACQUISITION AND COST**
HUGN is free for nonprofit use, but requires a license for commercial use (see http://www.gene.ucl.ac.uk/nomenclature/information/commercial.html )

# Part I – Team & Domain Scope Identification

## Target Vocabulary Domain

***Common name used to describe the clinical/medical domain or messaging standard requirement that has been examined.***

Genes and Proteins

***Describe the specific purpose/primary use of this standard in the federal health care sector (100 words or less)***

To allow the federal health care sector to exchange information regarding the role of genes in biomedical research and healthcare, using a single unambiguous genetic nomenclature. This information would be used to support the federal health care sector in a wide variety of emerging sectors such as pharmacogenomics, genomic medicine, genomic applications of clinical trials, early detection of malignancies, as well as a wide variety of uses in infectious disease such as epidemiology and disease surveillance. As an initial step it is necessary that a genetic nomenclature be adopted that allows the unambiguous assignment of a gene so that this can be correlated to other relevant information including the genes localization. The work group researched the current status of protein nomenclatures, and determined that none were sufficiently mature to warrant adoption at this time. Finally the work group researched the status of vocabularies that might bridge genomics with the sub-domains of Inherited Genetic Variation, Acquired Genetic Changes, and Infectious Disease. Again no vocabularies were found that were sufficiently mature to warrant adoption at this time, however the work group has determined that a focused effort to study the adequacy of other CHI endorsed vocabularies and see if new ones need be developed would greatly accelerate progress in this field. Additional recommendations may be found in the GAPS section.

## Scope

*(Content: Brief description of domain definition to include what is considered in scope and what is considered out of scope. Rationale and issues that were identified by team will be included.)*

| Domain/Sub-domain | In-Scope (Y/N) |
|---|---|
| Inherited Genetic Variation (e.g. Genetic disease, Pharmacogenomics, Disease susceptibility traits) | Y |
| Acquired Genetic Changes (e.g. Cancer) | Y |
| Infectious Disease: Genes/Proteins involved in pathogenesis, drug resistance, or identification | Y |
| Protein Nomenclature | Y |
| Gene Nomenclature | Y |

**Information Exchange Requirements (IERs)** *Using the table at appendix A, list the IERs involved when using this vocabulary.*

| |
|---|
| Customer Health Care Information |
| Care Management Information |
| Customer Risk Factors |
| Case Management Information |
| Body of Health Services Knowledge |
| Clinical Guidelines |
| Population Member Health Data |

**Team Members** *Team members' names and agency names with phone numbers.*

| Name | Agency/Department |
|---|---|
| **James Sorace MD MS (Team Lead)** | CMS |
| Frank Hartel | NCI |
| Sue Dubman | NCI |
| John Leighton Ph.D. | FDA |
| Donna Maglott | NCBI |
| Nancy Orvis | DoD |
| Tim Overman | VA |
| Jean Jenkins | NIH |

**Work Period** *Dates work began/ended.*

| Start | End |
|---|---|
| 9/12/03 | 11/19/03 |

# Part II – Standards Adoption Recommendation

**Recommendation** *Identify the solution recommended.*

Human Gene Nomenclature (HUGN) sponsored by the Human Genome Organization (HUGO).

**Ownership Structure** Describe who "owns" the standard, how it is managed and controlled.

HUGO is a non-profit body that is jointly funded by the UK Medical Research Council (40%) and the US National Institutes of Health, contract N01-LM-9-3533 (60%). It operates through the Chair and a small UK team of professional staff, with key policy advice from an International Advisory Committee (IAC) as well as a team of specialist advisors who provide support on specific gene family nomenclature issues. HUGO maintains and develops the Human Gene Nomenclature (HUGN) that currently has approved symbols for over 17,000 genes, approximately one half of the anticipated total of human genes. Individual new symbols are requested by scientists, journals (e.g. Genomics, Nature Genetics and Cytogenetics and Cell Genetics) and databases (e.g. RefSeq, OMIM, GDB, MGD and LocusLink), and groups of new symbols by those working on gene families, chromosome segments or whole chromosomes. As the human genome sequence analysis nears completion there is an increasing demand for the rapid approval of gene symbols. However, in all cases considerable efforts are made to use a symbol acceptable to workers in the field. HUGO holds regular nomenclature workshops, sometimes in conjunction with larger meetings e.g. American Society of Human Genetics (ASHG) and Human Genome Meeting (HGM). This ensures that names are determined in line with the needs of the scientific community. For details of previous and future workshops see http://www.gene.ucl.ac.uk/nomenclature/workshops.html. HUGN is free for nonprofit use, but requires a license for commercial use (see http://www.gene.ucl.ac.uk/nomenclature/information/commercial.html ).

