Feidhmeannacht na Seirbhíse Sláinte
Health Service Executive

Development of a Diabetes Register in Ireland: Feasibility Study – Recommendations for the Establishment of a Register

Authors: Mary Cooke, Gemma Leane, Mai Mannix, Orlaith O'Reilly

National Diabetes Programme
May 2010

Printed by Modern Printers: 056 7721739

Feidhmeannacht na Seirbhíse Sláinte
Health Service Executive

ISBN: 978-1-874218-84-5

Development of a Diabetes Register in Ireland: Feasibility Study – Recommendations for the Establishment of a Register

Authors: Mary Cooke, Gemma Leane, Mai Mannix, Orlaith O'Reilly

Feidhmeannacht na Seirbhíse Sláinte
Health Service Executive

National Diabetes Programme

May 2010

Foreword

The HSE National Diabetes Programme was established in 2010 to decrease morbidity and mortality from diabetes in Ireland. One of the key priorities of the programme is to set up a register of all people with diabetes in order to facilitate organised care and screening for complications.

This report provides valuable practical advice on how a register can best be established and maintained at the present time. Healthcare systems, and particularly clinical systems, in Ireland are not well integrated or developed. This lack of integration, standardisation and in particular the lack of a unique identifier has militated against the provision of disease registers in the past.

However, the extensive work carried out in this study on the practicalities of establishing a diabetes register from existing data sources and estimating its completeness provides clear recommendations on how the Diabetes Programme should proceed.

This work will also inform the National Diabetic Retinopathy Screening Information System Project.

The National Diabetes Programme endorses the findings of this report and would like to thank the Project Team and all who contributed to this study.

Professor Richard Firth,
National Clinical Lead Diabetes Programme

Preface

Diabetes is responsible for a significant burden of ill health for the population. Approximately 150,000 people in Ireland currently suffer from diabetes and this is estimated to increase by 62% in the next 10 years. Many people who have diabetes also suffer from other chronic diseases e.g. cardiac disease, stroke and high blood pressure. This represents a huge challenge to the Health Services. The key to the management of diabetes is; initial prevention, early diagnosis, structured management and prevention of complications.

In order to provide structured management, and in particular to prevent complications, a register of patients diagnosed with diabetes is essential.

The need for disease registers have long been recognised by the Irish Health Services, however this is the first study which actually teases out the practical implications and feasibility of developing such a register for people with diabetes, given our current information systems. This work has delivered two hugely important results; a practical method of register development and an actual register of patients for a defined geographical area.

These results provide the essential information for the national development of a Diabetic Retinopathy Screening Programme, and supporting information systems. It is now the challenge for the National Diabetes Programme to build on this work to enable the National Diabetes Retinopathy Screening Programme to be established throughout the country.

Dr. Orlaith O'Reilly
Chair of the Project Team

Contents

List of Tables

1. Introduction

Diabetes mellitus results from an inadequate effect of insulin either due to its absence (type 1) or resistance to its action (type 2). The International Diabetes Federation (IDF) estimate that approximately 285 million people worldwide, or 6.6% of the adult population aged 20-79 years, will have diabetes in 2010 [1]. With the epidemic of obesity that is currently being experienced in the developed world it is predicted that this figure will rise to 438 million, or 7.8%, by 2030 [1].

The actual prevalence of diabetes in Ireland is not known. A report published by the Institute of Public Health in Ireland provides the best available estimates of the population prevalence of diabetes (diagnosed and undiagnosed) in 2007 [2]. Approximately 144,000 persons in the Republic of Ireland are estimated to have adult diabetes (type 1 and 2 combined), i.e. 4.5% of all adults, aged 20 years and older. The estimated population prevalence for 2015 is 5.2% (193,000 adults with diabetes), increasing to 5.9% (233,000 adults with diabetes) in 2020, representing an increase of 62% over the 13 years [2]. This will be largely due to an increase in the incidence of type 2 diabetes owing to the increases in childhood and adolescent obesity.

