Fecondità e maternità: un sistema integrato per la misurazione di fenomeni sanitari e socio-demografici[1]

Tiziana Tuoto, Marina Attili, Alessandra Burgio, Rossana Cotroneo, Claudia Iaccarino, Sabrina Prati, Francesca Rinesi, Fabio Rottino, Laura Tosco, Luca Valentino

Sommario

Il lavoro descrive i passi iniziali di progettazione e sperimentazione di un complesso progetto di integrazione tra fonti, finalizzato alla realizzazione del "Sistema integrato sugli esiti del concepimento". L' integrazione delle fonti è indispensabile per ottenere un quadro completo e dettagliato sui principali aspetti demografici e socio-sanitari degli esiti dei concepimenti, dati i profondi cambiamenti degli ultimi decenni nella regolamentazione sulla raccolta dei dati amministrativi, legati a questioni di semplificazione e di privacy. Nel lavoro si delinea una strategia di integrazione, mettendone in rilievo gli aspetti metodologici e prediligendo soluzioni che possano essere facilmente estese, portate a regime ed inserite in processi di produzione corrente.

Parole chiave: integrazione di dati, record linkage, indicatori demo-socio-sanitari

Abstract

This paper describes the initial steps of design and testing of a complex and ambitious integration project between sources aimed at the implementation of a system called "Integrated system on pregnancy outcome". Given the deep changes in the regulation on administrative data collection occurred in the last decades and related to simplification and privacy issues, the integration of sources is essential to obtain a complete and detailed picture on the main social, health and demographic aspects of pregnancy outcome. The paper outlines a strategy for integration, highlights the methodological issues and proposes solutions that can be easily extended and included into current production processes.

Keywords: data integration, record linkage, demographic, social and health indicators.

[1] Gli articoli pubblicati impegnano esclusivamente gli Autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat. Sebbene il lavoro sia frutto dell'opera di tutti gli autori, sono da attribuire: i paragrafi 1, 3, 3.2, 4, 5 a Tiziana Tuoto; il paragrafo 2, 2.2, 2.3, 2.4 a Alessandra Burgio e Sabrina Prati, il paragrafo 2.1 a Marina Attili Alessandra Burgio, Rossana Cotroneo, Claudia Iaccarino, Tiziana Tuoto; il paragrafo 3.1 a Laura Tosco e Tiziana Tuoto; il paragrafo 4.1 a Alessandra Burgio, Rossana Cotroneo, Francesca Rinesi, Laura Tosco, Tiziana Tuoto e Luca Valentino, il paragrafo 4.2 a Marina Attili, Fabio Rottino, Claudia Iaccarino; il paragrafo 4.3 a Marina Attili, Claudia Iaccarino e Tiziana Tuoto. Per contattare gli autori: tuoto@istat.it.

1. Introduzione

L'integrazione di dati provenienti da fonti amministrative e/o da indagini statistiche è diventata negli ultimi anni una priorità per la statistica ufficiale: essa permette di migliorare la qualità dei dati e l'efficienza delle analisi attraverso un poderoso utilizzo di dati di fonte amministrativa e quindi disponibili senza ulteriori carichi economici e riducendo il carico statistico sui rispondenti. Tra le tecniche di integrazione di dati, quella del record linkage ha come obiettivo l'identificazione della stessa unità statistica, tipicamente memorizzata in archivi diversi e descritta con chiavi identificative non perfettamente coincidenti. Le soluzioni ai problemi di record linkage, studiate in letteratura e adottate nella pratica, si rifanno a svariati approcci e metodologie, che coinvolgono soluzioni euristiche, metodi probabilistici, approcci bayesiani, soluzioni basate sulle tecniche di data-mining e/o machine learning. Tuttavia i problemi di record linkage sono fortemente caratterizzati dalla natura dei dati da abbinare e dagli obiettivi dell'abbinamento e nessuna delle metodologie o delle tecniche proposte è la più efficace o la più efficiente, per tutte le diverse applicazioni, ma è necessario adattare il processo di linkage agli specifici requisiti dei dati in esame.

Il presente lavoro descrive un complesso progetto di integrazione tra fonti, finalizzato alla realizzazione di un "Sistema integrato sugli esiti del concepimento" il cui obiettivo è quello di misurare i principali aspetti socio-demografici e sanitari degli esiti dei concepimenti. Tale integrazione permetterà di superare il gap informativo venutosi a creare alla fine degli anni '90 con l'entrata in vigore delle leggi sulla semplificazione amministrativa e sulla privacy, che hanno generato profondi cambiamenti nella raccolta dei dati per le statistiche sanitarie e socio-demografiche. Solo attraverso l'integrazione di fonti di dati diverse è ora possibile fornire una visione completa e allo stesso tempo dettagliata dei fenomeni legati alla maternità e alla fertilità. Tale operazione però presenta diversi elementi di complessità, legati essenzialmente a due fattori: in primo luogo la disomogeneità delle fonti coinvolte, che sono di diversa natura (amministrativa, sanitaria, statistica), fanno riferimento a universi solo parzialmente sovrapposti, riportano dati raccolti con differente livello di accuratezza; in secondo luogo la mancanza, per la tutela della privacy, di informazioni "forti" per l'identificazione degli individui è un fattore estremamente rilevante per l'applicabilità stessa e l'efficacia delle tecniche di integrazione. In ogni caso, l'operazione di integrazione e di creazione di questo sistema integrato è irrinunciabile per la determinazione di misure sulle caratteristiche delle donne in gravidanza, sulla salute delle donne e dei bambini durante la gravidanza e nel periodo post-parto e sulla salute delle madri e dei bambini. In particolare, il sistema integrato permetterà di produrre indicatori sul tipo di parto, sui parti pretermine e multipli, sulla distribuzione del peso alla nascita e dell'età gestazionale. Il sistema renderà possibile studiare le principali relazioni tra tali fenomeni e la relazione tra questi e le grandezze che li determinano (stato civile, cittadinanza, livello d'istruzione e condizione lavorativa).

Il presente contributo è organizzato come segue: nel paragrafo 2 sono descritte le esigenze conoscitive legate al sistema integrato degli esiti dei concepimenti e definiti alcuni dei principali output attesi, insieme ad una sintetica ricognizione delle fonti e della loro qualità; nel paragrafo 3 vengono fornite la descrizione del sistema integrato, la definizione del fenomeno di riferimento adottata nel resto dell'articolo e le scelte alla base delle metodologie di integrazione sperimentate, evidenziando i requisiti della procedura di abbinamento e lo strumento utilizzato; il paragrafo 4 riporta i primi incoraggianti risultati

ottenuti con alcune sperimentazioni effettuate sui dati del 2007 per la regione Emilia-Romagna. Infine, nel paragrafo 5 sono raccolte alcune conclusioni, si prospettano i passi operativi e futuri indirizzi di ricerca.

2. L'esigenza conoscitiva

Il principale obiettivo del Sistema integrato sugli esiti dei concepimenti (SIEC) è consentire la costruzione di una molteplicità di indicatori su parti, nascite, interruzioni di gravidanza, mortalità perinatale mediante l'integrazione delle fonti demografiche e sanitarie. L'integrazione delle fonti permetterà di disporre di una base dati affidabile e di buona qualità attraverso la validazione e la eventuale correzione delle informazioni disponibili nei singoli flussi. Gli output dell'integrazione sono molteplici e differenziati a seconda che si assuma come unità di analisi la donna, il parto o l'esito del concepimento (nato vivo, nato morto, interruzione volontaria di gravidanza -IVG- e aborto spontaneo -AS), oppure un'ottica trasversale (più fonti con uno stesso riferimento temporale) o longitudinale (la stessa fonte in periodi diversi) per "seguire" nel tempo le storie riproduttive di differenti coorti di donne. Gli output del sistema possono essere inoltre classificati secondo una logica prevalentemente di processo o di prodotto.

Per output di processo si intende la possibilità di migliorare la qualità e la completezza delle informazioni rilevate in ciascuna fonte, nonché di armonizzare e rendere compatibili le informazioni desumibili per le stesse unità da diversi flussi produttivi. Esiste una gerarchia del livello di qualità delle informazioni che vengono rilevate nelle diverse fonti amministrative. In ogni fonte la qualità è più alta per quelle informazioni che sono disponibili nei registri di base e/o che sono necessarie per le finalità istituzionali del titolare della fonte. Ad esempio, per le rilevazioni di fonte anagrafica la qualità e la copertura delle informazioni demografiche desumibili direttamente dai registri di popolazione (data di nascita, stato civile, cittadinanza, luogo di nascita e di residenza) è massima. Nelle fonti amministrative sanitarie, ad esempio le SDO, la qualità è massima per tutte le variabili che descrivono il ricovero (età e sesso, diagnosi alla dimissione, procedure o interventi effettuati, tipo di ricovero, ecc.). Nel caso dei dati sulle interruzioni volontarie di gravidanza e sull'abortività spontanea sono di elevata qualità tutte le informazioni di carattere demografico e socio-sanitario che vengono desunte direttamente dalla cartella clinica.

