istat
working
papers

N.7
2016

Primi risultati sull'implementazione di Iris per la codifica delle cause di morte in Italia: opportunità e sfide

Chiara Orsi, Stefano Marchetti, Luisa Frova, Francesco Grippo

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N.7
2016

Primi risultati sull'implementazione di Iris per la codifica delle cause di morte in Italia: opportunità e sfide

Chiara Orsi, Stefano Marchetti, Luisa Frova, Francesco Grippo

Istat

Istat Working Papers

Primi risultati sull'implementazione di Iris per la codifica delle cause di morte in Italia: opportunità e sfide

N. 7/2016

ISBN 978-88-458-1892-9

Primi risultati sull'implementazione di Iris per la codifica delle cause di morte in Italia: opportunità e sfide[1]

Chiara Orsi, Stefano Marchetti, Luisa Frova, Francesco Grippo

Sommario

Iris è un software per la codifica automatica delle cause di morte e la selezione della causa iniziale. Attualmente è lo strumento maggiormente utilizzato a livello internazionale e contribuisce significativamente all'armonizzazione delle statistiche di mortalità per causa. Obiettivo del lavoro è descrivere le attività svolte per l'adozione di Iris in Italia e la valutazione dell'efficienza del sistema in termini di percentuale di codifica automatica. La costruzione del dizionario e dell'insieme delle regole di standardizzazione è stata effettuata per passi successivi. Nell'ultimo passo la percentuale di codifica automatica raggiunta del 77,6% è paragonabile a quella ottenuta con il sistema di codifica attuale, tuttavia Iris garantisce una maggiore confrontabilità internazionale e una migliore integrabilità nel processo di trattamento e produzione dei dati.

Parole chiave: cause di morte, codifica automatica, confrontabilità internazionale.

Abstract

Iris is a software for the automatic coding of the causes of death and the selection of the underlying cause. Currently it is the most widely used tool at international level and it significantly contributes to the harmonisation of mortality statistics by cause of death. Objective of this paper is to describe the activities carried out in order to implement Iris in Italy and to evaluate the efficiency of the software in terms of automatic coding percentage. The drafting of the dictionary and the setting up of standardisation rules was performed by subsequent steps. In the last step the percentage of automatic coding equal to 77.6%, is comparable to the percentage obtained with the coding system currently used at Istat (Italian National Institute of Statistics); Iris, however, provides a greater international comparability and a better integration in the treatment and production process of data on causes of death.

Keywords: causes of death, automatic coding, international comparability.

[1] Il lavoro è il frutto della collaborazione di tutti gli autori che hanno partecipato alla stesura e alla revisione di tutto il testo. In particolare sono da attribuire a Chiara Orsi i par. 4.2, 5.1 e 5.2; a Stefano Marchetti il cap. 2 e par 4.1; a Luisa Frova il par.5.3 e cap. 6; a Francesco Grippo i cap. 1 e 3.

Indice

1. Iris: uno strumento per la codifica delle cause di morte

Iris è un software internazionale utilizzato per la codifica automatica delle cause multiple di morte e per la selezione della causa iniziale (Iris Institute 2015). Esso ha come principale obiettivo quello di aumentare la comparabilità dei dati fra paesi rendendo agevole l'adozione delle procedure di codifica automatica nei vari contesti nazionali.

Iris è sviluppato sulla base di linee guida internazionali fornite dall'Organizzazione Mondiale della Sanità (OMS) nella Classificazione Internazionale delle Malattie, attualmente alla sua decima revisione (ICD10) (Ministero della Sanità 2000). L'adozione di queste linee guida garantisce che la selezione della causa iniziale, generalmente utilizzata per i confronti internazionali, sia armonizzata e dia luogo a informazioni il più possibile rilevanti dal punto di vista della salute pubblica. Le linee guida dell'ICD10 si concretizzano in un sistema di regole e istruzioni di codifica che configurano un complesso algoritmo con numerosi snodi decisionali (WHO 2015).

La complessità insita in queste regole può dare adito a interpretazione personale da parte dei singoli codificatori, introducendo una variabilità tra codificatori (Harteloh et al. 2010) A questo errore spesso si associa una componente sistematica dovuta al consolidamento di pratiche di codifica che si discostano dallo standard internazionale stabilito dall'ICD. Ciò può avvenire, ad esempio, a causa della applicazione parziale degli aggiornamenti previsti dall'OMS.

Come tutti i sistemi automatici in generale, Iris contribuisce a controllare queste fonti di errore. Innanzitutto, il trattamento automatico consente di limitare la variabilità inter-codificatore e questa componente resta solo sulle schede scartate dalla codifica automatica e codificate a mano.

Tra i sistemi di codifica automatica, il *Mortality Medical Data System* (MMDS), sviluppato alla fine degli anni '60 dall'NCHS negli Stati Uniti, ha costituito per decenni lo standard *de facto* (CDC, 2015). Le tavole di decisione utilizzate dal software, ovvero le tavole ACME, (CDC, 2014) sono state considerate il principale documento di riferimento anche per la codifica manuale. Tuttavia MMDS non è facilmente adattabile ai diversi contesti linguistici in quanto uno dei suoi moduli (MICAR) richiede un input in cui le espressioni diagnostiche devono essere trasformate in codici non internazionali (ERN, Entity Reference Number) legati alla terminologia medica americana. Il software successivamente utilizza questi codici per elaborare la codifica delle cause multiple necessaria per la selezione delle causa iniziale con il modulo ACME. La trasformazione del testo in ERN richiede, per i paesi non di lingua inglese, la traduzione del dizionario statunitense e lo sviluppo di procedure per la precodifica.

Iris supera invece questo limite separando le componenti lingua-dipendenti dai moduli che effettuano la codifica. Questo favorisce l'uso di Iris in diverse lingue.

Iris è utilizzato in molti paesi per la produzione delle statistiche ufficiali sulle cause di morte già da vari anni. Il consenso internazionale verso il software è anche il frutto della modalità di sviluppo e governance. Alcuni paesi (Francia, Germania, Italia, Svezia, Stati Uniti e Ungheria) collaborano attivamente allo sviluppo, mettendo a disposizione competenze informatiche e nosologiche. Questa collaborazione è formalizzata attraverso accordi che ciascun istituto nazionale ha siglato con il DIMDI (Istituto Tedesco di Documentazione Medica) per la fondazione dell'Iris Institute. Altri Paesi collaborano al progetto finanziariamente o partecipando attivamente alle riunioni annuali (User's group) in cui vengono riportate le esperienze di implementazione e vengono proposti miglioramenti. Iris è inoltre un software gratuito il cui download è libero.

Nell'ambito dell'User's group[2] sono discusse le strategie per la costruzione del dizionario e delle regole di standardizzazione, l'impatto sulla qualità dei dati e soprattutto l'impatto sulle serie storiche di mortalità per causa.