In 2006 the Department of Health and Children made a number of policy guidance recommendations in relation to the model of care for people with diabetes and recommendations for how services will prevent and manage diabetes in the population [3]. One recommendation was the development of a diabetes register, starting at local/regional level. The Institute of Public Health in Ireland also strongly recommended the establishment of national diabetes registers on the island, North and South [4].

Diabetes registers are key components of the information support needed for prevention, care and monitoring of diabetes. A disease register is a special form of clinical database. In compiling a population-based register, an attempt is made to identify and collect data on all cases of a disease (in this case diabetes mellitus) or other health condition within a defined population. Ideally it includes electronic feeds from laboratories, pharmacies and clinical encounters [5].

The National Diabetes Expert Advisory Group (EAG) was established in 2006 to advise the HSE on the development of diabetic services. The EAG identified the development of a register of diabetic patients as a priority which was necessary to underpin the National Diabetic Retinopathy Screening Programme and integrated care for diabetes [6]. The EAG National Diabetic Retinopathy Screening Committee [7] set up a feasibility study and project team to recommend a methodology to establish a diabetes register in Ireland. It was decided to develop the methodology in the Mid West region (Clare, Limerick and North Tipperary) and work commenced in April 2008.

The purpose of the register was to identify all people with diabetes mellitus, in the Mid West region, initially to support a Diabetic Retinopathy Screening Programme and with capability to expand to support other aspects of diabetes care. This was to be achieved by identifying client records from multiple electronic data sources in the Mid West region. It was recognised that there would be overlap of client records in the data sources and the data would be structured in disparate ways. This feasibility study was to assess the difficulty and potential for accuracy of these data sources in compiling the register. The results will inform the recommendations on a method of formulating and updating the diabetes register, relevant both regionally and nationally.

The National Diabetes Programme, established in 2010, identified the creation of a register of diabetes patients as a key priority. The programme accepts the recommendations of this report.

Aim

The aim of the feasibility study is to recommend a methodology to establish a diabetes register in the chosen test site, the Mid West region. It is proposed that the recommended methodology could be replicated in order to formulate a national diabetes register in Ireland. The feasibility study will address the question on how best to establish an accurate and complete register in an efficient and effective manner based on the practicalities of current Irish health service systems and resources available.

Objectives

1. To identify appropriate sources of data for the diabetes register.
2. To recommend a method of compiling and updating the diabetes register.
3. To recommend a method of quality assuring the diabetes register.
4. To make the above recommendations based on practical considerations of current Irish health service systems.
5. To identify resources required to formulate a diabetes register.
6. To make the above recommendations based on result of testing the accuracy and completeness of proposed data sources in the Mid West region.
7. To make recommendations in compliance with data protection legislation.
8. To recommend a minimum dataset needed to establish a diabetes register.

Members of the Project Team

Ms. Rachael Banques	Community Diabetes Care Facilitator	Nursing & Midwifery Planning & Development Unit, Limerick, HSE West
Ms. Mary Cooke	Information Systems Manager	Primary Community & Continuing Care (PCCC) Information Systems Dept, Limerick, HSE West
Ms. Alison Cullinan	Systems Analyst	PCCC Information Systems Dept, Limerick, HSE West
Ms. Gemma Leane	Research Officer	Public Health Dept, Kilkenny, HSE South
Dr. Mai Mannix	Specialist in Public Health Medicine	Chair of HSE West Diabetic Retinopathy Screening Committee
Mr. Tom Moore	Managing Director	EBCS Ltd
Mr. Michael Murphy	Senior Systems Analyst	PCCC Information Systems Dept, Limerick, HSE West
Dr. Orlaith O'Reilly	Director of Public Health Chair of the Project Team	Chair of National Diabetic Retinopathy Screening Committee
Mr. Tom Slater	Operations Manager	EBCS Ltd
Ms. Mairead Gleeson (member of project team initially for 12 months)	Project Manager	Public Health Dept, Kilkenny, HSE South
Dr. Sarah Doyle (member of project team initially for 6 months)	Specialist in Public Health Medicine	Public Health Dept, Kilkenny, HSE South