La qualità è inoltre variabile nel tempo e nello spazio. Esemplare è a questo proposito il caso della rilevazione sui Certificati di assistenza al parto (CEDAP). La rilevazione ha avuto un percorso faticoso, a distanza di 10 anni dall'inizio dell'acquisizione di questa preziosa fonte informativa si registra ancora una situazione molto eterogenea sul territorio, con punte di eccellenza in alcune regioni e situazioni meno soddisfacenti in altre.

Per output di prodotto si intende tutto ciò che è rilasciato, o rilasciabile, all'utenza esterna tramite l'accesso al sistema. Nei successivi sottoparagrafi 2.2, 2.3 e 2.4 si evidenzieranno alcuni output di prodotto propri dello sfruttamento integrato delle fonti reso possibile dal sistema e quindi aggiuntivi rispetto a quelli già elaborabili attingendo ad ogni singola fonte.

2.1 Ricognizione delle fonti

Affinché i risultati di un processo di integrazione siano di buona qualità, è necessario prevedere una preliminare e approfondita analisi dei dati provenienti dalle distinte fonti che si intende integrare. Infatti, per scegliere in modo efficiente le informazioni da utilizzare ai fini del record linkage, è fondamentale conoscere la natura dei dati, il numero di record da trattare, la presenza in essi di identificatori univoci e di variabili con alto potere discriminante, la presenza più o meno consistente di dati mancanti e/o con codifiche errate.

In prima istanza sono state considerate le seguenti fonti di dati, costituite da indagini, archivi amministrativi o registri di varia natura:

CEDAP – la rilevazione sui certificati di assistenza al parto (distinguendo ove possibile tra parti, nati vivi, nati morti);

P4 – la rilevazione Istat degli iscritti in anagrafe per nascita;

IVG – l'indagine Istat sulle interruzioni volontarie di gravidanza;

AS - l'indagine Istat sugli aborti spontanei;

SDO - le schede di dimissione ospedaliera (distinguendo parti, nati vivi, nati morti, interruzioni volontarie di gravidanza, aborti spontanei).

Le tavole successive riportano i principali risultati della ricognizione per le fonti considerate a questo stadio di progettazione del sistema. Nella tavola 2.1 per ogni fonte è indicato se l'Istat è l'ente titolare dei dati, mentre la successiva tavola 2.2 specifica il periodo di riferimento, la dimensione in termini di numero di eventi registrati in Italia, la presenza e la qualità di variabili identificative per ciascuna fonte di dati. Per motivi di privacy, le fonti sanitarie giungono all'Istat completamente prive degli identificativi delle persone coinvolte dall'evento.

Tavola 2.1 - Tipo di fonte ed ente gestore

	Nome della fonte	Titolarità dell'Istat	
		Si	No
1a	CEDAP parto		X
1b	CEDAP nato vivo		X
1c	CEDAP nato morto		X
3	P4	X	
4	IVG	X	
5	AS	X	
6 a	SDO parto		X
6 b	SDO nato vivo		X
6 c	SDO nato morto		X
6 d	SDO ivg		X
6 e	SDO as		X

Tavola 2.2 - Numerosità e qualità dell'informazione

	Nome della fonte	Anno di riferimento	Numero di unità in Italia (annuali approx)	Presenza di variabili identificative (*)	Qualità (**) delle variabili disponibili	
					Bassa	Buona
1a	CEDAP parti	2002-2007	520.000	NO	X	
1b	CEDAP nati vivi	2002-2007	525.000	NO	X	
1c	CEDAP nati morti	2002-2007	1.500	NO	X	
3	P4	1999-2009	570.000	SI		X
4	IVG	1982-2009	125.000	NO		X
5	AS	1982-2009	77.000	NO		X
6a	SDO parti	2001-2009	550.000	NO		X
6b	SDO nati vivi	2001-2009	560.000	NO		X
6c	SDO nati morti	2001-2009	1.700	NO		X
6d	SDO ivg	2001-2009	120.000	NO		X
6e	SDO as	2001-2009	85.000	NO		X

Note: (*) per variabili identificative di intendono i Codici Fiscali, nomi e cognomi (eventualmente standardizzati o codificati), codici univoci di evento …

(**) per qualità delle variabili si fa riferimento a valori mancanti, errori nella codifica, incompatibilità con altre variabili.

2.2 Linkage CEDAP e iscritti in anagrafe per nascita (P4)

L'integrazione tra CEDAP e iscritti in anagrafe per nascita consente, per il sottoinsieme di record comuni, ovvero i nati vivi della popolazione residente, di validare le informazioni demografiche di base del nato e dei genitori (output di processo). Un esempio è il caso della cittadinanza, informazione che in termini di qualità e completezza è maggiore nella rilevazione di fonte anagrafica.

Dal lato degli output di prodotto, invece, l'integrazione tra le due fonti consente di recuperare completamente il debito informativo su alcune relazioni tra nascite e parti creatosi a partire dal 1999 con la soppressione della rilevazione individuale dei nati effettuata dall'Istat fin dal 1926 presso lo stato civile. Infatti, mentre nelle attese del legislatore il contenuto informativo dei CEDAP avrebbe dovuto sostituire completamente le rilevazioni soppresse, le difficoltà di fatto incontrate nella rilevazione dei CEDAP, di cui si accenna al paragrafo precedente, hanno reso lacunoso il patrimonio informativo acquisito.

E' pertanto possibile elaborare e diffondere con periodicità annuale i principali indicatori sulla gravidanza, il parto e le caratteristiche dei nati della popolazione residente, distinti per le principali caratteristiche demografiche e sociali delle madri (stato civile, cittadinanza, titolo di studio, condizione professionale).

E' inoltre possibile mettere a disposizione dell'utenza una base di micro dati validati, priva di elementi che consentano di risalire agli individui (file standard o file per la ricerca) che fornisce agli studiosi strumenti per analizzare gli effetti delle principali determinanti socio-demografiche e sanitarie rispetto a diversi tipi di esiti (nato vivo, nato morto) oppure nati a rischio (fortemente pre-termine, fortemente sottopeso), nascite gemellari, nascite da parti non fisiologici (in particolare parti cesarei).

2.3 Linkage CEDAP e SDO

Le due rilevazioni (CEDAP e SDO), per quanto concerne l'evento parto, afferiscono alla stessa popolazione. In alcune regioni esiste una corrispondenza uno a uno in quanto se non viene compilato il CEDAP non viene rimborsato, o viene rimborsato in misura minore, il ricovero in caso di parto. L'integrazione può essere effettuata assumendo come unità di analisi il parto oppure il nato vivo. Il primo output di processo di rilievo dell'integrazione

SDO-CEDAP consiste nella possibilità di ricondurre la SDO parto alla SDO nato (attualmente i due flussi sono separati e non riconducibili allo stesso evento), ampliando così notevolmente il ventaglio informativo.

Si possono inoltre sfruttare tutte le informazioni delle SDO e dei CEDAP per effettuare controlli di coerenza tra le diverse fonti e correggere situazioni di incompatibilità dovute ad errori o scarsa qualità: si possono utilizzare le due fonti integrate sia per recuperare eventuali mancate risposte parziali, sia per recuperare le mancate risposte totali utilizzando le SDO come universo di riferimento.

Dal lato degli output di prodotto, l'integrazione permette di calcolare i principali indicatori sui parti. Infatti, le SDO arricchiscono i CEDAP con informazioni di tipo clinico (diagnosi principale e secondarie, procedure diagnostiche, interventi principali e secondari, informazioni sul ricovero e la dimissione) che permettono di studiare in maniera più dettagliata la medicalizzazione del percorso gravidanza-parto nonché di evidenziare l'adozione dei diversi protocolli di assistenza nei punti nascita in relazione alle principali caratteristiche dei centri nascita.

Di particolare rilievo sono le possibilità offerte dal sistema per lo studio dei parti cesarei. Ad esempio, è possibile procedere al calcolo delle classi di Robson[2] a livello regionale o sub-regionale per evidenziare le aree di maggiore criticità rispetto all'eccessivo ricorso al cesareo e metterle in relazione con le informazioni di contesto sul centro nascita.

Inoltre l'integrazione SDO-CEDAP, insieme ad una terza fonte di cui è titolare il Ministero della Salute e che si chiama "Struttura e attività degli Istituti di cura" permette di disporre di informazioni di contesto sul centro nascita, per sapere quanti posti letto sono disponibili nel reparto di ostetricia-ginecologia e quante culle nei nidi. Un tema estremamente attuale se si pensa alle chiusure previste per i centri di nascita che effettuano un numero di parti annui al di sotto di una soglia prefissata. nonché informazioni relative alla struttura dove è stato effettuato il ricovero ospedaliero.

Inoltre, l'integrazione permetterà di calcolare i principali indicatori sulla salute perinatale: ad esempio, gli indicatori raccomandati dal Progetto Europeristat (European Perinatal Health, 2008)., che consentono il confronto con gli altri Paesi europei, potranno essere calcolati anche a livello regionale e sub-regionale con periodicità annuale in modo da monitorare il fenomeno della salute riproduttiva sul territorio secondo quanto è richiesto dalle raccomandazioni internazionali.