Le strategie per lo sviluppo del dizionario sono eterogenee e poco generalizzabili per vari motivi. Oltre alla specificità della lingua è molto vario il contesto in cui Iris viene implementato. In al-

[2] I contributi dei paesi partecipanti sono disponibili sul sito dell'Iris Institute www.iris-institute.org

cuni paesi i dizionari sono stati realizzati a partire da tesauri sviluppati per altri scopi, come ad esempio il supporto alla codifica della morbosità (Weber e Özer 2008, Cristófori Martins 2012). Altri paesi hanno dovuto affrontare il problema del bilinguismo o multilinguismo (Spagna), mentre in Repubblica Ceca non è stato sviluppato un dizionario in quanto Iris è utilizzato esclusivamente per la selezione della causa iniziale (Poppová 2011).

Particolarmente significative risultano le esperienze di valutazione sull'impatto dell'introduzione di Iris sulle serie storiche (Poppová 2012, ONS 2014). Queste esperienze mettono in evidenza che l'impatto principale è dovuto ad una migliore applicazione delle regole di selezione rispetto alle prassi precedenti.

2. Adozione di Iris in Italia: contesto e vantaggi attesi

L'attuale processo di produzione dei dati sulle cause di morte prevede che i medici riportino le cause su modelli cartacei[3]. I medici sono tenuti a certificare, scrivendo per esteso, tutte le cause e le malattie che secondo il loro giudizio hanno contribuito al decesso. I modelli compilati (oltre 600 mila modelli annui) vengono poi raccolti mensilmente e inviati alla registrazione in service. La trasformazione dei certificati cartacei in informazioni su supporto elettronico è una fase particolarmente onerosa e delicata considerando che gli addetti alla registrazione non sono esperti di terminologia medica e a volte è difficile interpretare la grafia manuale. Le attuali procedure utilizzate per la codifica automatica non sono completamente generalizzate e richiedono in fase di registrazione un pretrattamento manuale dei dati per garantire performance soddisfacenti del sistema di codifica automatico. Per limitare gli errori nelle fasi di acquisizione e pretrattamento delle informazioni il personale addetto alla registrazione in service dei modelli viene adeguatamente formato e monitorato nel tempo.

Il software attualmente utilizzato per la codifica (CodSanII) permette la lavorazione automatica di circa 78% dei decessi (anche grazie al pretrattamento manuale) e la codifica manuale interattiva degli scarti. CodSanII è un software sviluppato ad hoc che include un prodotto di codifica generalizzato (ACTR) per il trattamento del testo in italiano e il sistema statunitense MMDS (in particolare i moduli Micar-Acme) realizzato dall'NCHS, per la codifica delle cause di morte.

La necessità di adottare Iris dismettendo l'attuale sistema per la codifica delle cause di morte in Italia scaturisce dall'insieme di due fattori, ognuno dei quali rilevante e vincolante.

Il primo problema che renderà a breve obsoleto l'attuale sistema scaturisce dalla decisione dell'NCHS di non rilasciare il software con gli aggiornamenti successivi al 2009 ed in linea con l'evoluzione della Classificazione. Pertanto i dati prodotti con MMDS non sono completamente confrontabili con quelli della maggioranza dei paesi che hanno implementato aggiornamenti dell'ICD-10 più recenti. Inoltre, questo gap di confrontabilità aumenta con il passare del tempo.

Contemporaneamente si è resa opportuna, se non inevitabile, la sperimentazione di nuove modalità di trattamento dei dati conseguenti a una radicale innovazione del flusso di produzione, che prevedrà la certificazione elettronica delle cause di morte da parte dei medici[4].

In vista di una tale innovazione, sarà richiesto ai medici di riportare le cause di morte in testo libero. Secondo raccomandazioni internazionali (WHO 2010) non potranno, infatti, essere utilizzati sistemi di autocompilazione che, sebbene agevolerebbero la successiva fase di codifica, avrebbero un effetto sulla riduzione sistematica della variabilità e sulla specificità delle cause riportate introducendo una distorsione delle statistiche di mortalità per causa. L'introduzione della certificazione via web prevedrà pertanto la codifica delle cause in testo libero, modalità che l'attuale sistema, a differenza di Iris, non è in grado di gestire.

Inoltre l'utilizzo di un software progettato e realizzato specificatamente per agevolare la gestione della codifica delle cause di morte nei diversi Paesi comporta una serie di ulteriori importanti

[3] Modelli Istat D4 e D4bis, adattamenti nazionali dello standard Oms.

[4] DPCM 10 novembre 2014, n. 194 prevede che tra i servizi dell'Anagrafe Nazionale della Popolazione Residente (ANPR) sia prevista l'acquisizione digitale delle cause di morte.

vantaggi. Con Iris la lavorazione di una scheda riguarda contemporaneamente il riconoscimento del testo, la codifica delle cause multiple e la selezione della causa iniziale. In caso di mancato successo delle procedure automatiche anche le attività di intervento manuale si inseriscono in un contesto di gestione complessiva della scheda. Viene così meno la necessità di gestire la codifica in passi successivi e il conseguente rischio di più interventi manuali sulla stessa scheda.

Le procedure di check sono contestuali alla fase di codifica e l'insieme delle regole di compatibilità della causa con il sesso e l'età del deceduto viene fornito dall'Iris Institute con il rilascio delle versioni di Iris in linea con gli aggiornamenti ICD. Questo evita di dover riconsiderare a livello nazionale il piano di check ad ogni rilascio di versioni aggiornate, fermo restando che è sempre possibile integrare le regole corrette a livello internazionale con gli opportuni adattamenti nazionali (Istat 2013).

La gestione integrata di tutte le fasi di codifica rende agevole anche, in caso di evidenti errori di battitura, modificare direttamente il testo delle espressioni diagnostiche ed eseguire nuovamente la codifica automatica.