**Summary Basis for Recommendation** *Summarize the team's basis for making the recommendation (300 words or less).*

HUGN is a recognized standard for human gene nomenclature that has a systematic process for establishing genetic nomenclature. It contains names for approximately one half of the expected number of protein coding human genes* using established criteria (see http://www.gene.ucl.ac.uk/nomenclature/guidelines.html). HUGO has also approached issues regarding non-structural genes. The federal government already extensively utilizes HUGN. For example LocusLink an NCBI resource supports the HUGN as well as Online Mendelian Inheritance in Man (also supported by NIH funding). Thus it is the *de facto* standard for human genomic nomenclature. HUGO works closely with a wide variety of scientific organizations including publishers to assure that its nomenclature is consistently updated. For example, authors may request the assignment of new gene symbol prior to the publication of a manuscript, thus assisting in the establishment of a consistent non-redundant nomenclature. Further, the cross linking of

| |
|---|
| the HUGN with the annotation efforts of NCBI assures that physicians, and scientist can obtain updated information regarding genes, their sequences, and their map positions as well as efforts to map known human genetic variations such as Single Nucleotide Polymorphisms (SNPs). This degree of integration with other genomic resources for human is a primary and unique function of the HUGO Gene Nomenclature Committee (HGNC).<br><br>*Traditionally genes have been defined as stretches of DNA that are transcribed to RNA and then translated to proteins (thus protein coding). Recent scientific advances have also indicated that genes that are transcribed to RNA, but not subsequently translated to proteins, serve important biological functions. Even more recently, highly conserved DNA sequences (e.g. almost identical between humans and mice) with unknown but probably significant biological functions have been defined. Thus the HGNC will not be complete even when it achieves 100% coverage of the portion of genes that code for proteins. |

**Conditional Recommendation** *If this is a conditional recommendation, describe conditions upon which the recommendation is predicated.*

| |
|---|
| There are no conditions. |

**Approvals & Accreditations**

*Indicate the status of various accreditations and approvals*:

| Approvals & Accreditations | Yes/Approved | Applied | Not Approved |
|---|---|---|---|
| Full SDO Ballot | | | |
| ANSI | | | |

**Options Considered** *Inventory solution options considered and summarize the basis for not recommending the alternative(s). SNOME-CT must be specifically discussed.*

| |
|---|
| SNOMED CT® was given consideration, but SNOMED CT® does not specifically address the naming of either genes or proteins in a systematic way. Further recommendations concerning SNOMED CT® may be found in the gaps section. |
| The Gene Ontology (GO) nomenclature was also considered as an alternative. However, this nomenclature is predominantly oriented towards cell biology, and lacks coverage of clinical disease states as well as actual gene names. It may be desirable to reconsider this GO in the context of cell physiology. |

**Current Deployment**

| |
|---|
| HUGN is extensively deployed and it is the de facto standard for the scientific literature. For example NCBI uses HUGN, and only assigns an in house temporary name until an |

official one becomes available. As of 11/7/03 17003 approved gene symbols have been entered in the HUGN database. HUGO is thus approximately halfway through the process of naming the know genes. HUGO is an international organization with headquarters in Britain. The HUGN currently contains approved symbols for over 17,000 genes, approximately one half of the anticipated total of human genes. Individual new symbols are requested by scientists, journals (e.g. Genomics, Nature Genetics and Cytogenetics and Cell Genetics) and databases (e.g. RefSeq, OMIM, GDB, MGD and LocusLink), and groups of new symbols by those working on gene families, chromosome segments or whole chromosomes. As the human genome sequence analysis nears completion there is an increasing demand for the rapid approval of gene symbols. However, in all cases considerable efforts are made to use a symbol acceptable to workers in the field. HUGO holds regular nomenclature workshops, sometimes in conjunction with larger meetings e.g. American Society of Human Genetics (ASHG) and Human Genome Meeting (HGM).