Sarà infine possibile creare una base dati individuale anonima che consenta agli esperti che operano nel settore e agli studiosi di analizzare le relazioni tra le principali caratteristiche individuali e di contesto e l'evento nascita con i suoi possibili esiti.

2.4 Linkage SDO e IVG, SDO e AS

Il principale output di processo derivante dall'integrazione tra SDO-IVG-AS è la possibilità di validare alcune informazioni socio-demografiche di rilievo per l'analisi dei

[2] Questa classificazione, prendendo in esame la precedente storia ostetrica (parità), l'età gestazionale, la presentazione e le modalità del travaglio, propone 10 categorie mutuamente esclusive che quantificano 10 sottopopolazioni, entro ciascuna delle quali è possibile analizzare la frequenza di ricorso al parto cesareo (Can we reduce the caesarean section rate? Michael Stephen Robson, Best Practice & Research Clinical Obstetrics & Gynaecology Vol. 15, No. 1, pp. 179-194, 2001Harcourt Publishers Ltd).

fenomeni come lo stato civile e il titolo di studio, rilevate da tempo e affidabili nelle fonti Istat sull'abortività; viceversa per la fonte SDO sono da considerarsi di buona qualità le informazioni sul ricovero (in regime ordinario con pernottamento o in day hospital senza pernottamento, durata della degenza e intervento effettuato).

La fonte SDO viene attualmente utilizzata come universo di riferimento per la correzione della mancate risposte totali presenti nei flussi IVG e AS. Con l'integrazione delle fonti in esame sarà possibile analizzare i protocolli utilizzati negli ospedali per codificare l'IVG e l'AS (analisi della diagnosi principale e delle diagnosi secondarie, dell'intervento/procedura principale e degli interventi/procedure secondari). Questo può fornire dei criteri più mirati per selezionare i casi di IVG e AS nelle SDO e quindi migliorare la costruzione dell'universo di riferimento per la stima dei dati mancanti.

In termini di output di prodotto, l'integrazione permette di arricchire e dettagliare le informazioni medico-sanitarie rilevate con i moduli Istat e calcolare indicatori sulle caratteristiche del ricovero per le donne che subiscono una interruzione della gravidanza o un aborto spontaneo: interventi, terapie e procedure effettuate, presenza di eventuali complicazioni durante il ricovero.

3. Le metodologie di integrazione per il sistema integrato sugli esiti dei concepimenti

Per la realizzazione del prototipo del sistema integrato sugli esiti dei concepimenti, è stata individuata come unità di riferimento l'evento "esito del concepimento": per tale fenomeno il sistema è in grado di distinguere le seguenti modalità

- nato vivo;
- nato morto;
- aborto spontaneo;
- interruzione volontaria di gravidanza.

In questa prima fase di progettazione e sperimentazione della fattibilità dell'intero sistema, eventi riferiti a concepimenti diversi che coinvolgono la stessa donna, nell'arco dello stesso anno, sono distinti e non immediatamente riconducibili. Si è deciso quindi di costruire il sistema integrato coinvolgendo principalmente le basi di dati riguardanti i certificati di assistenza al parto (CEDAP), distinguendo ove possibile tra parti, nati vivi, nati morti; gli iscritti in anagrafe per nascita (P4); le indagini Istat sulle interruzioni volontarie di gravidanza e sugli aborti spontanei (IVG); le schede di dimissione ospedaliera (SDO), distinguendo le interruzioni volontarie di gravidanza, gli aborti spontanei, i parti, i nati vivi e i nati morti. Come esposto nel paragrafo 2, è possibile ampliare le potenzialità conoscitive del sistema collegando eventi che vedono protagonista la stessa donna, anche in tempi diversi; a tal fine è necessario un approfondimento successivo che studi la possibilità di agganciare eventi diversi riferiti alla stessa donna, anche in un approccio longitudinale, per la ricostruzione della vita riproduttiva.

Il sistema sugli esisti del concepimento può essere definito secondo diversi livelli di integrazione, ad esempio è possibile prevedere integrazioni di tipo macro (a livello di indicatori e di aggregati) o di tipo micro (a livello di singolo record per ogni evento considerato). L'integrazione a livello micro, laddove possibile, è quella che salvaguarda il patrimonio informativo più ampio, quindi si è deciso di privilegiare inizialmente questo tipo

di metodologia, mettendo in campo tutte le risorse e le competenze per valutare la fattibilità del sistema in cui ogni fonte è integrata a livello del singolo record. Opportune valutazioni relative ad integrazioni di tipo macro e ad eventuali interazioni tra le due operazioni possono essere ulteriormente prese in considerazione.

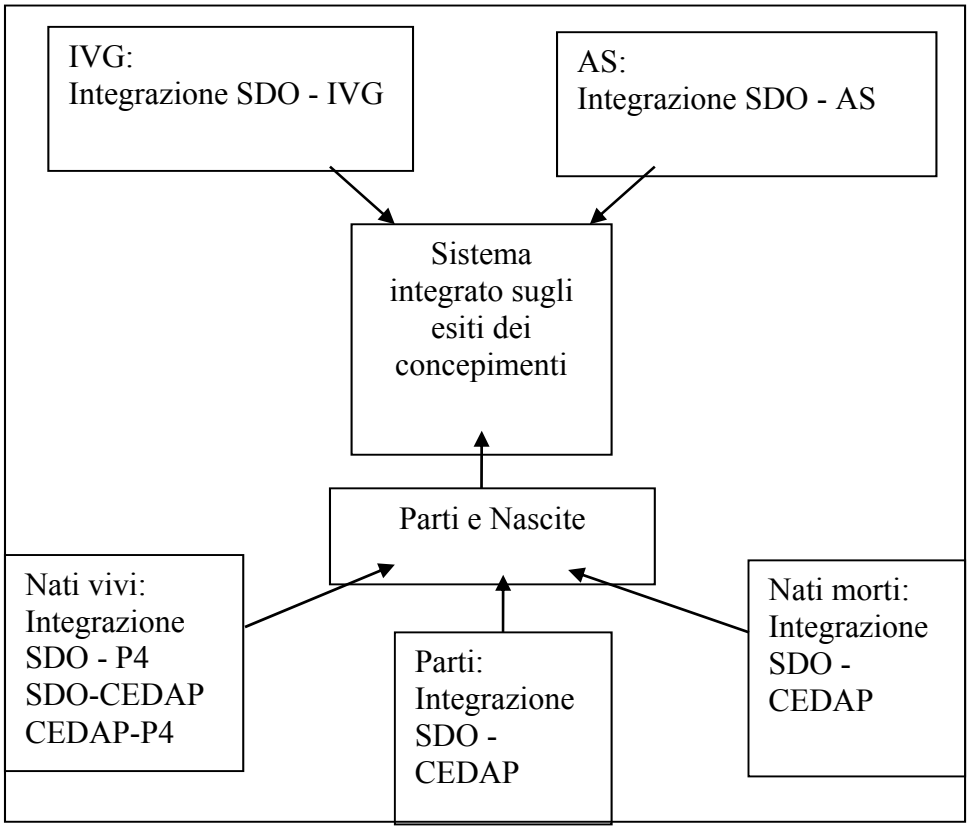
Nella scelta delle metodologie da testare per lo studio di fattibilità e la realizzazione, almeno in fase prototipale, del sistema integrato sugli esiti del concepimento, si persegue l'obiettivo di mettere a punto una strategia di integrazione che possa essere facilmente estesa, portata a regime ed inserita in un processo di produzione corrente.

Nell'ambito delle metodologie per l'integrazione dei microdati, il record linkage probabilistico permette di corredare il sistema integrato con una serie di informazioni e misurazioni relative al processo di integrazione che i dati hanno subito, tra cui ad esempio la probabilità di corretto abbinamento per ogni singolo record considerato nel sistema e altre misure di qualità complessive relative all'intero processo di integrazione. Le applicazioni di record linkage devono essere corredate da informazioni sulla qualità del linkage da utilizzare con apposite metodologie di stima, volte ad assicurare la qualità delle analisi condotte sui dati abbinati. Infatti, trattare tali dati ignorando il processo di integrazione e gli eventuali errori da questo generati, può portare, in generale, a stime distorte. Tali informazioni sono di fondamentale importanza per la corretta analisi e interpretazione statistica dei fenomeni legati alla fecondità e alla maternità che attraverso il sistema si vogliono studiare.