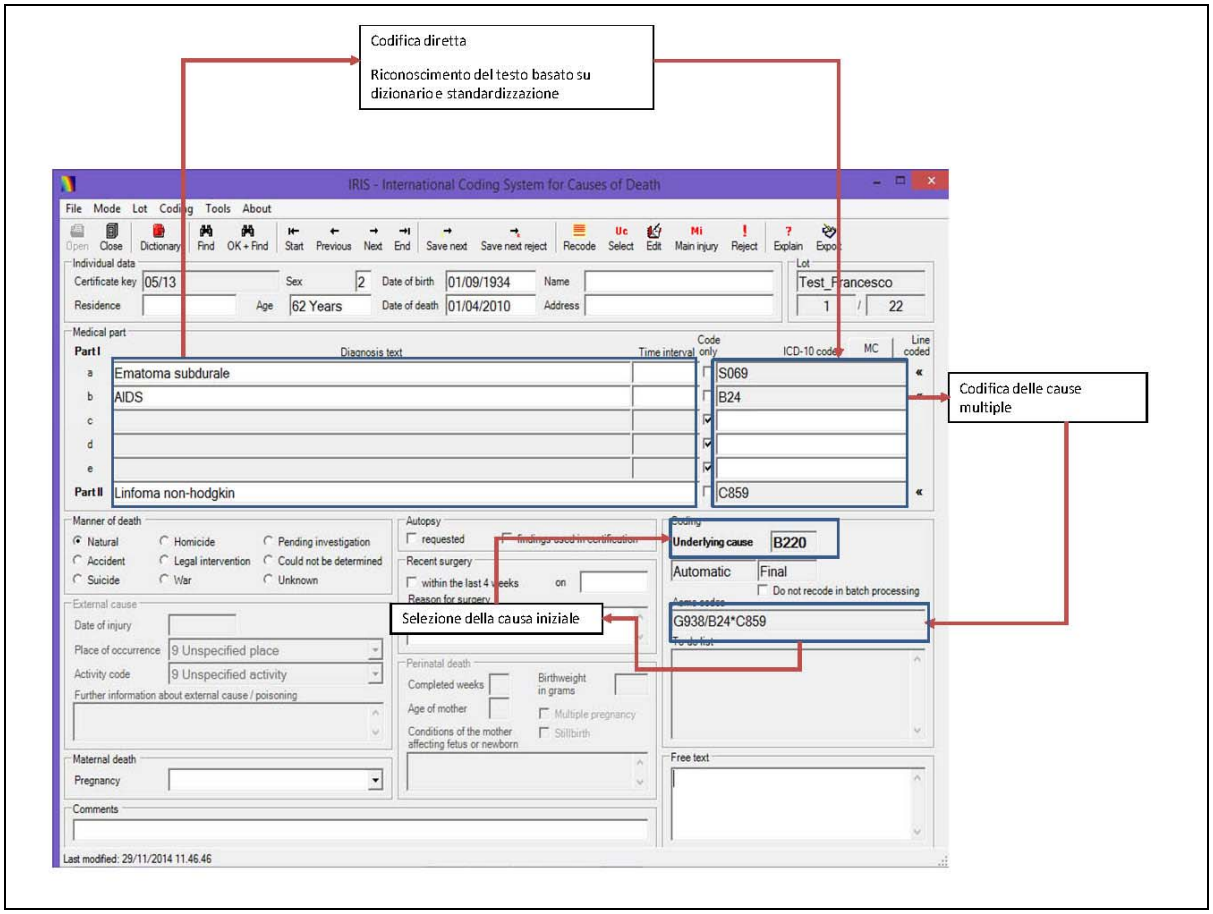
In ultimo, Iris dispone di dettagliati strumenti di interazione con l'utente, reportistica e help online, che supportano e istruiscono il codificatore durante le fasi di lavorazione.

3. Funzionamento di Iris

Lo schema di funzionamento e la maschera di Iris sono presentate in figura 1. Le fasi di elaborazione possono essere così riassunte (Iris Institute 2014):

- codifica delle singole patologie (codifica "diretta") ovvero riconoscimento del testo riportato sul certificato e, attraverso un apposito dizionario, attribuzione di un codice ICD10 a ciascuna patologia (questa fase è facoltativa fornendo ad Iris direttamente i codici ICD10);
- modifica dei codici "diretti" (codifica delle cause multiple) secondo regole definite che tengono conto delle caratteristiche del deceduto (età e sesso) e delle altre informazioni riportate sul certificato (modalità di decesso, durata delle patologie, insieme delle patologie riportate, ecc.);
- selezione della causa iniziale attraverso le regole dell'OMS (ICD10) di selezione e modifica.

Figura 1. Schema di funzionamento di Iris



La fase descritta nel primo punto, la codifica di ciascuna patologia con un opportuno codice diretto, è legata alla lingua in cui vengono redatti i certificati di morte. Per poter effettuare automaticamente questa fase, bisogna fornire al software due strumenti:

- un dizionario di termini diagnostici associato ai codici ICD10;
- un sistema di regole di standardizzazione (questo strumento non è strettamente necessario, ma garantisce una gestione più efficiente).

Le regole di standardizzazione modificano le espressioni diagnostiche riportate sui certificati in espressioni "equivalenti" con lo scopo di renderle uguali a quelle presenti nel dizionario. Tali regole agiscono su vari aspetti del testo: sinonimi, abbreviazioni, termini irrilevanti, acronimi, singolare/plurale/maschile/femminile, ecc. Un esempio di regola di standardizzazione è riportato nella figura 2.

Figura 2. Esempio di regola di standardizzazione

L'algoritmo di standardizzazione utilizzato da Iris richiede che le regole di standardizzazione siano espresse sotto forma di espressioni regolari (regular expression, "regex") che sono delle stringhe scritte con una sintassi standardizzata utilizzate per il riconoscimento di specifiche porzioni di testo (Weber 2006).

La standardizzazione rappresenta quindi uno strumento che aumenta la probabilità di codifica della singola patologia e che consente di limitare il numero di termini inclusi nel dizionario. Infatti, molti termini diversi possono essere standardizzati allo stesso modo e accoppiarsi ad un unico termine del dizionario. Più è efficiente il sistema di standardizzazione più le dimensioni del dizionario possono essere contenute. Un dizionario di dimensioni contenute è anche più facilmente gestibile sia per la correzione di eventuali errori sia per l'aggiornamento dei codici derivante dal continuo aggiornamento dell'ICD.

Le fasi successive della codifica con Iris sono indipendenti dalla lingua, pertanto, tutti gli aggiornamenti riguardanti questa parte sono sviluppati dall'Iris Institute e non sono un onere per i singoli Paesi.

Va ricordato che la versione 4.4.1 di Iris, utilizzata per questo report, incorpora i due moduli core di MMDS (Micar e Acme). Attualmente l'Iris Insitute sta predisponendo una nuova versione con moduli sviluppati ex novo con l'obiettivo di superare alcune criticità legate soprattutto a MICAR, come la codifica delle cause esterne, della mortalità perinatale e della mortalità dovuta a complicanze di interventi chirurgici (Eckert 2014).

4. Implementazione di Iris in Italia

4.1 Necessità di costruire un nuovo dizionario

L'adozione di Iris in Italia per la produzione delle statistiche ufficiali di mortalità richiede una serie di attività che garantiscano qualità della codifica, efficienza e manutenibilità del sistema.

In questo lavoro vengono descritte le attività finalizzate alla costruzione di un dizionario che permetta il raggiungimento di una percentuale di codifica automatica confrontabile con le performance del sistema attuale (circa 78%). Questo obiettivo passa attraverso la predisposizione di un adeguato sistema di regole di standardizzazione del testo.

Nel corso del lavoro sono state ottenute diverse versioni del dizionario, ciascuna delle quali è stata sottoposta a valutazione della performance come dettagliato nei successivi paragrafi. Il dizionario necessario alla codifica automatica con Iris consiste di un insieme di espressioni diagnostiche[5] ad ognuna delle quali è associato un codice della classificazione ICD10.