Acknowledgements

General Practitioners, Practice Nurses and Practice Secretaries who facilitated the study

PCCC Information Systems Department Staff		Limerick, HSE West
Dr. Ned Barrett	Consultant Clinical Biochemist	Mid Western Regional Hospital Limerick, HSE West
Ms. Trina Dooley	Regional HIPE/Casemix Coordinator	HIPE Dept, Mid Western Regional Hospital Limerick, HSE West
Mr. Oliver Power	Laboratory Information System Manager	Mid Western Regional Hospital Limerick, HSE West
Prof. Kevin Balanda	Associate Director	Institute of Public Health in Ireland
Ms. Lorraine Fahy	Research Analyst	Institute of Public Health in Ireland
Dr. Margaret Morgan	Community Ophthalmic Physician	Diabetic Retinopathy Screening Service, Letterkenny, HSE West
Staff in the Diabetic Retinopathy Screening Service		Letterkenny, HSE West

2. Methodology

The feasibility study commenced in April 2008 with the final phase completed in September 2009. There were six phases to the study which are briefly summarised below.

Phase 1: Identify existing data sources/information systems in the Mid West region which could be used to identify people with diabetes. Data sources/information systems which would have individually identifiable information together with any of the following:

- A diagnosis of diabetes
- Attendance at hospital outpatient specialised diabetic clinics
- People prescribed drugs/blood glucose testing strips used in diabetes care

Phase 2: Capturing additional data fields for matching purposes, data cleansing and de-duplication within each of the identified data sources.
Two different methods were used in this phase:

- Manual de-duplication within four data sources
- Fuzzy de-duplication (using Fuzzy Grouping) within one data source

Phase 3: Data source hierarchy. Decision on which data sources to take priority when merging the data sources into one single file.

Phase 4: De-duplication across data sources. This involved the merging of each individual data source, based on data source hierarchy, using fuzzy de-duplication to produce a single record for each client within a single master file.

Phase 5: Refinement of the master file. Remove records for clients identified as deceased since the data was extracted, clients resident outside the Mid West region, clients who had a HbA_{1c} test within the upper part of the normal range.

Phase 6: Verification of a sample of the client records within the master file. Three different methods were used in this phase:

- Surname Verification
- Laboratory Information System Verification
- GP Visit Verification

The six phases are discussed in greater detail.

Phase 1: Data Sources Used

The first phase of the feasibility study was to identify existing data sources/information systems in the Mid West region which could be used to identify people with diabetes. Data sources/information systems which would have individually identifiable information together with any of the following:

- A diagnosis of diabetes
- Attendance at hospital outpatient specialised diabetic clinics
- People prescribed drugs/blood glucose testing strips used in diabetes care

The following data sources/information systems were selected as appropriate sources which would be used to detect people with diabetes:

- Medical Card Prescriptions (MCP)
- Long Term Illness (LTI) Scheme
- Patient Administration System (PAS)
- Hospital In-Patient Enquiry (HIPE) Scheme
- Laboratory Information System (LIS)

A description of the data and the data extraction criteria for each data source/information system are presented in Tables 2.1-2.5.