Con l'ottica di risolvere il problema dell'integrazione delle fonti relative agli esiti dei concepimenti in modi facilmente riproducibili in contesti di produzione dei dati, è stato utilizzato il toolkit RELAIS (RELAIS, 2011), come strumento privilegiato per la realizzazione dei processi. Tale strumento mette a disposizione una serie di metodi e tecniche diverse per l'esecuzione di ogni singola fase di un processo di record linkage. Lo strumento è quindi particolarmente indicato per la sperimentazione e realizzazione di progetti di integrazione che coinvolgono fonti di dati così diverse per quantità e qualità, e quindi per la creazione di un sistema complesso e articolato come quello integrato per gli esiti del concepimento. In fase sperimentale, una ulteriore valutazione della robustezza dei risultati dell'abbinamento probabilistico è stata effettuata attraverso il confronto con i risultati di una procedura di abbinamento deterministico, (descritta nel successivo paragrafo 4.2) che è stata sviluppata *ad hoc* nel contesto delle attività di abbinamento deterministico di dati demo-sociali. Tale procedura è stata arricchita e perfezionata nel corso degli anni e ha dimostrato di perseguire ottimi risultati, testati soprattutto in situazioni analoghe (Attili e Valentino, 2010). In ogni caso, per la costruzione del sistema integrato degli esiti dei concepimenti è più opportuno l'utilizzo di strumenti di integrazione che siano basati su metodologie statistiche ben consolidate e validate dalla comunità scientifica.

In questa fase sperimentale finalizzata alle valutazioni della fattibilità del sistema integrato, tutte le fonti che riportano informazioni relative allo stesso fenomeno-evento sono state abbinate a coppie e, successivamente, sono stati considerati i risultati di ogni integrazione per individuare sia l'insieme minimo che quello massimo di eventi abbinati. Queste operazioni, seppure più dispendiose rispetto ad abbinamenti successivi in sequenza o a cascata, permettono di tenere sotto controllo il potere informativo di ogni fonte e predisporre le operazioni successive avvalendosi del massimo dell'informazione possibile. Il progetto di integrazione per la costruzione a livello micro del prototipo del sistema integrato dei concepimenti prevede i seguenti abbinamenti come riportati nello schema 3.1.

Schema 3.1. Sintesi delle integrazioni tra fonti per la costruzione del sistema sugli esiti dei concepimenti



Operativamente, seguire questo schema significa procedere parallelamente con gli abbinamenti:

1. SDO – IVG, il cui risultato intende costruire la parte del sistema integrato relativa alle interruzioni volontarie di gravidanza
2. SDO – AS, il cui risultato intende costruire la parte relativa agli aborti spontanei per cui è stato necessario rivolgersi ad una struttura ospedaliera
3. SDO - P4, il cui risultato intende costruire la parte del sistema integrato relativa ai nati vivi
4. CEDAP - P4, il cui risultato intende costruire la parte relativa ai nati vivi
5. SDO - CEDAP, il cui risultato intende costruire la parte relativa ai parti, ai nati vivi e ai nati morti.

I risultati degli abbinamenti 3), 4) e 5) andranno a loro volta integrati e analizzati per verificare:

- i casi linkati in tutte le applicazioni
- i casi linkati solo da una o due applicazioni

L'integrazione dei risultati parziali 3), 4) e 5) permetterà di ricostruire l'archivio dei parti e delle nascite e a sua volta andrà integrato con i risultati 1) e 2) per costituire il sistema integrato sugli esiti dei concepimenti.

3.1 Il record linkage e lo strumento RELAIS

Come noto, il record linkage indica un processo di abbinamento di record che ha come obiettivo l'identificazione della stessa unità statistica, memorizzata in archivi diversi o presente più volte nella stessa lista, anche in assenza di identificatori univoci o quando questi sono affetti da errori. L'identificazione dell'unità in archivi di diversa natura avviene attraverso chiavi comuni, presenti nei vari file; le chiavi possono essere anche non perfettamente corrispondenti. La complessità del record linkage dipende da molteplici aspetti, principalmente legati all'assenza di identificatori univoci o alla presenza di errori negli identificatori stessi.

Formalmente, l'obiettivo del linkage è identificare un'unità che può essere rappresentata in maniera differente in due diverse fonti dati A e B. In generale, le coppie che si intende classificare come abbinamenti (ossia a e b sono la stessa unità) o non abbinamenti (a e b sono due differenti unità) sono quelle dell'insieme Ω, prodotto cartesiano di A e B. Tale insieme ha cardinalità $n_A \times n_B$ ed è costituito da tutte le possibili coppie a, b (a,b | a $\in$ A, b $\in$ B). Per individuare le coppie che si riferiscono alla stessa unità, gli abbinamenti, si ricorre al confronto tra k variabili, "variabili di match", comuni alle due fonti di dati e associate alle unità. Tali variabili identificano in maniera univoca le unità, a meno, ovviamente, di errori o valori mancanti nelle variabili stesse; proprio a causa delle imperfezioni nelle variabili di match, l'abbinamento non può essere risolto attraverso l'utilizzo di un semplice "join" fra le due liste in esame. Il confronto tra le variabili viene effettuato per mezzo di un'opportuna funzione, scelta in base al tipo di variabile e alla sua qualità (in termini di completezza e correttezza). Per ogni coppia (a,b) $\in \Omega$, si definisce un vettore γ, detto "vettore dei confronti", i cui k elementi sono il risultato del confronto tra le variabili di match. Nel modello probabilistico per l'individuazione degli abbinamenti, si ipotizza che la distribuzione del vettore dei confronti sia una mistura di due distribuzioni, una generata dalle coppie (a,b) che effettivamente rappresentano la stessa unità, distribuzione *m*, e una generata dalle coppie (a,b) che rappresentano unità diverse, distribuzione *u*. A partire dalla stima di tali distribuzioni, è possibile costruire il peso composto di abbinamento (Fellegi and Sunter, 1969), dato dal rapporto delle verosimiglianze

$$r = \frac{m(\gamma)}{u(\gamma)} = \frac{\Pr(\gamma \mid M)}{\Pr(\gamma \mid U)},$$

dove *M* è l'insieme delle coppie che rappresentano degli abbinamenti e *U* è l'insieme delle coppie che non rappresentano degli abbinamenti, con $M \cup U = \Omega$ e $M \cap U = \varnothing$. In generale, la stima dei parametri delle distribuzioni viene generalmente ottenuta per mezzo

dell'applicazione dell'algoritmo EM (Jaro 1989). La stima dei parametri delle distribuzioni risulta molto complessa o addirittura impraticabile quando la dimensione dello spazio dei confronti Ω è dell'ordine dei milioni. A tali dimensioni si arriva rapidamente dato che lo spazio Ω, creato come prodotto cartesiano dei file da abbinare, cresce in maniera quadratica rispetto alla dimensione dei file di partenza. In tali casi si procede usualmente alla riduzione dello spazio di ricerca delle coppie attraverso l'applicazione di metodi di bloccaggio, sorting o indexing (Cibella, Tuoto, 2012).

Sulla base del rapporto r, le coppie sono ordinate e sottoposte ad un processo di classificazione negli insiemi M ed U:

- se il peso r è maggiore di una certa soglia T_m allora la coppia viene classificata come match;
- quando il suo peso è inferiore alla soglia T_u la coppia viene classificata come non match;
- per le unità il cui peso cade nell'intervallo $I=(T_u , T_m)$ non è possibile stabilire lo stato di abbinamento ma è necessario procedere ad un'ispezione manuale o comunque ad ulteriori analisi.

Secondo lo schema di decisione impostato da Fellegi e Sunter le due soglie, T_u e $T_{m,}$, sono fissate in modo che siano minimizzati sia gli errori di classificazione che la dimensione dell'area tra le soglie per cui non viene presa una decisione.

In numerose applicazioni, attraverso il linkage si mira ad individuare tra le coppie solo legami del tipo 1 a 1, in cui, cioè, una unità del file A viene abbinata con una sola unità del file B; in questi casi è necessario introdurre metodi di ottimizzazione che consentano di selezionare, tra tutte le coppie che coinvolgono le stesse unità della lista A e della lista B, quelle che rispettano il vincolo 1:1 e massimizzano la somma dei pesi r.

RELAIS (REcord Linkage At IStat) è un toolkit sviluppato presso l'Istat che mette a disposizione un insieme di tecniche per affrontare e risolvere problemi di record linkage (Cibella *et al.* 2007). RELAIS si basa sull'idea che un processo di record linkage può essere visto come costituito da diverse fasi per ognuna delle quali possono essere adottate diverse tecniche risolutive afferenti a diverse aree di conoscenza. La scelta della tecnica più appropriata da applicare dipende dal dominio di applicazione. RELAIS fornisce diverse tecniche per le diverse fasi di un processo di record linkage, consentendo di combinare tali tecniche in modo da ottenere il processo lavorativo ottimale per la specifica applicazione.

Nella costruzione del sistema integrato dei concepimenti è stata usata la versione 2.3 beta di RELAIS. Le caratteristiche specifiche del sistema possono essere trovate nel manuale utente, disponibile all' indirizzo http://www.istat.it/it/strumenti/metodi-e-software/software/relais. Rispetto alla versione 2.2, la versione 2.3 beta è stata rilasciata in maniera informale specificatamente per le attività di questo lavoro, poiché per le caratteristiche specifiche dei dati in esame è stata arricchita con la funzione di distanza, denominata window equality, ideata in particolare per il confronto tra numeri interi e definita secondo la regola: se (|x-y| <= w) i due numeri sono considerati uguali altrimenti i due numeri sono considerati diversi, dove w è la dimensione della finestra definita dall'utente e x e y sono i due numeri da confrontare.