# Part III – Adoption & Deployment Information

*Provide all information gathered in the course of making the recommendation that may assist with adoption of the standard in the federal health care sector. This information will support the work of an implementation team.*

**Existing Need & Use Environment**

*Measure the need for this standard and the extent of existing exchange among federal users. Provide information regarding federal departments and agencies use or non-use of this health information in paper or electronic form, summarize their primary reason for using the information, and indicate if they exchange the information internally or externally with other federal or non-federal entities.*

Column A: Agency or Department Identity (name)
Column B: Use data in this domain today? (Y or N)
Column C: Is use of data a core mission requirement? (Y or N)
Column D: Exchange with others in federal sector now? (Y or N)
Column E: Currently exchange paper or electronic (P, E, B (both), N/Ap)
Column F: Name of paper/electronic vocabulary, if any (name)
Column G: Basis/purposes for data use (research, patient care, benefits)

| Department/Agency | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| Department of Veterans Affairs | | | | | | |
| Department of Defense | | | | | | |
| HHS Office of the Secretary | | | | | | |
| Administration for Children and Families (ACF) | | | | | | |
| Administration on Aging (AOA) | | | | | | |
| Agency for Healthcare Research and Quality (AHRQ) | | | | | | |
| Agency for Toxic Substances and Disease Registry (ATSDR) | | | | | | |
| Centers for Disease Control and Prevention (CDC) | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Centers for Medicare and Medicaid Services (CMS) | | | | | | |
| Food and Drug Administration (FDA) | | | | | | |
| Health Resources and Services Administration (HRSA) | | | | | | |
| Indian Health Service (IHS) | | | | | | |
| National Institutes of Health (NIH) | | | | | | |
| Substance Abuse and Mental Health Services Administration (SAMHSA) | | | | | | |
| Social Security Administration | | | | | | |
| Department of Agriculture | | | | | | |
| State Department | | | | | | |
| US Agency for International Development | | | | | | |
| Justice Department | | | | | | |
| Treasury Department | | | | | | |
| Department of Education | | | | | | |
| General Services Administration | | | | | | |
| Environmental Protection Agency | | | | | | |
| Department of Housing & Urban Development | | | | | | |
| Department of Transportation | | | | | | |
| Homeland Security | | | | | | |

**Number of Terms**

HUGN has approved symbols for over 17,000 genes, approximately one half of the anticipated total of human genes. Individual new symbols are requested by scientists, journals (e.g. Genomics, Nature Genetics and Cytogenetics and Cell Genetics) and databases (e.g. RefSeq, OMIM, GDB, MGD and LocusLink), and groups of new symbols by those working on gene families, chromosome segments or whole chromosomes. As the human genome sequence analysis nears completion there is an increasing demand for the rapid approval of gene symbols. However, in all cases considerable efforts are made to use a symbol acceptable to workers in the field. HUGO holds regular nomenclature workshops, sometimes in conjunction with larger meetings e.g. American Society of Human Genetics (ASHG) and Human Genome Meeting (HGM). Terms are updated daily, with the public FTP site updated twice a week.

**Range of Coverage**

HUGN is currently approximately at the 50 % mark in naming known protein coding genes. It is impossible to forecast exactly when this effort will be concluded. Further the definition of genes may be expanding to include other than the traditionally defined protein coding genes (e.g. RNA genes). Never the less the NCBI as well as the general scientific community have accepted the standard. It is important to recognize that HUGN is a highly focused standard dealing with the nomenclature of human genes. HUGO defines genes as:

*A gene is a DNA segment that contributes to phenotype/function. In the absence of demonstrated function a gene may be characterized by sequence, transcription or homology*.