Table 2.1: Description of MCP data and data extraction criteria

Data Source	**Medical Card Prescriptions (MCP)**
Extraction Criteria	Clients resident in the Mid West region prescribed diabetic drugs and/or blood glucose testing strips with ATC level codes: A10A/A10B or V04CA91 A10A: Insulins and analogues A10B: Blood glucose lowering drugs, excluding insulins V04CA91: Blood glucose test strips
Data Period	December 2006 to November 2007 inclusive
Total Clients Extracted	7083
Data Fields Available	Medical Card Number including card position, ATC Level, Sex, Date of Birth, GMS GP Code, GP Name, GP Address, GP Area
Notes	• Data related to medical card holders only • Data in respect of diabetic clients was extracted by Galmac (company who provide the Drug Prescribing Analysis System in the HSE areas) from the full file of Medical Card Prescriptions for the Mid West received from the Primary Care Reimbursement Service (PCRS). One single record extracted in respect of each client, as each client could have multiple records in the prescriptions file • MCP did not include client demographic data • Diabetics would be expected to have at least one prescription supply in a one year period • November 2007 was the latest file available when exercise undertaken • PCRS have a national medical card prescriptions database

Table 2.2: Description of LTI scheme and data extraction criteria

Data Source	**Long Term Illness (LTI) Scheme**
Extraction Criteria	Clients resident in the Mid West region with a LTI card diagnosed with diabetes mellitus
Data Period	All current active cards on Careworks Schemes System as at 12/05/2008
Total Clients Extracted	4306
Data Fields Available	System Number, LTI Number, Surname, Forename, Date of Birth, PPSN, Address, CCA, Illness, GMS GP Code, GP Name, Medical Card Number including card position
Notes	• LTI scheme is a module in the Careworks Schemes System • Careworks Schemes System is not a national system. It is used in the Mid West area and a number of other HSE areas • LTI data is sent to PCRS from each HSE area, PCRS have a national LTI database

Table 2.3: Description of PAS and data extraction criteria

Data Source	**Patient Administration System (PAS)**
Extraction Criteria	Any clients attending outpatient specialised diabetic clinics at the following Mid West hospitals: Mid Western Regional Hospital Limerick Clinic 1. Diabetic Clinic Clinic 2. Young Persons Diabetic Clinic Clinic 3. Diabetic Eye Clinic Clinic 4. Paediatric Diabetic Clinic Clinic 5. Diabetic Dietetic Clinic Ennis General Hospital Clinic 6. Diabetic Clinic
Data Period	Commencement date of the clinic on PAS up to and including 14/05/2008 Clinic 1. March 1995 to 14/05/2008 Clinic 2. October 2003 to 14/05/2008 Clinic 3. June 1995 to 14/05/2008 Clinic 4. January 2002 to 14/05/2008 Clinic 5. February 2007 to 14/05/2008 Clinic 6. January 2005 to 14/05/2008
Total Clients Extracted	Clinic 1. 4287 Clinic 2. 419 Clinic 3. 2806 Clinic 4. 148 Clinic 5. 86 Clinic 6. 920
Data Fields Available	Clinic Name, GP Name, GP Address, Medical Record Number (Chart Number), Surname, Forename, Date of Birth, Address, Phone Number, Last Attendance Date, Deceased Date
Notes	• No specific separate diabetic clinic on PAS at Nenagh General Hospital, however diabetics are seen in general medical clinics • Assumption that all clients attending a specialist diabetic clinic had a diagnosis of diabetes • Same PAS system used locally but separate PAS sites set up for each hospital • Variety of PAS and HIS products in hospital sites nationally. In the Mid West the PAS supplier is Irish Medical Systems Maxims (IMS)

Table 2.4: Description of HIPE scheme and data extraction criteria

Data Source	**Hospital In-Patient Enquiry (HIPE) Scheme**
Extraction Criteria	Any clients discharged from the Mid Western Regional Hospital Limerick, Ennis General Hospital and Nenagh General Hospital, day cases or inpatients, with a principal or secondary diagnosis of diabetes mellitus (ICD-10-AM codes E10-E14)
Data Period	January 2005 to December 2007 inclusive
Total Clients Extracted	6171
Data Fields Available	Medical Record Number, Surname, Date of Birth, Sex, Date In (Admission Date), Date Out (Discharge Date), RIP Flag, Date of Death, Area of Residence, Medical Card Flag, Medical Card Number
Notes	• ICD coding moved to ICD-10-AM in January 2005. For extraction purposes it was agreed to use ICD-10-AM codes only • Each hospital site has a separate HIPE database, however the same system is used in acute hospitals nationally • HIPE data is sent to the HIPE Unit in the Economic and Social Research Institute (ESRI) on a monthly basis • HIPE has very limited client demographic data and has no GP demographic/identification data