3.2 I requisiti della procedura di abbinamento

La realizzazione del sistema integrato sugli esiti del concepimento ha l'obiettivo principale di permettere di misurare i più importanti aspetti socio-sanitari degli esiti dei

concepimenti attraverso un insieme di indicatori. Ciò impone che la base di dati di riferimento sia la più ampia possibile e che non presenti "distorsioni" rispetto alle fonti originarie che la compongono. In termini di linkage, questi requisiti si traducono innanzitutto in un "alto" livello del match rate (o tasso di abbinamento), definito, nel caso dell'abbinamento tra due fonti, come il rapporto tra numero dei record abbinati e il numero dei record presenti nella più piccola delle fonti considerate[3]. Per questo motivo, particolare attenzione in questo studio di fattibilità deve essere prestato alle sperimentazioni relative all'abbinamento delle fonti di dimensioni più contenute, IVG e AS e al sottoinsieme dei nati morti, poiché, in questi casi, un basso valore del match rate potrebbe comportare a livello integrato una eccessiva riduzione delle osservazioni relative al fenomeno in esame.

D'altro lato, a patto di raggiungere un "buon" livello del match rate, è possibile valutare meno severamente l'eventuale perdita di ulteriori veri link, a meno che questi non siano caratterizzati in modo tale da rendere distorte le successive analisi sui dati integrati. Ciò significa poter tenere in diversa considerazione gli errori di falso abbinamento e quelli di mancato abbinamento[4]: i primi saranno considerati più gravi, in quanto introducono eventi non esistenti nei dati, mentre i mancati abbinamenti possono essere considerati meno gravi, anche se comunque da evitare, sotto l'ipotesi, da verificare, che non comportino distorsione nei risultati, ossia che i mancati abbinamenti non si concentrino in maniera evidente in particolari categorie o sotto-popolazioni. Le procedure di abbinamento, sperimentate e descritte nel seguito, hanno quindi posto vincoli più stringenti sul tasso di falso abbinamento e meno restrittivi sul tasso di mancato abbinamento. Per analizzare l'eventuale introduzione di un effetto distorsivo dovuto al mancato abbinamento di alcuni record, gli esiti dell'abbinamento probabilistico sono comparati a livello di singola coppia individuata con le risultanze dell'abbinamento deterministico proposto nel successivo paragrafo 4.2; inoltre verranno comunque confrontate le distribuzioni di frequenza dei dati abbinati e di quelli di partenza rispetto alle principali variabili di interesse per la costruzione di indicatori socio-sanitari.

[3] Il tasso di abbinamento o match rate, che è una delle misure di qualità di un processo di linkage, è propriamente definito come il rapporto tra il numero totale di record abbinati e il numero vero (ignoto) di abbinamenti. Tale indicatore viene usualmente calcolato sostituendo al totale ignoto di veri abbinamenti la sua stima fornita dal modello o la dimensione del più piccolo tra i file da abbinare, che costituisce comunque il massimo degli abbinamenti possibili, sotto l'ipotesi che tutti i record debbano essere abbinati. In questo modo si fornisce una misura del minimo del tasso di abbinamento, ed è questa la valutazione cautelativa della procedura di linkage che è stata adottata nel seguito del presente lavoro.

[4] Gli errori di classificazione nel modello di linkage proposto da Fellegi e Sunter sono di due tipi: gli abbinamenti errati (o falsi abbinamenti – false matches nella più diffusa terminologia inglese), quando vengono abbinate unità che corrispondono a entità differenti e gli abbinamenti mancati (false non-matches), quando record corrispondenti ad una stessa entità non vengono abbinati. In generale, gli abbinamenti errati si suddividono, a loro volta, in: accoppiamenti tra due unità che non dovrebbero essere abbinate tra loro ma con altri record e accoppiamenti tra unità che non dovrebbero essere abbinate affatto. Gli errori di abbinamento, sia abbinamenti errati che abbinamenti mancati, giocano un ruolo fondamentale per la valutazione della bontà dei risultati delle procedure di linkage e devono essere tenuti nella massima considerazione nelle successive analisi sui dati linkati, in quanto possono influire significativamente su di esse. Misure sintetiche della qualità del linkage, basate su tali errori, sono i tassi di mancato abbinamento e di falso abbinamento, definiti, il primo, come il rapporto tra numero stimato di mancati abbinamenti e il totale stimato di veri abbinamenti e, il secondo come il rapporto tra il numero stimato di falsi abbinamenti e il totale di abbinamenti individuati.

4. Primi risultati: una sperimentazione sull'Emilia Romagna

La dimensione del fenomeno oggetto di studio è rilevante in termini di numero di individui coinvolti. Come riportato nella tavola 2.2 del paragrafo 2, alcune delle fonti principali coinvolgono un numero di record molto elevato (circa 500 mila unità, annualmente a livello Italia). Le analisi e le sperimentazioni per lo studio della fattibilità sono state avviate sulla regione Emilia Romagna, come riferimento territoriale per gli eventi da considerare. La selezione di questa regione è legata principalmente a due fattori:

- la numerosità degli eventi presi in esame è cospicua, in Emilia Romagna sono concentrati il 10% circa degli eventi nazionali relativi alle nascite;
- la buona qualità delle fonti disponibili, per quanto riguarda i dati sanitari che arrivano all'Istat dopo essere stati raccolti a livello regionale.

La scelta di restringere, in fase sperimentale, il campo di osservazione ad un numero ridotto di casi tende a far crescere il potere identificativo di variabili come le date di nascita per quanto riguarda gli individui e le date relative agli eventi, che per alcune fonti sono le informazioni con più alto potere identificativo disponibili.

Per quanto riguarda il riferimento temporale, la sperimentazione del sistema integrato è stata avviata selezionando come anno di riferimento degli eventi il 2007. La scelta dell'anno di riferimento è stata dettata dall'opportunità di scegliere un periodo in cui la gestione e la qualità delle fonti coinvolte abbia raggiunto una certa stabilità; inoltre, si prevede la possibilità, in una fase successiva, di incrementare il sistema con gli eventi relativi agli anni seguenti, così da gettare le basi per la creazione di un sistema che permetta anche di seguire nel tempo i comportamenti riproduttivi e i relativi esiti in un'ottica longitudinale.

Si evidenzia che la scelta dell'anno è stata anche dettata dal fatto che le fonti hanno una tempistica di rilascio dei dati diversa per cui il 2007 era l'anno più recente comune a tutte le fonti disponibili nel momento in cui è iniziata la progettazione delle attività descritte in questo documento (primo semestre 2012). Nella successiva tavola 4.1 si riportano le numerosità relative alla regione Emilia Romagna per gli eventi occorsi nell'anno 2007 e la percentuale delle numerosità di tale regione rispetto al totale Italia.

Tavola 4.1. Numerosità degli eventi in Emilia Romagna nel 2007

	Nome della fonte	Numero eventi in Emilia Romagna	Percentuale sugli eventi in Italia
1a	CEDAP parti	39.792	7.65
1b	CEDAP nati vivi	40.022	7.63
1c	CEDAP nati morti	114	5.19
3	P4	39.744	7.15
4	IVG	11.267	9.01
5	AS	5.872	7.61
6a	SDO parti	40.242	7.22
6b	SDO nati vivi	41.637	7.44
6c	SDO nati morti	154	9.16
6d	SDO ivg	11.661	8.85
6e	SDO as	6.448	7.25

E' importante notare che, nell'anno 2007 e in una regione come l'Emilia Romagna che rappresenta un'eccellenza per la gestione e la qualità delle fonti oggetto di analisi, alcune di

queste riportano dati in maniera ancora poco precisa ed affidabile sia in termini di numerosità che di informazione registrata. E' il caso ad esempio della fonte CEDAP, soprattutto per la rilevazione della nati-mortalità. Queste criticità emergono chiaramente anche dagli indicatori di qualità delle procedure di linkage, come si può notare nel successivo paragrafo 4.1. Per questo motivo è importante che le analisi successive sui dati abbinati tengano conto e includano le misure della qualità dell'integrazione effettuata. Infatti, a queste criticità si può ovviare in diverse fasi del processo di produzione dell'informazione statistica. In fase di stima, si devono adottare stimatori che tengono conto esplicitamente della scarsa qualità delle fonti coinvolte. Questo implica che la qualità delle fonti sia valutata anche in termini quantitativi. In questo senso, il processo di integrazione di tipo probabilistico è il più indicato perché fornisce delle misure quantitative (la probabilità di corretto abbinamento e quella di mancato abbinamento) che possono essere direttamente inserite nella fase di stima degli indicatori forniti dal sistema integrato. A livello organizzativo, d'altro canto, è necessario che queste evidenze sulle cadute di qualità per particolari fonti o sottopopolazioni, siano sfruttate per identificare i campi di azione e gli attori su cui intervenire prioritariamente per aumentare la qualità dell'input.