Thus HUGO uses a deliberately vague set of criteria to define a gene, however this definition is flexible enough to embrace evolving concepts. HUGN is not a nomenclature for annotating genes function or cellular biology. These concepts are however beginning to be addressed by the Gene Ontology consortia (see below). Further HUGN does not cover clinical issues such as naming disease states or describing diseased phenotypes.

The HUGN is limited to covering human gene sequences, and it dose not provide a pathway for pathogen gene nomenclature. The committee recommends additional efforts in this area (see gaps below). Finally HUGN does not name proteins.

**Acquisition:** *How are the data sets/codes acquired and use licensed?*

The HUGN is free via FTP for nonprofit uses. Commercial use requires a license.

**Cost**

The HUGN is free via FTP for nonprofit uses.

**Systems Requirements**

There are no specific systems requirements

**Guidance:**

HUGO maintains the HUGN. HUGO runs a website that allows users to request a gene name, and responds to user questions.

**Maintenance:**

HUGO holds regular nomenclature workshops, sometimes in conjunction with larger meetings e.g. American Society of Human Genetics (ASHG) and Human Genome Meeting (HGM). Terms are updated daily, with the public FTP site updated twice a week.

**Customization**: Not applicable

**Mapping Requirements:**

The HUGN is mapped by NCBI to the physical map of the human genome.

**Compatibility**

Identify the extent of off-the-shelf conformity with other standards and requirements:

| Conformity with other Standards | Yes (100%) | No (0%) | Yes with exception |
|---|---|---|---|
| NEDSS requirements | | | |
| HIPAA standards | | | |
| HL7 2.x | | | |

**Implementation Timeframe**

The HUGN is already widely used within the federal government, and is readily accessible for further implementation.

**Gaps**

While the HUGN represents a useful genetic nomenclature, it will not address every

vocabulary need in this field. In order to analyze these needs in a systematic manner, the group has divided genes and proteins into three domains: inherited phenotypes, acquired genetic mutations, and infectious disease. Human disease states are complex and may frequently involve more than one domain. Realizing that the genomic map and genome nomenclature is a developing basic science standard that will be maintained by NCBI and HUGO, we have focused on healthcare delivery and looked backward towards translational and basic research rather than the opposite approach. Further, genomic medicine will require the use of many other vocabularies that CHI has studied. For example, it may be more advantageous to expand a known list of diseases and clinical conditions to include pharmacogenomic concepts related to drug metabolism, rather than create a domain-specific vocabulary *de novo*. Similarly expressing pharmacogenomic concepts would require a vocabulary for medications. Thus the committee recommends a follow up effort during the next phase of the CHI initiative to determine when existing vocabularies can be expanded to fill this need (and if so how), or when additional vocabularies may need to be developed. Further the committee strongly recommends that NLM and NIH contractors (such as OMIM and GeneTest) be encouraged to both participate in this process and required to implement these vocabularies in a timely manner. The work group has concluded that current efforts appear dispersed and that such follow on action is critical if translational research is to occur in a timely manner. With this background a brief summary of the three domains is presented.

**Inherited Genetic Variation:** This is an extremely important domain, and its clinical uses include genetic diseases as well as pharmacogenomics etc. Currently there are several databases with a substantial user base. Examples include OMIM, GeneTest and HUGE Net. Issues identified in examining these data base efforts include content, vocabulary structure and ability to integrate other standards. OMIM (operating out of Johns Hopkins University under NLM/NIH contract) curates the literature and assigns an OMIM number to specific phenotypes. These OMIM numbers are then linked to a name, a text review of the literature, and in collaboration with the University of Washington, a gene test list that provides information regarding laboratories running relevant assays. OMIM is widely referenced and is a featured link in many NIH related websites. Its summaries are textual and do not use a structured vocabulary. GeneTest allows users to determine were to find the actual testing services for a genetic disease as well as additional information regarding the disease itself. HUGH Net is a CDC initiative that consists of a database of disease susceptibility genes. None of these three databases uses a formally structured vocabulary. Examples for follow on activity in this field include expanding LOINC® codes when necessary to cover genetic test, the use of SNOMED CT® as a source for naming inherited genomic traits including diseases. A careful comparison of disease states and genetic phenotypes found in OMIM/GeneTest with those available in SNOMED CT® and UMLS® would be a substantive contribution on this area. The GeneTest group has developed XML DTDs to better structure the free text found in its database. Extending this approach