Table 2.5: Description of LIS and data extraction criteria

Data Source	**Laboratory Information System (LIS)**
Extraction Criteria	Any clients with a HbA_{1C} test where the results were either High (greater than 6.0%) or in the upper part of the normal range (NR) (between 5.5% and 6.0%)
Data Period	January 2007 to December 2007 inclusive
Total Clients Extracted	14624
Data Fields Available	Forename, Surname, Sex, Date of Birth, Patient ID, Address, Result, Clinician, GP Code (lab), GP County
Notes	• Assumption that most diabetics would have at least one HbA_{1C} test per year • Some diabetics may have a result in the normal range • Number of different LIS applications used nationally, Mid West use the iSoft iLab application • The Mid West has a single laboratory serving 5 hospital sites in the area: Mid Western Regional Hospital Limerick, Ennis General Hospital, Nenagh General Hospital, Mid Western Regional Maternity and Mid Western Regional Orthopaedic Hospital. There are individual PAS sites set up for each hospital and each of these PAS sites has an unilateral interface to the LIS for new clients added to PAS and any updates made to existing clients. There is no automated process on LIS to merge duplicate clients, merging tends to be on an ad hoc basis and therefore there may be multiple records for the 'same' client on LIS • Lab in most cases only receives a specimen and request form and staff cannot validate the client demographic information • At the time of data extraction some GPs from North Clare tended to use Galway rather than the Limerick lab for testing as the HSE specimen collection from GP practices was not in place

Data Sources Not Available

The Primary Care Reimbursement Service (PCRS) hold the Drugs Payment Scheme (DPS) data and unfortunately at the time of the feasibility study this could not be accessed from PCRS. Under the Drugs Payment Scheme, an individual or family have to pay a fixed amount each month (in 2008 this was €90 a month) on approved prescribed drugs, medicines and certain appliances for use by that person or his or her family before availing of the scheme. Any approved purchases over that fixed amount are covered. PCRS would only hold prescription data in respect of clients who exceeded the fixed amount a month (in 2008 this was €90). This scheme is aimed at those who do not have a medical card and normally have to pay the full cost of their medication. The prescribed drugs, medicines and appliances for use by people with diabetes could have been identified within the scheme. While all diabetic clients are entitled to a LTI card not everyone may be aware of this and could possibly be holding a DPS card.

St. John's Voluntary Hospital, Limerick was not included in the study.

Phase 2: Capturing Additional Data Fields, Data Cleansing and De-Duplication Within Each Data Source

The second phase involved capturing additional data fields for matching purposes, data cleansing and de-duplication within each of the five identified data sources. Two different methods were used in de-duplication:

- Manual de-duplication within four data sources
- Fuzzy de-duplication (using Fuzzy Grouping) within one data source

Capturing Additional Data Fields for Matching Purposes

Two data sources did not have sufficient data fields available to facilitate the de-duplication process. These data sources, MCP and HIPE, were matched to other existing local databases (for the feasibility study) to capture client demographic data and GP demographic/identification data as required.

Additional data fields captured and details on the matching criteria used are presented in Tables 2.6-2.7.