4.1 I risultati del linkage probabilistico

Le analisi per la valutazione della fattibilità comprendono anche numerose applicazioni di integrazione delle fonti. I dettagli delle procedure di integrazione sperimentate sono descritti in un rapporto tecnico a diffusione interna dal titolo "Progettazione e realizzazione del prototipo del sistema integrato degli esiti del concepimento" (Istat, 2012). Di fatto si è proceduto abbinando singole coppie di fonti che riguardano lo stesso fenomeno, secondo lo schema 3.1. Le variabili disponibili per l'abbinamento possono essere raggruppate in 3 gruppi:

- le variabili riguardanti la donna protagonista dell'evento: l'età (in poche fonti la data di nascita completa), la provincia e/o il comune/stato estero di residenza, la provincia e/o il comune/stato estero di nascita;
- le variabili riguardanti la localizzazione geografica dell'evento: il codice dell'ospedale, il comune e la provincia;
- le variabili riguardanti la data dell'evento: giorno e mese.

Le strategie di abbinamento più efficaci sono risultate quelle in cui nella selezione delle variabili di linkage e delle variabili utilizzate per la riduzione dello spazio di ricerca delle coppie candidate è considerata almeno una variabile da ciascun gruppo. In realtà, i diversi processi di integrazione applicati alle varie coppie di dati hanno richiesto metodi di riduzione delle spazio di ricerca differenti (di bloccaggio o di ordinamento) basati su variabili diverse, in funzione soprattutto della dimensione dei dati da trattare. Conseguentemente, per le varie procedure di abbinamento sono state selezionate variabili di linkage diverse, evitando di scegliere all'interno dello stesso gruppo quelle con potere identificativo aggiuntivo troppo scarso, perché fortemente correlate (come ad esempio il codice dell'ospedale e il comune dell'evento). Per alcune variabili, quali l'età della donna e il giorno dell'evento, sono state prese in considerazione funzioni di confronto che, accettando concordanze meno stringenti dell'esatta uguaglianza, consentono di definire degli intervalli di somiglianza così da aumentare la probabilità di aggancio tra record.

La tavola 4.2 riporta i risultati delle sperimentazioni che hanno fornito gli esiti migliori, in base ai criteri definiti nel paragrafo 3.2, in termini di match rate e probabilità di corretto

abbinamento.

Tavola 4.2. Quadro riassuntivo degli abbinamenti ottenuti con record linkage probabilistico

Abbinamenti	Dimensione file di partenza	Numero abbinamenti da RL probabilistico	Match rate
AS – SDO as	5.872 – 6.448	3.864	0.66
IVG - SDO ivg	11.267 – 11.661	10.938	0.97
P4 - SDO parti	39.744 – 40.242	27.711	0.70
CEDAP natimorti – SDO natimorti	114 - 154	80	0.70
CEDAP parti - SDO parti	39.792 – 40.242	33.622	0.84
P4 - CEDAP nati	39.744 – 40.370	37.469	0.94

I risultati degli abbinamenti tra le fonti considerate sono sicuramente incoraggianti. Infatti, per tutte le procedure di abbinamento considerate, le probabilità medie stimate che ciascuna coppia individuata sia un vero match è generalmente superiore al 95% (Tavola 4.3).

Tavola 4.3. La qualità degli abbinamenti in termini di falso e mancato abbinamento

Abbinamenti	Numero record abbinati	Probabilità media di corretto abbinamento	Numero stimato di falsi abbinamenti	Numero massimo stimato di mancati abbinamenti
AS – SDO as	3.864	0.97	116	1314
IVG - SDO ivg	10.938	0.93	766	328
P4 - SDO parti	27.711	0.99	277	8.313
CEDAP natimorti – SDO natimorti	80	0.99	1	15
CEDAP parti - SDO parti	33.622	0.99	336	5.380
P4 - CEDAP nati	37.469	0.99	375	2.248

La eventuale presenza di falsi abbinamenti nel sistema integrato sugli esiti dei concepimenti costituisce una criticità per lo studio e l'analisi dei fenomeni che da questo si vogliono interpretare soprattutto se i falsi abbinamenti introducono distorsione nei dati. Di fatto, non ci sono ragionevoli motivi per ritenere che il processo di abbinamento sia distorto rispetto a qualcuna delle variabili di riferimento per le stime degli indicatori di interesse. In ogni caso il tasso di falso abbinamento e la probabilità di corretto abbinamento sono misure di cui bisognerà tenere conto esplicitamente nelle analisi successive sui dati abbinati, per una corretta modellizzazione del processo di generazione dei dati stessi.

Il numero di mancati abbinamenti è anch'esso contenuto, dati gli alti valori del match rate.

Riguardo alla stima del numero di mancati abbinamenti, si deve considerare che non tutti i record dei file di partenza che non si abbinano sono mancati abbinamenti, dato che nella maggior parte dei casi non si verifica una perfetta sovrapposizione delle popolazioni di riferimento. Ad esempio, i record riportati nelle due fonti P4 - SDO parti non sono riferiti esattamente allo stesso universo di riferimento, per varie ragioni, tra cui:

- la fonte P4 riporta tanti record quanti sono i nati vivi, indipendentemente dal fatto che provengano da parto singolo o multiplo, quindi, nel caso di parti singoli, c'è coincidenza con le informazioni sul parto riportate nei record della fonte SDO parti mentre i record del P4 generati da parti multipli "eccedono" rispetto ai record SDO parti per tutti i gemelli dal secondo in poi;
- nella fonte SDO ci sono record di parti di bambini nati morti che la fonte P4 non

registra;
- nella fonte SDO ci sono record relativi a parti avvenuti in Emilia Romagna di bimbi iscritti in anagrafe fuori dalla regione o anche non iscritti, come nel caso delle donne straniere;

Di conseguenza, è attesa una buona sovrapposizione dei file, d'altra parte è ammesso che un numero di mancati abbinamenti non costituisca errore del processo ma sia dovuto al diverso universo di riferimento delle fonti trattate. Per quanto riguarda il fenomeno delle nascite, la fonte SDO nati potrebbe avere una maggiore aderenza alla fonte P4 rispetto a SDO parti; purtroppo, dalla ricognizione delle fonti effettuata, è emerso che allo stato attuale le variabili identificative utili ai fini dell'abbinamento riportate in SDO nati sono troppo esigue e nella maggior parte dei casi non valorizzate.

Infine, occorre ricordare che sia nel record linkage tra le fonti considerate, che in quello effettuato sui residui si è scelto di operare un abbinamento di tipo 1:1: in caso di parto plurimo che ha dato luogo a due o più nati vivi ci si aspetta che una stessa scheda di dimissione ospedaliera per parto si agganci a più schede di Iscrizione in Anagrafe per nascita. L'abbinamento di tipo 1:1 consente, per questa specifica coppia di fonti, di creare un sistema integrato che ha come unità di riferimento la singola donna e non il risultato dei suoi concepimenti. Il fenomeno "esito del concepimento" per ciascuna donna può essere ricostruito evitando la riduzione degli abbinamenti da 1:n a 1:1, integrando con la fonte SDO nati morti e prendendo in considerazione l'integrazione con le fonti SDO nati (quando sarà maggiormente popolata di informazioni essenziali per il linkage) e CEDAP.

Infine, è ragionevole ritenere che ulteriori abbinamenti saranno recuperati nel momento in cui verranno prese in esame tutte le regioni italiane, dato che mancate coincidenze nella variabile che indica la regione di evento impediscono di confrontare record e riconoscere ulteriori coppie. Nel momento in cui si provvederà alla messa a punto del sistema per l'intera Italia, si può immaginare di confrontare a livello nazionale i record che non si abbinano all'interno della singola regione, così da recuperare queste ulteriori coppie.

4.2 L'applicazione di linkage deterministico

Sui dati in esame è stata applicata anche una procedura di abbinamento deterministico, che è stata sviluppata nel contesto delle attività di abbinamento deterministico di dati demo-sociali e nel tempo è stata arricchita e perfezionata, dimostrando di perseguire ottimi risultati. Il metodo viene brevemente descritto in questo paragrafo e può definirsi un modello a "chiavi integrate" corredato di un indicatore di "qualità degli abbinamenti" ottenuti.

Il metodo consente l'utilizzo contemporaneo, in maniera integrata all'interno della stessa procedura, di una serie di chiavi di linkage, che si ottengono dal "concatenamento testuale" di un gruppo di variabili di linkage, opportunamente scelte, nell'ottica del problema trattato.

È opportuno precisare che una delle peculiarità del metodo in questione è quella di prevedere due diversi tipi di chiavi di linkage:
- la chiave di linkage completa, che è la chiave più restrittiva che si possa costruire, ossia quella che si ottiene concatenando tutte le variabili di linkage individuate;
- le "altre" chiavi di linkage o "derivate", che sono chiavi alternative alla chiave completa e in un certo modo da essa "derivate" e per questo "meno restrittive" della stessa. Esse si ottengono dalla chiave completa togliendo una variabile di linkage alla volta.