Per la costruzione del dizionario di Iris non è stato possibile basarsi su quello utilizzato nell'attuale sistema fondato sugli ERN, poiché Iris richiede un dizionario basato sui codici ICD10.

Inoltre, il dizionario precedente era stato realizzato tenendo conto delle caratteristiche di ACTR, in particolare delle regole di standardizzazione utilizzabili che sono in parte diverse da quelle di Iris. I dizionari infatti devono essere predisposti per essere utilizzati insieme alla standardizzazione del sistema di codifica automatica scelto.

4.2 Creazione del dizionario e definizione delle regole di standardizzazione

Per lo sviluppo del dizionario di Iris si è partiti da un insieme di termini diagnostici codificati con un codice ICD10 e successivamente sono state adottate delle strategie per poterne ridurre le dimensioni e aumentarne la qualità.

L'archivio di partenza (A_0) è costituito dalle espressioni diagnostiche presenti sulla casistica reale dei certificati di decesso redatti dai medici e sottoposte alla registrazione con pretrattamento manuale del testo[6]. In particolare vi sono tutte le patologie riportate sui certificati di morte del 2011 e dei primi 6 mesi del 2012[7] di cui si dispone della relativa codifica ERN ottenuta con il sistema di codifica automatica correntemente in uso. Queste espressioni corrispondono a 106.476 termini diversi tra loro. Il corrispondente codice ICD10 è stato ottenuto con la tabella di corrispondenza

[5] Nel testo vengono utilizzati diversi modi di identificare i termini del dizionario a seconda del contesto: "espressioni diagnostiche", "stringhe", "patologie", "righe del dizionario", "terminologia medica". Considerare tali termini come sinonimi.

[6] La registrazione con pretrattamento è brevemente descritta nel paragrafo 2.

[7] Nel complesso sono stati considerati circa 900.000 certificati di morte con una media di 3,4 patologie per certificato.

ERN-ICD10 fornita dall'NCHS.

4.2.1 Sviluppo delle regole di standardizzazione e generazione di diverse versioni del dizionario

Partendo dall'archivio A_0 si è proceduto per passi successivi ottenendo a ciascun passo:

1. una versione corretta dell'archivio (A_i), che costituisce l'input per il passo successivo (i+1);
2. una dizionario standardizzato (D_i) che rappresenta lo strumento da utilizzare per la codifica automatica e che viene sottoposto a valutazione della performance.

Nel primo passo un insieme iniziale di regole di standardizzazione viene applicato a ciascun termine dell'archivio A_0. Questo produce un dizionario con termini standardizzati che in alcuni casi possono essere duplicati. Ciò permette di individuare eventuali incongruenze (due o più stringhe associate nell'archivio di partenza a codici diversi che danno origine, dopo la standardizzazione, alla stessa stringa) oppure di mettere in evidenza e correggere errori nelle regole di standardizzazione. Correggendo le incongruenze dovute al dizionario viene generato un archivio corretto A_1 (archivio di partenza senza le incongruenze emerse al passo 1), a partire dal quale, sostituendo il testo con quello standardizzato e eliminando le ridondanze si genera il dizionario D_1 di dimensioni ridotte. Si specifica che le varie versioni di A_i contengono sempre lo stesso numero di stringhe grezze e si differenziano tra loro solo per la correzione dei codici associati.

I successivi passi ripercorrono iterativamente quanto fatto nel passo 1. La procedura iterativa prevede, quindi, che ogni passo i sia costituito dalle seguenti fasi (schematizzate in figura 3):

1. sviluppo delle regole di standardizzazione $\{r_i\}$.;
2. integrazione con le regole corrette ottenute nei passi precedenti $R_i = U_i \{\hat{f}_{i-1}\} \{r_i\}$;
3. standardizzazione dell'archivio A_{i-1} corretto al passo precedente con l'insieme delle regole R_i;
4. controllo e correzione degli errori della standardizzazione ottenendo $\{\hat{f}_i\}$;
5. individuazione e correzione degli errori dell'archivio A_{i-1} (incongruenze) e produzione di A_i;
6. applicazione delle regole di standardizzazione corrette $\hat{R}_i = U_i \{\hat{f}_i\}$;
7. individuazione ed eliminazione dei termini standardizzati duplicati ottenendo la versione del dizionario standardizzato D_i;
8. valutazione della performance di Iris con il dizionario D_i separatamente sui due file di dati.

Con questa procedura si ottiene un archivio di termini codificati sempre più corretto e viene generato un dizionario standardizzato di ridotte dimensioni e senza incongruenze.

Per completezza, di seguito vengono analizzati in dettaglio i casi di incoerenza e ridondanza, le cause che li generano e le modalità di gestione.

4.2.2 Casi di incongruenza

Si ha un'incongruenza quando due o più stringhe associate nell'archivio di partenza a codici diversi danno origine, dopo la standardizzazione, alla stessa stringa.

Una volta individuate le incongruenze, si analizza a cosa sono dovute. Ci sono tre possibilità:

1. L'incongruenza è dovuta a un errore nell'archivio (A). Un esempio di incongruenza dovuta a un errore nell'archivio è riportata nella tavola 1: le stringhe "BPC ostruttiva" e "BPCO ostruttiva" erano associate erroneamente al codice J44.9, il codice corretto per queste stringhe è J44.8. Questo errore è emerso perché entrambe queste stringhe sono state standardizzate in BPCO e quest'ultimo termine era già presente nell'archivio (A) con il codice corretto J44.8.

Tavola 1 – Esempio di incongruenza dovuta a errore nell'archivio iniziale

Stringa	Codice assegnato alla stringa nell'archivio	Regole di standardizzazione applicate	Risultato della standardizzazione	Codice corretto
BPC ostruttiva	J44.9	BPC Ostruttiva=BPCO	BPCO	J44.8
BPCO ostruttiva	J44.9	BPCO ostruttiva=BPCO	BPCO	J44.8
BPCO	J44.8	Nessuna	BPCO	J44.8

2. L'incongruenza è dovuta a un errore nelle regole di standardizzazione. Un esempio di incongruenza dovuta a un errore nelle regole di standardizzazione è riportato nella tavola 2: in un primo momento era stata sviluppata una regola che eliminava il termine "lieve" (e sinonimi), perché ritenuto ininfluente, ma questo processo ha permesso di evidenziare che il termine non è ininfluente per alcune patologie. La regola di standardizzazione è stata quindi eliminata dall'insieme.