and integrating it with structured vocabularies when feasible represents a very useful approach that should be considered.

**Acquired Genetic Changes:** This area focuses on acquired genetic changes that are typically found in malignant disease. Unlike genetic variation above, the group has not found a suitable standard. The NCI Thesaurus is currently in limited use for such purposes within NCI, but it is not ready for wider clinical use. NCI is continuing development and refinement of the Thesaurus. In time NCI expects that the Thesaurus will be a viable candidate standard, and within 6 months expects to have a version of the NCI Thesaurus in production that ought to be reviewed for adequacy as a CHI standard. Currently, there are relatively few acquired changes that are known to be clinically significant. This is an area of translational research, which may increase rapidly in clinical importance.

**Infectious Disease**: The committee wishes to stress the importance of a uniform taxonomy for viral, bacterial and other pathogens. The current NCBI initiative should be evaluated critically by other agencies especially the CDC with regards to its ability to support epidemiological and public health databases. Example database that would benefit from such support in this field include PulseNet and The Universal Virus Database ICTVdB. Beyond taxonomy the use of controlled vocabularies to annotate a genes role in pathogen detection, pathogenicity traits, or treatment selection (i.e. resistance) would be very desirable. Again close coordination between the various branches of the NIH as well as the CDC should be encouraged. Further, the committee was not able to find an adequate standardized gene nomenclature for pathogens. This is not surprising given the range of organisms (viral through malarial), and the dispersed nature of academic research. The committee recommends that simple alternatives such as archiving bacterial and viral genes in a database and assigning them a unique ID might represent a useful initial effort. The committee has discussed these issues with the NIAID. There is general agreement that these issues require timely action.

**Summary:** appropriately addressing the needs of this field is of immense importance in translational research.

1) Translational research would be greatly accelerated if implementation of the HUGN standard were coupled with close coordination with other CHI vocabularies. The NCI is working actively in trying to bridge the gaps between basic and clinical science in these fields, but similar efforts by other entities appears uncoordinated.
2) The field of infectious disease represents a very significant gap in current planning. More active coordination between government agencies is necessary not only for translational research, but also for disease surveillance and bio-defense. CDC input on these issues is of great importance, as is coordinating efforts with NIAID as well as other institutes.
3) Genomic medicine will require the adoption of structured vocabularies by content providers. Data standards should be developed with the active participation of the content providers/clinicians, with implementation mandated

when possible by the NIH/NLM.

4) Finally, in addition to the development of the standardized vocabularies noted above, messaging standards such as HL7® will need to incorporate genomic and proteomic content into their data models.

**Obstacles**

The success of the HUGN will require continued funding (from the NIH and the UK), and active management.