Table 2.6: Details of MCP match to Careworks Schemes client index daily extract

Data Source	**Medical Card Prescriptions (MCP)**
Data Fields Available (source file)	Medical Card Number including card position, ATC Level, Sex, Date of Birth, GMS GP Code, GP Name, GP Address, GP Area
Task Undertaken	The Mid West extracts a file of active clients from the local Careworks Schemes Client Index daily. The MCP file was matched to this daily file to extract the relevant demographic information on the client i.e. PPSN, Client ID (Schemes), Name and Address. The following matching criteria was used: • Medical Card Number, Date of Birth & Sex • Medical Card Number & Date of Birth • Medical Card Number & Sex • Medical Card Number Only
Data Fields Available after Matching	Medical Card Number, Careworks ClientID, Forename, Surname, Address, Date of Birth, Sex, PPSN, ATC Level, GMS GP Code, GP Name, GP Address, GP Area
Notes	• 1009 clients were not found on the daily active client extract file from Careworks Schemes Client Index

Table 2.7: Details of HIPE match to PAS for each Mid West hospital

Source	**Hospital In-Patient Enquiry (HIPE) Scheme**
Data Fields Available (source file)	Medical Record Number, Surname, Date of Birth, Sex, Date In (Admission Date), Date Out (Discharge Date), RIP Flag, Date of Death, Area of Residence, Medical Card Flag, Medical Card Number
Task Undertaken	HIPE file for each site was matched against the PAS site for the specific hospital (i.e. Mid Western Regional Hospital Limerick, Ennis General Hospital and Nenagh General Hospital) to extract demographic information for each client, i.e. Forename and Address, and GP demographic/identification data where available. The following matching criteria was used: • Medical Record Number PAS maintains a Referring GP field on the Admission Screen and also the clients usual GP on the Client Details Screen. The GP information was extracted from the "Referring Doctor" field on the corresponding admission record on PAS if recorded. Where it was not recorded on the admission the PAS Client Details Screen was checked to capture GP if available
Data Fields Available after Matching	Medical Record Number, Surname, Forename, Date of Birth, Sex, Address, Phone Number, GP Name, GP Address, Date In (Admission Date), Date Out (Discharge Date), RIP Flag, Date of Death, Area of Residence, Medical Card Flag, Medical Card Number
Notes	• The task of matching the Mid Western Regional Hospital HIPE file against the Mid Western Regional Hospital PAS file highlighted variation in the way some Medical Records were recorded on PAS and HIPE. This occurred where the Medical Record Number was of the format C + 5 digits on PAS. In HIPE these numbers were stored as C + 6 digits (leading zeros were inserted to fill out the number). For example, C01111 on PAS was C001111 on HIPE; C95000 on PAS was C095000 on HIPE. Numbers over C100000 matched on both. It was necessary to make a small change to the matching task and rerun it for these specific Medical Record Numbers so that the appropriate client demographic data and GP demographic/identification data could be obtained from PAS

Data Cleansing and De-Duplication Within Each Data Source

The data fields in each data source were reviewed and minimum data requirements set for each data source. Records which did not meet the minimum data requirements were deemed incomplete records and were removed.

The format of the GP demographic/identification data varied across the data sources. It was recognised that a common GP identifier (code) would be required to assist with verifying GP diabetic client lists as outlined in Phase 6.

A task was undertaken to generate a master list of GPs for the Mid West area. Locally available GP lists were reviewed to determine the highest quality, most comprehensive and complete list available. The GP list from MCP was selected as the master GP list as this file was relatively current and GMS GP code was included for most GPs.

The GP records on each of the other four data sources were then manually matched to the master GP list and the relevant GMS GP code recorded. Any new GPs identified during the process were added to the master GP list with their GMS GP code. Where a GP was not found, and efforts to identify the GP as a valid GP failed, they were marked as 'Unknown GP' and GPs from outside the Mid West area were marked 'Outside Area GP'. The client records on each data source were then updated with the relevant GMS GP code.

This task was extremely time consuming and labour intensive for PAS, HIPE (PAS GP recorded where available) and LIS. The quality of the GP demographic/identification data available on these data sources was poor and the lack of a nationally recognised GP identifier made processing of these GPs very slow. MCP and LTI, with the availability of a GMS GP code for clients (where GP was recorded), required very little effort to match.