Tavola 2 – Esempio di incongruenza dovuta a un errore in una regola di standardizzazione

Stringa	Codice assegnato alla stringa nell'archivio	Regole di standardizzazione applicate	Risultato della standardizzazione	Codice corretto
Ritardo mentale	F72	Nessuna	Ritardo mentale	F72
Lieve ritardo mentale	F71	Eliminare "lieve" e sinonimi	Ritardo mentale	F71

3. Una combinazione dei due casi precedenti: l'incongruenza è dovuta sia a un errore nell'archivio che a un errore nelle regole di standardizzazione, un esempio è riportato nella tavola 3. In questo caso l'eliminazione del termine "severo" porta a codificare nello stesso modo la malnutrizione e la malnutrizione di grado severo. Inoltre nell'archivio A alla stringa "stato severa malnutrizione" era associato un codice errato. Si è quindi intervenuto correggendo l'archivio ed eliminando la regola di standardizzazione che eliminava il termine "severo".

Tavola 3 – Esempio di incongruenza mista

Stringa	Codice assegnato alla stringa nell'archivio	Regole di standardizzazione applicate	Risultato della standardizzazione	Codice corretto
Malnutrizione	E46	Nessuna	Malnutrizione	E46
Malnutrizione grado severo	E43	Eliminare "severo" (e sinonimi)	Malnutrizione	E43
Stato severa malnutrizione	E46	Eliminare "severo" (e sinonimi)	Malnutrizione	E43

Gli errori individuati sia nell'archivio che nelle regole di standardizzazione vengono corretti e il procedimento viene ripetuto, fino al completo esaurimento delle incongruenze.

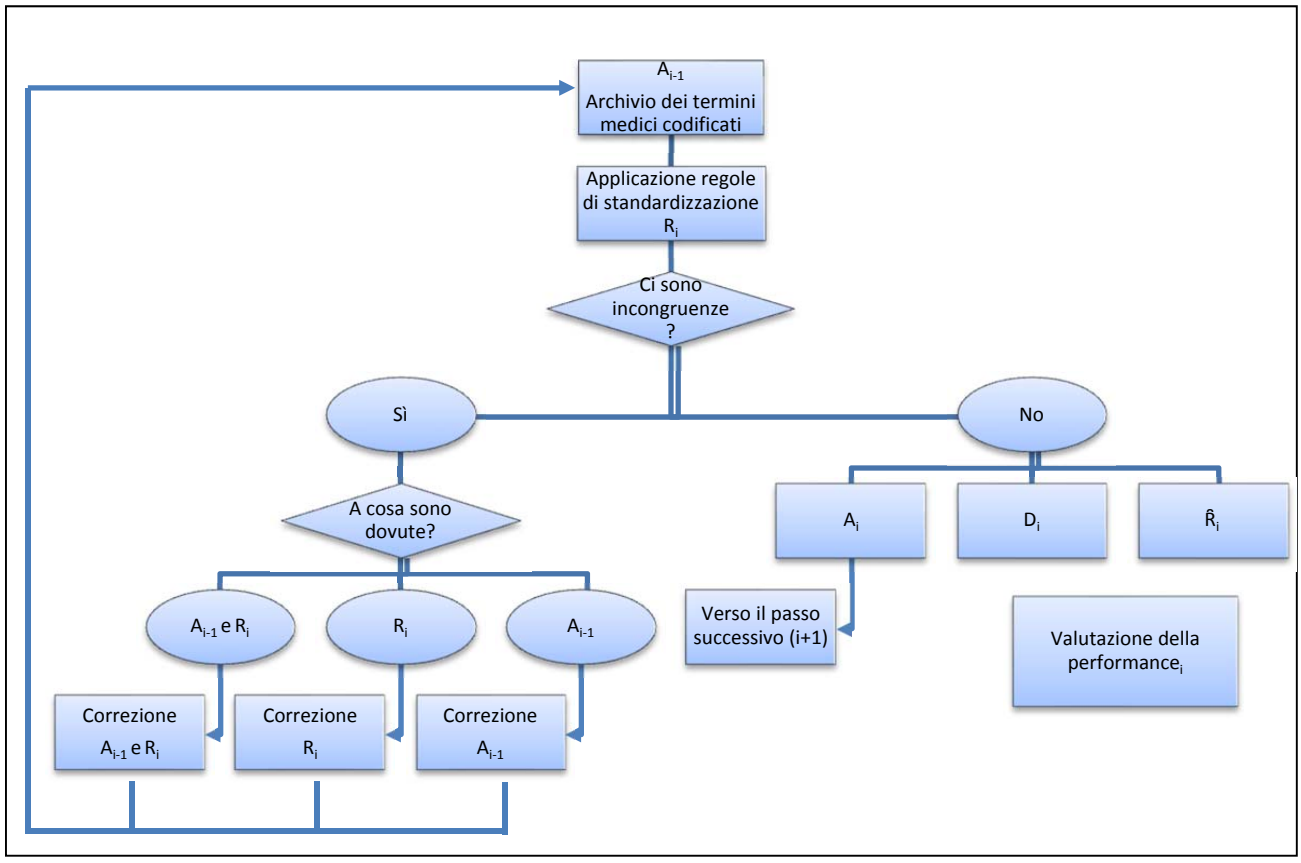
4.2.3 Casi di ridondanza

Si ha una ridondanza quando due o più stringhe dell'archivio, a cui è associato lo stesso codice ICD10, danno origine, dopo la standardizzazione, alla stessa stringa. Una volta individuate tutte le ridondanze, vengono mantenute soltanto le stringhe standardizzate eliminando le duplicazioni, dando luogo alla versione successiva del dizionario, di dimensioni minori rispetto a quella precedente. Un esempio di ridondanza è riportato nella tavola 4.

Tavola 4 – Esempio di ridondanza

Stringa	Codice assegnato alla stringa nell'archivio	Regole di standardizzazione applicate	Risultato della standardizzazione
Insufficienza cardiovascolare	I51.6	Insufficienza=Insuff	Insuff cardiovascolare
Insufficienza cardio vascolare	I51.6	Insufficienza=Insuff Cardio vascolare=cardiovascolare	Insuff cardiovascolare
Insuff cardiovascolare	I51.6	Nessuna	Insuff cardiovascolare
Insufficienza cardiovascolare irreversibile	I51.6	Insufficienza=Insuff Eliminare "irreversibile"	Insuff cardiovascolare
Insuff cardio vascolare	I51.6	Cardio vascolare=cardiovascolare	Insuff cardiovascolare
Insuf cardiovascolare	I51.6	Insuf=Insuff	Insuff cardiovascolare
Iinsuffic cardiovascolare	I51.6	Iinsuffic=Insuff	Insuff cardiovascolare
Insuf cardio vascolare	I51.6	Insuf=Insuff Cardio vascolare=cardiovascolare	Insuff cardiovascolare
Insufficienza cardio vascolare irreversibile	I51.6	Insufficienza=Insuff Cardio vascolare=cardiovascolare Eliminare "irreversibile"	Insuff cardiovascolare
Insufficienza presumibile cardiovascolare	I51.6	Insufficienza=Insuff Eliminare "presumibile"	Insuff cardiovascolare

Figura 3 – Creazione della versione i del dizionario (Di) a partire dalla versione i dell'insieme di regole di standardizzazione (Ri)



Il procedimento decritto evidenzia come, a ogni passo, l'insieme di regole di standardizzazione più ampio permette di codificare un maggior numero di stringhe. Contemporaneamente si è ottenuto un archivio A migliore (potrà in futuro essere il punto di partenza in caso di revisione della ICD) e un dizionario di termini standardizzati D di dimensioni ridotte e quindi più gestibile.