***Appendix A***

**Information Exchange Requirements (IERs)**

| Information Exchange Requirement | Description of IER |
|---|---|
| Beneficiary Financial / Demographic Data | Beneficiary financial and demographic data used to support enrollment and eligibility into a Health Insurance Program. |
| Beneficiary Inquiry Information | Information relating to the inquiries made by beneficiaries as they relate to their interaction with the health organization . |
| Beneficiary Tracking Information | Information relating to the physical movement or potential movement of patients, beneficiaries, or active duty personnel due to changes in level of care or deployment, etc. |
| Body of Health Services Knowledge | Federal, state, professional association, or local policies and guidance regarding health services or any other health care information accessible to health care providers through research, journals, medical texts, on-line health care data bases, consultations, and provider expertise. This may include: (1) utilization management standards that monitor health care services and resources used in the delivery of health care to a customer; (2) case management guidelines; (3) clinical protocols based on forensic requirements; (4) clinical pathway guidelines; (5) uniform patient placement criteria, which are used to determine the level of risk for a customer and the level of mental disorders (6) standards set by health care oversight bodies such as the Joint Commission for Accreditation of Health Care Organizations (JCAHO) and Health Plan Employer Data and Information Set (HEDIS); (7) credentialing criteria; (8) privacy act standards; (9) Freedom of Information Act guidelines; and (10) the estimated time needed to perform health care procedures and services. |
| Care Management Information | Specific clinical information used to record and identify the stratification of Beneficiaries as they are assigned to varying levels of care. |
| Case Management Information | Specific clinical information used to record and manage the occurrences of high-risk level assignments of patients in the health delivery organization.. |
| Clinical Guidelines | Treatment, screening, and clinical management guidelines used by clinicians in the decision-making processes for providing care and treatment of the beneficiary/patient. |

| | |
|---|---|
| Cost Accounting Information | All clinical and financial data collected for use in the calculation and assignment of costs in the health organization . |
| Customer Approved Care Plan | The plan of care (or set of intervention options) mutually selected by the provider and the customer (or responsible person). |
| Customer Demographic Data | Facts about the beneficiary population such as address, phone number, occupation, sex, age, race, mother's maiden name and SSN, father's name, and unit to which Service members are assigned |
| Customer Health Care Information | All information about customer health data, customer care information, and customer demographic data, and customer insurance information. Selected information is provided to both external and internal customers contingent upon confidentiality restrictions. Information provided includes immunization certifications and reports, birth information, and customer medical and dental readiness status |
| Customer Risk Factors | Factors in the environment or chemical, psychological, physiological, or genetic elements thought to predispose an individual to the development of a disease or injury. Includes occupational and lifestyle risk factors and risk of acquiring a disease due to travel to certain regions. |
| Encounter (Administrative) Data | Administrative and Financial data that is collected on patients as they move through the healthcare continuum. This information is largely used for administrative and financial activities such as reporting and billing. |
| Improvement Strategy | Approach for advancing or changing for the better the business rules or business functions of the health organization. Includes strategies for improving health organization employee performance (including training requirements), utilization management, workplace safety, and customer satisfaction. |
| Labor Productivity Information | Financial and clinical (acuity, etc.) data used to calculate and measure labor productivity of the workforce supporting the health organization. |
| health organization Direction | Goals, objectives, strategies, policies, plans, programs, and projects that control and direct health organization business function, including (1) direction derived from DoD policy and guidance and laws and regulations; and (2) health promotion programs. |
| Patient Satisfaction Information | Survey data gathered from beneficiaries that receive services from providers that the health organization wishes to use to measure satisfaction. |

| | |
|---|---|
| Patient Schedule | Scheduled procedure type, location, and date of service information related to scheduled interactions with the patient. |
| Population Member Health Data | Facts about the current and historical health conditions of the members of an organization. (Individuals' health data are grouped by the employing organization, with the expectation that the organization's operations pose similar health risks to all the organization's members.) |
| Population Risk Reduction Plan | Sets of actions proposed to an organization commander for his/her selection to reduce the effect of health risks on the organization's mission effectiveness and member health status. The proposed actions include: (1) resources required to carry out the actions, (2) expected mission impact, and (3) member's health status with and without the actions. |
| Provider Demographics | Specific demographic information relating to both internal and external providers associated with the health organization including location, credentialing, services, ratings, etc. |
| Provider Metrics | Key indicators that are used to measure performance of providers (internal and external) associated with the health organization. |
| Referral Information | Specific clinical and financial information necessary to refer beneficiaries to the appropriate services and level of care. |
| Resource Availability | The accessibility of all people, equipment, supplies, facilities, and automated systems needed to execute business activities. |
| Tailored Education Information | Approved TRICARE program education information / materials customized for distribution to existing beneficiaries to provide information on their selected health plan. Can also include risk factors, diseases, individual health care instructions, and driving instructions. |