Each data source was then examined for duplicate client records and these were removed. Duplicates at this stage would only be those with an exact match based on a variety of matching criteria.

The data fields available varied for each data source. Combinations of these data fields were used to identify possible duplicates. In the case of the 'Date of Birth' field, errors can occur in recording the correct date and to increase the number of possible duplicates 'Date of Birth' was split into day, month and year to allow matches on partial date of birth (i.e. combinations of day & month, day & year, and month & year were used). The de-duplication combinations chosen were sequenced in order to capture the duplicates with the highest probability of being definite duplicates first.

The process involved in manual de-duplication was as follows:

- Specific de-duplication task was run (as per Tables 2.8-2.11)
- An output file of the potential duplicates was produced
- Each file was manually reviewed by HSE staff and the actual duplicates on each file confirmed
- Each data source was then updated and the confirmed duplicates removed.

This was a very time consuming and labour intensive process, and would not be practical on a national scale.

The outcome of data cleansing and manual de-duplication within the data sources are presented in Tables 2.8 – 2.11.

Table 2.8: Description of data cleansing and manual de-duplication within MCP file

Medical Card Prescriptions (MCP)	Numbers	Notes
Total Clients Extracted	7083	
Incomplete Records Removed	1009	• Records not found on the daily active client extract file from Careworks Schemes Client Index. These clients were inactive as they were deceased, moved out of the region or had their medical card withdrawn at the time of the matching exercise
Duplicates	3	• De-duplication combinations used for matching included: Medical Card Number & Date of Birth Forename, Surname & Date of Birth Surname & Date of Birth Surname, Forename, Year of Date of Birth & Day of Date of Birth Surname, Forename, Year of Date of Birth & Month of Date of Birth Surname, Forename, Month of Date of Birth & Day of Date of Birth • The exercise identified 49 potential duplicates but following manual review this was reduced to 3 confirmed duplicates
Records after Manual De-Duplication	6071	• 498 records had no GP recorded

Table 2.9: Description of data cleansing and manual de-duplication within LTI file

Long Term Illness (LTI)	Numbers	Notes
Total Clients Extracted	4306	
Incomplete Records Removed	219	• Records without a valid Date of Birth were removed
Duplicates	11	• De-duplication combinations used for matching included: PPSN Medical Card Number & Date of Birth Forename, Surname & Date of Birth Surname & Date of Birth Surname, Forename, Year of Date of Birth & Day of Date of Birth Surname, Forename, Year of Date of Birth & Month of Date of Birth Surname, Forename, Month of Date of Birth & Day of Date of Birth
Records after Manual De-Duplication	4076	• 665 records had no GP recorded

PAS data was captured in two separate files, one for the five specialist diabetic Mid Western Regional Hospital Limerick clinics and the other for the Ennis General Hospital diabetic clinic. Duplicate records within each individual PAS file were first removed. The files were then merged into one file and the de-duplication process run again.

Table 2.10: Description of data cleansing and manual de-duplication within PAS file

Patient Administration System (PAS)	Numbers	Notes
Total Clients Extracted	7348	
Incomplete Records Removed	0	
Duplicates	132	• De-duplication combinations used for matching included: Medical Record Number Medical Record Number, Forename, Surname & Date of Birth Medical Record Number, Surname & Date of Birth Forename, Surname, Date of Birth & First Address Line Forename, Surname & Date of Birth Surname, Forename, Year of Date of Birth & Day of Date of Birth Surname, Forename, Year of Date of Birth & Month of Date of Birth Surname, Forename, Month of Date of Birth & Day of Date of Birth Surname & Date of Birth • Where duplicates were confirmed priority was given to the client record with the latest attendance date • The exercise identified 270 potential duplicates but following manual review this was reduced to 132 confirmed duplicates
Records after Manual De-Duplication	7216	• 124 records had no GP recorded (or GP recorded as unknown) • Match of PAS GPs to the master GP list resulted in 6916 records having a valid GMS GP code