5. Valutazione della performance

5.1 Metodologia

Per valutare l'efficacia e l'efficienza di questo nuovo sistema di codifica automatica -adattato alla casistica italiana- e per valutare la performance e la qualità del dizionario e della standardizzazione ottenuti ai diversi passi, sono stati analizzati diversi aspetti.

Per ogni versione del dizionario e per ciascun sottoinsieme di regole di standardizzazione è stata calcolata la percentuale di certificati codificati con Iris sulle due diverse basi di dati: la prima di tipo "tradizionale" è costituita dai certificati di morte relativi al mese di dicembre 2012 pre-trattati durante la registrazione (file di input per ACTR), la seconda base di dati è invece di tipo "sperimentale". Quest'ultima è stata preparata registrando su formato elettronico gli stessi certificati di morte (dicembre 2012) senza trattamento manuale, quindi senza intervento da parte dell'operatore di modifica o standardizzazione del testo. Questa seconda sperimentazione consente di valutare le potenzialità del software Iris anche in vista dell'introduzione del certificato elettronico, ovvero quando sarà necessario lavorare con stringhe esclusivamente digitate on-line dai medici certificatori senza particolari istruzioni da seguire durante la compilazione.

Per ogni passo è stata valutata la riduzione delle dimensioni del dizionario rispetto al passo precedente e rispetto all'archivio iniziale. Sono state inoltre contate le incongruenze individuate in ogni passo e il conseguente numero di stringhe corrette nell'archivio iniziale (ciascun errore individuato comporta la correzione di uno o più termini del dizionario).

5.2 Principali risultati

La tavola 5 riporta, per ciascun passo, la percentuale di certificati codificati da Iris sia sull'invio tradizionale che su quello sperimentale. La percentuale di certificati codificati dell'invio tradizionale è passata da 75,5% (nei passi iniziale e nel passo 1) a 77,6% (nei passi 2 e 3). La percentuale di certificati dell'invio sperimentale codificati è passata da 57,7% nel passo 0 a 64,7 nel passo 1 a 65,9% nel passo 2 a 73.2% nel passo 3. È riportata inoltre, come termine di paragone, la percentuale di certificati codificati con il sistema di codifica attuale (ACTR+MMDS), pari a 78%.

Tavola 5 – Valutazione delle performance di IRIS nei 4 passi di ottimizzazione del sistema. Percentuale di certificati codificati per tipo di dato.

	Tipo di dato in input ad IRIS	
Codifica	File tradizionale (dati pre-trattati) %	File sperimentale (dati non trattati) %
Passo 0: A_0+R_0[a]	75,5	57,7
Passo 1: D_1+R_1	75,5	64,7
Passo 2: D_2+R_2	77,6	65,9
Passo 3: D_3+R_3	77,6	73,2
ACTR+MMDS	78.0	-

(a) Nel passo 0 le regole di standardizzazione non intervengono sulla terminologia medica ma consistono in un set minimale di istruzioni necessarie ad Iris per l'interpretazione delle durate e di una regola che trasforma l'espressione "dovuto a" nel simbolo "\" (simbolo utilizzato da Iris come separatore causale).

La percentuale di certificati codificati da Iris lievemente inferiore a quella dell'attuale sistema di codifica è un risultato atteso. Una parte degli scarti di Iris è infatti dovuta ad alcuni codici inclusi nel dizionario in uso relativi alla versione 2009 dell'ICD-10, mentre la versione Iris utilizzata per questa sperimentazione si basa sulla versione 2014 e pertanto non riconosce alcuni codici diventati obsoleti.

La notevole differenza che sussiste inizialmente tra la percentuale di certificati codificati nell'invio sperimentale rispetto al risultato ottenuto sul file tradizionale è una misura dell'impatto

del pretrattamento manuale del testo sulla performance della codifica automatica. D'altro canto, l'aumento della percentuale di certificati codificati sul file sperimentale nei passi successivi evidenzia la capacità delle nuove regole di standardizzazione di sostituire gradualmente le istruzioni applicate oggi manualmente in fase di registrazione; garantendo, di conseguenza, la possibilità di utilizzare i dati acquisiti attraverso strumenti di certificazione elettronica senza diminuire in maniera significativa la performance.

La tavola 6 riporta il numero di termini inclusi nel dizionario standardizzato e la riduzione percentuale della sua dimensione rispetto al passo precedente. Dalla versione iniziale, nella quale D_0 coincide con l'archivio delle stringhe A_0, all'ultima (D_3) il dizionario è passato da 106.476 a 74.131 termini, con una riduzione complessiva del 30,4%.

Tavola 6 – Effetto delle strategie di standardizzazione sulla riduzione della dimensione del dizionario e sull'individuazione e correzione degli errori. Numero di termini e riduzione percentuale della dimensione del dizionario.

Codifica	Termini nel dizionario Di	Riduzione % dei termini di D_i rispetto al passo precedente	Riduzione % dei termini di D_i rispetto ad A	Numero di Incongruenze individuate	Numero di correzioni apportate su A
Passo 0	106.476	-	-	-	-
Passo 1	85.649	19,6	19,6	446	668
Passo 2	76.683	10,5	28,0	484	805
Passo 3	74.131	3,3	30,4	55	81

La tavola 6 riporta anche, per ciascun passo, il numero di incongruenze individuate nel dizionario e il conseguente numero di correzioni sull'archivio. Le correzioni sono maggiori degli errori individuati in quanto ad una incongruenza possono corrispondere a più stringhe da correggere (cfr tavola 2 paragrafo precedente).

5.3. Strategie di standardizzazione e impatto sulla performance

Ogni passaggio ha comportato la costruzione di nuove regole di standardizzazione e le corrispondenti versioni del dizionario.

Nel passo 1 sono state formulate regole di standardizzazione per l'eliminazione di termini ininfluenti e la gestione di variazioni sintattiche (genere, plurali, abbreviazioni). Questo primo passo ha avuto come effetto più rilevante la riduzione delle dimensioni del dizionario, pur restando costante la performance di codifica dell'invio tradizionale (75,5%). La riduzione delle dimensioni si accompagna anche alla messa in evidenza di molte incongruenze (446).