Per il calcolo dell'indicatore di qualità degli abbinamenti, occorre definire le variabili "di controllo", anch'esse comuni alle due fonti, meno discriminanti delle variabili di linkage, ma lo stesso utili per la valutazione della qualità dei risultati ottenuti. Infatti, il suddetto indicatore di qualità si ottiene conteggiando il numero di concordanze totali che si verificano tra i valori assunti dalle variabili comuni alle due fonti (sia di linkage che di controllo) sull'abbinamento individuato. Le concordanze vengono conteggiate solo per i valori diversi dai valori mancanti su entrambe le fonti. Questo indicatore costituisce uno strumento prezioso in due momenti di applicazione del metodo:

- in una fase intermedia del processo, quando una unità di una delle due fonti si lega a più unità dell'altra; in questo caso il metodo sceglierà l'abbinamento in cui l'indicatore di qualità presenti il valore massimo (massima concordanza delle informazioni disponibili);
- nella fase finale di valutazione, quando occorre decidere se gli abbinamenti dichiarati dal metodo debbano essere giudicati veri, falsi o dubbi.

Per le sperimentazioni relative al sistema integrato degli esiti dei concepimenti, le diverse chiavi su cui si basa il metodo sono state costruite a partire dalle stesse variabili utilizzate per l'applicazione del linkage probabilistico.

4.3 Analisi e confronto dei risultati

In questo paragrafo si analizzano in maniera comparativa i risultati delle due tecniche di linkage sperimentate: quella probabilistica e quella deterministica. Dal confronto emergono punti di forza e di debolezza di entrambi gli approcci, ma soprattutto tale analisi mette in luce il lavoro ancora da fare per il miglioramento della qualità delle fonti di partenza, soprattutto in un'ottica di utilizzo integrato dei dati. In ogni caso, il confronto dei risultati delle due procedure di abbinamento sperimentate conferma la validità della strategia delineata per la costruzione del sistema integrato sugli esiti dei concepimenti, poiché in generale, salvo poche eccezioni discusse nel seguito, la percentuale degli abbinamenti individuati da entrambe le procedure è superiore al 70%, con punte di 97%, a conferma della robustezza dei processi di linkage individuati.

Nella successiva tavola 4.4 si riassumono i risultati, per ogni coppia di fonti considerate: riportando ancora la dimensione dei file di partenza per ogni abbinamento, vengono presentati il totale di abbinamenti individuati dall'approccio deterministico, il numero di abbinamenti trovati dai modelli probabilistici e il numero di coppie comuni alle due procedure; l'ultima colonna della tavola riporta in termini percentuali il rapporto tra il numero di abbinamenti comuni ai due metodi e la dimensione del più piccolo insieme di abbinamenti individuati.

Tavola 4.4. Quadro riassuntivo degli abbinamenti secondo i metodi di record linkage sperimentati

Fonti abbinate	Dimensione file di partenza	RL deterministico	RL probabilistico	Link comuni	% Link comuni
AS - SDO as	5.872–6.448	4.746	3.864	720	18,6
IVG - SDO ivg	11.267–11.661	9.650	10.938	7.497	77,7
P4 - SDO parti	39.744–40.242	33.752	27.711	19.266	69,5
CEDAP natimorti - SDO natimorti	114–154	102	80	79	98.7
CEDAP parti - SDO parti	40.242–39.792	36.481	33.622	27.844	82.8
P4 - CEDAP nati	39.744–40.370	38.081	37.469	36.572	97.6

Da un primo esame della tavola risulterebbe che in generale il metodo deterministico tenda ad individuare un numero maggiore di abbinamenti rispetto ai modelli probabilistici sperimentati. In realtà, questo risultato necessita un'analisi più approfondita poiché dipende da diversi fattori. In primo luogo incide il metodo deterministico applicato, infatti questo non si limita a dichiarare abbinamenti tutte quelle coppie che coincidono sull'insieme completo di variabili di confronto selezionate ma procede ad abbinare anche le coppie che coincidono su un insieme più limitato di variabili, rimuovendo il vincolo di uguaglianza per una o più delle variabili di confronto selezionate, purché il numero di concordanze totali sia superiore ad una soglia prefissata. In secondo luogo, il minor numero di abbinamenti individuati dai modelli probabilistici è condizionato dal fatto che questi sono stati applicati in via sperimentale senza iterazioni successive sui record rimasti non abbinati al primo passo, al contrario della pratica consueta nell'uso di metodi probabilistici. Ciò è confermato in particolare dal risultato relativo all'abbinamento tra le fonti IVG e SDO-ivg, per le quali le procedure probabilistiche sono state impiegate come di prassi in più passi successivi, con relativo abbinamento dei record rimasti non abbinati nei passi precedenti. In questo caso in particolare sono state sperimentate due iterazioni e ciò ha consentito al metodo probabilistico di individuare un numero di abbinamenti maggiore rispetto al metodo deterministico. Tale constatazione suggerisce di estendere la pratica di applicazioni ripetute e iterate dei modelli di linkage probabilistico anche alle altre coppie di dati, poiché è ancora possibile processare i dati non abbinati per recuperare corretti abbinamenti. In ogni caso, restano valide le considerazioni conclusive del paragrafo 4.1, sulla possibilità di individuare nuovi abbinamenti estendendo il campo di osservazione dell'attuale sperimentazione di linkage probabilistico all'intero territorio nazionale.

Infine, al di là della valutazione comparativa sulle performance dei due diversi metodi sperimentati, il confronto mette in luce un aspetto fondamentale nell'ottica di messa a regime del sistema integrato sugli esiti dei concepimenti. A tal proposito, la tavola 4.4 evidenzia chiaramente come la comparabilità dei risultati sia fortemente legata alle fonti considerate nel linkage: prendendo in considerazione le fonti per cui la sovrapposizione degli eventi e degli universi di riferimento è più alta (P4 - CEDAP nati, CEDAP parti - SDO parti, CEDAP natimorti - SDO natimorti) i risultati delle due procedure sono simili, con percentuali di sovrapposizione fino al 98%. La comparabilità dei risultati scende fino al 70% circa per le fonti relative ad universi di riferimento solo parzialmente omogenei (si vedano le considerazioni conclusive del paragrafo 4.1 relative al linkage tra P4 - SDO parti). Un discorso a parte merita il confronto dei risultati degli abbinamenti AS - SDO e IVG - SDO: data la natura delle fonti e la qualità delle informazioni in esse riportate, ci si

aspettava una percentuale di sovrapposizione dei risultati analoga tra i due abbinamenti. La notevole differenza nelle percentuali di abbinamenti comuni individuati dalla procedura deterministica e da quella probabilistica e il numero contenuto di abbinamenti per le fonti AS - SDO sono da imputare alla non disponibilità per l'anno di riferimento della variabile "giorno di intervento". Ulteriori approfondimenti sono disponibili su altri lavori effettuati in Istat (Cotroneo, Tuoto, Loghi, 2012). In ogni caso la variabile "giorno di intervento" si è rivelata di fondamentale importanza per la costruzione dei modelli di linkage tanto che è stata tempestivamente attivata la procedura per il suo inserimento nella rilevazione Istat degli aborti spontanei.

5. Conclusioni e prospettive future

I risultati del complesso processo di integrazione per la valutazione della fattibilità della costruzione di un sistema integrato sugli esiti dei concepimenti sono molto incoraggianti. La sperimentazione sulla regione Emilia Romagna e per l'anno di riferimento considerato suggeriscono di continuare sul percorso individuato, dato l'elevato numero di record abbinati e l'alta qualità degli abbinamenti. Il sistema permette di produrre indicatori su temi quali la medicalizzazione del percorso gravidanza, nascita e allattamento in relazione al contesto socio-sanitario, alle caratteristiche socio-demografiche delle donne, al contesto socio-economico familiare; il percorso gravidanza, nascita e parto (fisiologico/naturale vs fortemente medicalizzato).

In questo paragrafo sono riportate alcune conclusioni sul lavoro svolto e si evidenziano delle proposte per possibili sviluppi, relativamente agli aspetti definitori, ai dati considerati e alle procedure di integrazione.

Per quanto riguarda gli aspetti definitori, la strategia sperimentata per la messa a punto del sistema integrato sugli esiti dei concepimenti fa riferimento al fenomeno "esito del concepimento", rispetto alle modalità: nato vivo, nato morto, aborto spontaneo, interruzione volontaria, come precisato nel paragrafo 3. Le potenzialità offerte dal sistema fin qui illustrate fanno riferimento quindi all'adozione di un'ottica trasversale. Come possibile sviluppo si vuole indagare nei prossimi passi la possibilità di ricondurre ciascun fenomeno a quelli ad esso correlati (ad esempio, parti multipli e anche parti, interruzioni di gravidanza e aborti relativi ad una stessa donna) ossia spostare l'attenzione dal fenomeno "esito" del singolo concepimento alla "donna", in quanto soggetto del concepimento. Il prototipo studiato e sperimentato in questo lavoro è comunque un passaggio obbligato dell'estensione proposta, come ampiamente illustrato nel paragrafo 2, per "seguire" nel tempo l'evoluzione delle storie riproduttive delle donne, anche in prospettiva longitudinale. Per quanto riguarda i dati da utilizzare per la costruzione del sistema integrato sugli esiti dei concepimenti, l'attuale sperimentazione ha condotto un'analisi puntuale della fattibilità dell'integrazione a livello micro per le varie fonti considerate, da cui è emerso un insieme minimo di variabili fondamentali a garantire abbinamenti di elevata qualità e la conseguente necessità di richiedere tali variabili per le fonti che ancora non ne sono provviste. Tale richiesta dovrebbe essere evasa facilmente, in quanto le variabili fondamentali sono già rilevate dall'ente che fornisce i dati e non sono soggette a vincoli legati alla privacy. E' il caso ad esempio della variabile "giorno di intervento" per la fonte AS, evidenziato nel paragrafo precedente.