Sull'invio sperimentale, il passo 1 ha avuto un impatto significativo aumentando la performance del 7% (da 57,7 al 64,7). Le regole di standardizzazione che hanno maggiormente contribuito a questo risultato sono quelle che riguardano l'eliminazione delle preposizioni (di, del, al, ecc…) e di parole ininfluenti (paziente, ecc.) che fanno parte delle regole di pre-trattamento manuale e quindi non si trovano nell'invio tradizionale.

Nel passo 2, sono state aggiunte ulteriori regole per la gestione di variazioni sintattiche. Nello sviluppo di questo passo si è tenuto conto dei risultati del test di performance del dizionario ottenuto al passo 1. In particolare è stata analizzata la casistica dell'invio tradizionale scartata da Iris al passo 1. Le nuove regole di standardizzazione sono state quindi costruite in modo da poter trattare questa casistica. L'effetto di questo approccio è reso evidente dall'aumento della performance di codifica sull'invio tradizionale (si passa da 75,5 a 77,6% al passo 2).

Nel passo 3 nuove regole di standardizzazione sono state sviluppate dopo un'attenta analisi dei motivi di scarto dell'invio sperimentale al passo precedente. È stato inoltre ampliato lo sviluppo delle regole attualmente applicate manualmente in fase di data entry e di quelle riguardanti i termini di congiunzione tra le patologie (per esempio con, accompagnato da, ecc.). Questo ultimo set di istruzioni è indispensabile per poter codificare stringhe non pre-trattate con sistemi automatici. Questa attività ha avuto come conseguenza un ulteriore incremento della performance di codifica dell'invio sperimentale che è passata da 65,9 a 73,2 (+ 7,3%).

Su una lavorazione annuale, è possibile stimare che ciascun punto percentuale di miglioramento della performance corrisponda a circa 6 mila certificati, equivalenti a un risparmio di 30 giorni/persona lavorativi.

6. Prossimi passi

Le attività descritte nel presente lavoro hanno consentito il raggiungimento dell'obiettivo primario di questa prima fase: ottenere percentuali di codifica automatica paragonabili al sistema corrente. La successiva fase di studio dovrà essere, quindi, maggiormente incentrata sulla valutazione e il miglioramento della qualità.

L'utilizzo di Iris è previsto per la codifica dei dati ufficiali di mortalità riferiti all'anno 2015. Questo significa che il software dovrà entrare in completo esercizio nei primi mesi del 2016, dopo un periodo di formazione del personale addetto alla codifica.

Alla data attuale, sebbene siano state raggiunte delle performance adeguate, è possibile prevedere nei prossimi mesi un ulteriore contributo all'ampliamento delle regole di standardizzazione finalizzate soprattutto al riconoscimento del testo non pretrattato.

Propedeutiche all'adozione di Iris per le statistiche ufficiali di mortalità sono anche altre attività. Tra queste vi è l'adeguamento del dizionario agli aggiornamenti della ICD10 fino al 2015, poiché l'attuale versione è aggiornata alla Classificazione del 2009. Tale attività avrà un ulteriore impatto positivo sulla performance del sistema, in quanto gli attuali codici dell'ICD10 obsoleti che Iris non riconosce saranno sostituiti da codici validi e pertanto accettati da Iris. Sarà inoltre necessaria la revisione della codifica di alcune condizioni morbose (tumori, cause esterne) che richiedono l'utilizzo di un codice accessorio (connected codes) per il loro corretto trattamento (Iris Institute, 2014).

Parallelamente dovranno essere completate tutte le operazioni per l'integrazione del software all'interno del sistema di produzione e dovranno essere avviati gli studi di bridge coding (Istat 2011) che consentiranno di misurare l'impatto della nuova lavorazione sulle serie storiche di mortalità per causa.

Tra le motivazioni più forti, oltre a quelle già illustrate nel lavoro, che hanno spinto questo Istituto a intraprendere l'onerosa strada dell'adozione di Iris vi è sicuramente la ricerca verso il continuo miglioramento della qualità delle statistiche in termini di comparabilità e armonizzazione soprattutto in un contesto internazionale.

Sono sempre più numerosi infatti i paesi che utilizzano Iris e quelli che hanno manifestato il loro interesse verso quello che *de facto* è oggi lo standard.

Dopo l'implementazione di Iris sarà auspicabile inaugurare una nuova linea di attività volta a valutare l'incremento atteso della qualità delle statistiche italiane di mortalità per causa in termini di maggiore confrontabilità con l'Europa, ma anche con altri paesi del Mondo che utilizzano lo stesso sistema.

Glossario

ACME (Automated Coding of Medical Entities). Modulo del sistema di codifica automatico **MMDS** che automatizza il processo di selezione della **causa iniziale di morte**, avendo come input i codici **ICD10**. Si basa su **tavole di decisione**.

ACTR (Automated Coding by Text Recognition). Software generalizzato per il riconoscimento del testo sviluppato dall'Istituto di Statistica canadese. Viene attualmente utilizzato per riconoscere e codificare le espressioni mediche in italiano. ACTR consente di associare a ciascuna espressione medica in italiano un **ERN** (Entity Reference Number). Per funzionare, ACTR utilizza un **dizionario di codifica**.

Bridge coding. Codifica della **causa iniziale di morte** di uno stesso pool di dati con due sistemi diversi, per esempio utilizzando due diverse classificazioni. Consente di misurare le diversità nella causa iniziale introdotte dal cambiamento del sistema di codifica.

Causa iniziale di morte. Patologia o lesione o traumatismo individuata come quella che ha dato origine alla sequenza di patologie o lesioni o traumatismi che ha portato alla morte.

Cause multiple di morte. Tutte le patologie riportate sul **certificato di morte**.

Certificato (o scheda) di morte. Modello su cui si basa la rilevazione sulle cause di morte (Istat D4 e D4 bis) compilato dal medico curante o necroscopo per ogni decesso avvenuto sul territorio italiano. Sulla scheda di morte il medico che certifica il decesso deve, tra le varie informazioni, indicare la sequenza morbosa che ha condotto alla morte e gli eventuali altri stati morbosi rilevanti. Le informazioni di carattere demografico e sociale sono successivamente riportate dall'ufficiale di Stato civile del comune di decesso.

Classificazione internazionale delle malattie. La classificazione ICD (International Classification of Diseases) è la classificazione statistica internazionale delle malattie e dei problemi sanitari correlati, stilata dall' Organizzazione Mondiale della Sanità e utilizzata dall'Istat per la codifica delle cause di morte. L'ICD è uno standard di classificazione per gli studi statistici ed epidemiologici. È oggi alla decima edizione (**ICD10**).

Codifica delle cause di morte. Processo che consente di "tradurre" le informazioni riportate sul **certificato di morte** in codici **ICD10**.

CodSanII. Software sviluppato presso l'Istat per la codifica delle cause di morte. Include **ACTR**, il sistema di codifica automatica **MMDS**, in particolare **MICAR** e **ACME**, e le procedure per la codifica manuale assistita.

DIMDI. Istituto Tedesco di Documentazione Medica, che ospita **l'Iris Institute**.

Dizionario di codifica. Nel sistema di codifica attuale, dizionario che associa a un'espressione diagnostica un **ERN**. In Iris, invece, a ogni espressione diagnostica è associato un codice **ICD10**.

ERN (Entity Reference Number). Un ERN è un codice di 6 cifre che corrisponde a un'espressione diagnostica. Viene utilizzato dal sistema **MMDS**, in particolare dal modulo **MICAR**, nelle fasi intermedie della codifica automatica.

ICD10. Con l'acronimo ICD10 viene indicata la decima edizione della **classificazione internazionale delle malattie**.

Iris Institute. Istituto che nasce da una collaborazione internazionale per la distribuzione, la manutenzione e lo sviluppo del software Iris.

MICAR (Mortality Medical Indexing, Classification and Retrieval). Modulo del sistema di codifica automatico **MMDS** che automatizza il processo di codifica delle **cause multiple di morte**, ossia attribuisce un codice **ICD10** a ciascuna espressione medica presente sul certificato di morte (tenendo conto dell'età e del sesso del defunto, della presenza di altre condizioni, di altre informa-

zioni pertinenti presenti sul **certificato di morte**).

MMDS (Mortality Medical Data System). Sistema di codifica automatica sviluppato dal National Center for Health Statistics (NCHS) degli Stati Uniti. Include due moduli principali: **MICAR** e **ACME**.

Pretrattamento manuale del testo. Modifica manuale del testo che viene effettuata sulle espressioni diagnostiche riportate sui **certificati di morte** prima che questi vengano sottoposti al processo di codifica automatica. Viene effettuata seguendo regole ben precise ed è finalizzato a rendere interpretabile da **ACTR** il linguaggio medico naturale riportato sui certificati di morte.

Procedure di check. In questo lavoro questo termine sta ad indicare tutte le procedure statistiche messe in atto durante e dopo la fase di codifica per controllare che i codici **ICD10** della causa di morte siano compatibili con l'età e il sesso del deceduto e con le modalità di decesso indicate dal certificatore (naturali, accidentali, ecc.).

Regole di standardizzazione. Scritte utilizzando le **regular expression**, consentono di trasformare un'espressione diagnostica in un'altra "equivalente", agendo su vari aspetti: sinonimi, singolare/plurale, maschile/femminile, eliminazione di termini ininfluenti, ecc.

Regular expression (espressione regolare). Un'espressione regolare definisce una funzione che prende in ingresso una stringa e restituisce in uscita un valore del tipo sì/no, a seconda che la stringa segua o meno un certo pattern. Vengono utilizzate da Iris per la costruzione delle **regole di standardizzazione.**

Sistema di codifica automatica. Insieme di moduli che consentono di codificare i **certificati di morte** in maniera automatica.

Tavole di decisione (di) ACME. Strumento sviluppato per **ACME** dal National Center for Health Statistics (NCHS) degli Stati Uniti. Rappresentano una traduzione delle istruzioni testuali per la selezione della **causa iniziale di morte** fornite dall'Organizzazione Mondiale della Sanità in relazioni tra codici e sono riconosciute e utilizzate a livello internazionale.

Bibliografia

CDC (Center for Diseases Control and Prevention), 2014. "ICD-10 ACME Decision tables for classifying underlying causes of death, 2014, 4,361 pp." Ultima cons. 2 febbraio. http://www.cdc.gov/nchs/nvss/instruction_manuals.htm

CDC (Center for Diseases Control and Prevention), 2015. "About the mortality medical data system." Ultima cons. 2 febbraio. http://www.cdc.gov/nchs/nvss/mmds/about_mmds.htm

Cristófori Martins, R. 2012. "Codificação automática da causa de morte e seleção da causa básica: a adaptação para o Brasil do software Iris." Disertação apresentada ao programa pós-graduação, Universidade de São Paulo, Faculdade de Saúde Pública.

Eckert, O. 2014. Improvement of mortality statistics by MUSE. Poster presentato alla conferenza: European conference on quality in official statistics, Vienna, 2-5 giugno.

Harteloh P, de Bruin K, Kardaun J., 2010. The reliability of cause-of-death coding in The Netherlands. Eur J Epidemiol. 2010 Aug;25(8):531-8.

Iris Institute, 2014. "Iris user's reference manual V4.4.1S1." Ultima cons. 2 febbraio. http://www.dimdi.de/dynamic/en/klassi/irisinstitute/downloadcenter/manuals/user-guide/

Iris Institute, 2015. "About Iris." Ultima cons. 2 febbraio. http://www.dimdi.de/static/en/klassi/irisinstitute/about-iris/index.htm

Istat, 2011. Analisi del bridge coding Icd-9 - Icd-10 per le statistiche di mortalità per causa in Italia. Metodi e Norme N.50. Istat, Roma 2011.

Istat, 2013. L'indagine sulle cause di morte – Nuovo piano di controllo e correzione dei dati di mortalità per causa e fasi procedurali. Letture statistiche, metodi. Istat, Roma 2013.

Ministero della Sanità 2000. Classificazione Statistica Internazionale delle Malattie e dei Problemi Sanitari correlati dell'OMS. Decima Revisione. Istituto poligrafico e Zecca dello Stato, Roma 2000.

ONS (Office for National Statistics), 2014. "Impact of the implementation of IRIS software for ICD-10 cause of death coding on mortality statistics, England and Wales." Ultima cons. 2 febbraio. http://www.ons.gov.uk/ons/rel/subnational-health3/impact-of-the-implementation-of-iris-software-for-icd-10-cause-of-death-coding-on-mortality-statistics/england-and-wales/stb-impact-of-the-implementation-of-iris.html

Poppová, M. 2011. Iris: language-independent coding software – Implementation in Czech Republic. *Demografie, 53(4).*

Poppová, M. 2012. Changes in coding practice between 2010 and 2011 in the Czech Republic. *Demografie 54(4): 427-433.*

Weber, S. 2008. *Regular expressions. A tutorial for the standardisation of medical terms in automated coding of mortality data*. http://www.dimdi.de/dynamic/en/klassi/irisinstitute/downloadcenter/manuals/regular-expression/

Weber, S., Özer, O. 2006. "Language standardization for mortality coding. A German approach. Relazione presentata al convegno: WHO Family of International Classifications Network Meeting, Tunis (Tunisia) 29 ottobre- 4 novembre.

WHO (World Health Organization), 2015. "ICD-10 Interactive Self Learning Tool." Ultima cons. 2 febbraio. http://apps.who.int/classifications/apps/icd/icd10training/

WHO (World Health Organization), 2010. International Statistical Classification of Diseases and Related Health Problems 10th Revision. Volume 2 Instruction manual 2010 Edition. Disponibiole su http://www.who.int/classifications/icd/en/