Un ulteriore utilizzo dei dati disponibili, non ancora indagato in modo approfondito, prevede infine la possibilità di collegare tra di loro gli abbinamenti individuati tra le singole coppie di fonti attraverso identificativi esatti condivisi tra alcune fonti, laddove la corrispondenza sia del tipo uno a uno (si pensi ai nati vivi registrati nei CEDAP, negli iscritti in anagrafe per nascita e nelle SDO). Questo passaggio richiede la definizione di controlli di coerenza tra le diverse fonti per correggere situazioni di incompatibilità dovute ad errori o scarsa qualità.

Si rimanda ad approfondimenti successivi anche il trattamento dei record non abbinati delle varie fonti considerate dalle procedure di linkage. Queste valutazioni sono legate soprattutto alla definizione degli universi di riferimento delle fonti considerate e agli obiettivi conoscitivi del sistema integrato, oltre che alle procedure di integrazione messe in atto.

Per quanto riguarda gli aspetti strettamente connessi alle procedure di integrazione, le sperimentazioni effettuate hanno fornito risultati molto confortanti sia in termini di numero di record abbinati che di qualità degli abbinamenti, dichiarando quindi fattibile la realizzazione di un sistema integrato sugli esisti dei concepimenti. I lavori dovrebbero proseguire estendendo l'abbinamento all'intero territorio nazionale. In quest'ottica, sulla base delle sperimentazioni condotte, un ulteriore filone di attività consiste nella possibilità di introdurre nelle procedure di abbinamento dei vincoli che garantiscano un'elevata qualità dei risultati soprattutto rispetto a particolari fenomeni o sottopopolazioni su cui il sistema integrato voglia fornire dei focus (ad esempio, il fenomeno della natimortalità o la sotto-popolazione delle donne straniere).

Un successivo sviluppo dell'attuale sistema prevede, come anticipato nel paragrafo 2, l'integrazione di ulteriori fonti di dati, ad esempio quelle relative all'indagine campionaria sulle nascite e all'indagine Istat sulle cause di morte, con riferimento ai morti nel primo anno di vita e negli anni successivi. L'integrazione di queste ulteriori fonti di dati permetterebbe di arricchire gli obiettivi conoscitivi coperti dal sistema integrato e recuperare il grave vuoto informativo sulla salute perinatale. Al momento, infatti, non è possibile ad esempio ricondurre la mortalità nel primo anno di vita alle informazioni sulla gravidanza e il parto. Includere nel sistema anche l'indagine Istat sulle cause di morte permetterebbe di calcolare i principali indicatori sulla salute perinatale (mortalità neonatale precoce e tardiva), come raccomandato dal Progetto Europeristat, e consentire il confronto con gli altri Paesi europei oltre che fornire valutazioni anche a livello regionale e sub-regionale. Inoltre l'integrazione nel sistema della rilevazione delle cause di morte dopo il primo anno di vita consentirebbe di mettere in relazione la mortalità materna per cause connesse alla gravidanza o al parto con le informazioni sulla gravidanza e il parto contenute nei CEDAP. Anche questo tipo di indicatori è fortemente raccomandato a livello internazionale (European Perinatal Health Report, 2008). E' possibile includere nel sistema integrato anche le indagini campionarie sulle nascite e le madri, dove la popolazione di riferimento sono le nascite della popolazione residente riferite agli anni di calendario 2003 e 2009-2010. Tale operazione consentirebbe di:

a) verificare la qualità e la copertura delle indagini rispetto ad alcune variabili chiave come il livello di istruzione della popolazione;
b) valutare la bontà dell'ordine di nascita stimato utilizzando la variabile "numero di componenti minorenni" registrata nelle schede anagrafiche di iscrizione per nascita (previo opportuno trattamento statistico di validazione), con l'informazione diretta

fornita dalle intervistate e la parità desumibile dai CEDAP in modo da migliorare ulteriormente le procedure di stima;

c) studiare la qualità delle informazioni sanitarie sul parto rilevate attraverso le interviste Cati rispetto a quelle rilevate per le stesse donne dal personale sanitario che compila il CEDAP (per verificare, ad esempio, se le donne tendono a sovrastimare il peso dei figli alla nascita o a non riferire eventi sensibili come esiti negativi di precedenti concepimenti o nascite non viventi nel caso di parto gemellare con nati vivi) .

In generale sarebbe possibile valutare il grado di affidabilità delle informazioni raccolte con le indagini campionarie anche in vista di successive occasioni di indagini. Ciò consentirebbe di disporre di una base dati di amplissimo potere informativo per l'analisi delle nascite e dei parti in relazione alle determinanti socio-sanitarie e a quelle demo-sociali.

Infine, si vuole ricordare che le successive analisi statistiche basate sui dati riportati nel sistema integrato dovranno necessariamente tenere nella debita considerazione il processo di integrazione che ha generato tali dati. Infatti, sebbene i metodi di integrazione di microdati siano uno strumento estremamente potente e ampiamente utilizzato, il linkage, come in genere molte fasi del processo di produzione del dato statistico, produce risultati non sempre privi di errori. Come evidenziato nel paragrafo 3, le applicazioni di record linkage devono quindi essere corredate da opportune informazioni sulla qualità del linkage; tali indicatori di qualità, a partire da quelli riportati nel paragrafo 4, devono essere inseriti con opportune metodologie nelle successive fasi di stima basata su dati linkati, così da assicurare la qualità delle analisi finali. Infatti, trattare i dati abbinati ignorando il processo di integrazione e gli eventuali errori da questo generati, può portare, in generale, a stime distorte (Neter *et al,* 1965, Winkler e Scheuren, 1993, 1997). Bisognerà quindi studiare e implementare opportuni metodi di stima di indicatori e di relazioni statistiche tra variabili che tengano nella dovuta considerazione il processo di integrazione alla base della costruzione del sistema integrato.

Riferimenti bibliografici

Attili M., Valentino L. (2010). "Le informazioni sulle nascite e i parti. Esperienze di integrazione tra dati di fonte anagrafica e sanitaria", presentato al seminario "REcord Linkage At IStat: Applicazioni con RELAIS 2.0", 22 aprile 2010, Istat, Roma.

Cibella N., Fortini M., Spina R., Scannapieco M., Tosco L., Tuoto T. (2007). "Relais: An open source toolkit for record linkage", Rivista di Statistica Ufficiale n. 2-3/2007, pp.55-68

Cibella N., Tuoto T. (2012) "Statistical perspectives on blocking methods when linking large data-sets", in A. Di Ciaccio et al. (eds.), Advanced Statistical Methods for the Analysis of Large Data-Sets, Springer-Verlag Berlin Heidelberg.

Cotroneo R., Tuoto T., Loghi M. (2012). "L'importanza della scelta delle variabili nel record linkage: il caso delle Interruzioni volontarie di gravidanza e degli Aborti spontanei" presentato alle Giornate di Studio sulla Popolazione (GSP) 2013, Brixen, Febbraio 6 –8, 2013, disponibile all'indirizzo web http://www.sis-aisp.it/ocs-2.3.4/index.php/gsp2013/gsp2013/paper/view/233

European Perinatal Health Report. (2008). Disponibile all'indirizzo web www.europeristat.com

Fellegi, I.P., Sunter, A.B. (1969). "A Theory for Record Linkage", Journal of the American Statistical Association, 64, pp. 1183-1210.

Jaro, M. A. (1989). "Advances in record linkage methodology as applied to the 1985 census of Tampa Florida", Journal of the American Statistical Society, 84 (406), pp.414–20.

Istat (2012) "Progettazione e realizzazione del prototipo del sistema integrato degli esiti del concepimento" relazione finale del Gruppo di lavoro "Metodi e tecniche di record linkage tra fonti demografiche e sociali"

Neter, J., Maynes, E.S., Ramanathan, R. (1965). The effect of mismatching on the measurement of response error. Journal of the American Statistical Association, 60, 1005-1027.

RELAIS (2011). User's guide version 2.2. Disponibile all'indirizzo web http://www.istat.it/it/strumenti/metodi-e-software/software/relais e http://joinup.ec.europa.eu/software/relais/release/22

Scheuren, F.,Winkler, W.E. (1993). Regression analysis of data files that are computer matched – Part I. Survey Methodology, 19, pp. 39-58.

Scheuren F.,Winkler W.E. (1997). Regression analysis of data files that are computer matched- part II, Survey Methodology, 23, pp. 157-165.