Table 2.11: Description of data cleansing and manual de-duplication within LIS file

Laboratory Information System (LIS)	Numbers	Notes
Total Clients Extracted	14624	
Incomplete Records Removed	17	• Records rejected because they did not have a surname. These records also had no address
Duplicates	2923	• De-duplication combinations used for matching included: Patient ID, Forename, Surname, Date of Birth & First Address Line Patient ID, Forename, Surname & Date of Birth Patient ID, Surname & Date of Birth Patient ID Forename, Surname, Date of Birth & First Address Line Forename, Surname & Date of Birth Surname, Forename, Year of Date of Birth & Day of Date of Birth Surname, Forename, Year of Date of Birth & Month of Date of Birth Surname, Forename, Month of Date of Birth & Day of Date of Birth Surname & Date of Birth • The above de-duplication exercises were undertaken first. However, as the volume of clients still remaining on the file was very high a review was undertaken to identify how other potential duplicates could be found • The following de-duplication combinations were subsequently identified and used: Forename, Surname & First Address Line Forename & First Address Line Surname & First Address Line • Where duplicate records had a mix of 'High' and 'NR' in the Result field priority was given to the record with the 'High' result • The exercise identified 6924 potential duplicates but following manual review this was reduced to 2923 confirmed duplicates
Records after Manual De-Duplication	11684	• 3826 records had no GP recorded • Match of LIS GPs to the master GP list resulted in 7818 records having a valid GMS GP code

Fuzzy De-Duplication (using Fuzzy Grouping) Within One Data Source

Following manual de-duplication of MCP, LTI, PAS and LIS data files it was recognised that the process used, while manageable for the volumes involved in the individual data sources, was extremely labour intensive and time consuming and would not be practical for the volumes which would be involved once merging of the data sources into one single file commenced.

A toolkit, called Fuzzy Logic provided with Microsoft SQL Server 2005 was reviewed to consider if it might be a suitable tool to assist with de-duplication across data sources. Fuzzy Logic provides a way of comparing records which are unlikely to be exact matches; it assumes there will be sufficient matching elements in the data to be able to determine the probability of a match. Fuzzy Logic uses scoring (algorithms) to identify duplicates. Fuzzy Logic provides two options: Fuzzy Lookup which allows the comparison of records in two different files and Fuzzy Grouping which is used to compare records within the same file. As Fuzzy Logic was being considered for the across source de-duplication it was decided to use the tool to de-duplicate the HIPE file using Fuzzy Grouping as part of the evaluation of the tool. Fuzzy Logic is outlined in detail in Phase 4.

The outcome of data cleansing and de-duplication using Fuzzy Grouping within HIPE is presented in Table 2.12.

Table 2.12: Description of data cleansing and de-duplication using Fuzzy Grouping within HIPE file

Hospital In-Patient Enquiry (HIPE)	Numbers	Notes
Total Clients Extracted	6171	
Incomplete Records Removed	0	
Duplicates	1783	• De-duplication combinations used for matching included: – Medical Record Number (1.0) + Surname (1.0): an exact match – Surname (≥0.7) + Date of Birth (≥0.75) + overall match score (≥0.7) – Forename (no minimum) + Surname (no minimum) + Date of Birth (1.0) + Full Address (no minimum) + overall match score (≥0.8) • A high level of duplication would be expected given that the data spanned a three year period • The exercise identified 1783 potential duplicates, all of which were confirmed as duplicates following manual review
Records after Fuzzy Grouping De-Duplication	4388	• HIPE does not maintain GP demographic/identification data but as part of matching exercise to PAS, to capture additional data fields, GP identifiers were extracted where available. Following the match to PAS 163 records had no GP recorded (or GP recorded as unknown) • Match of HIPE GPs to the master GP list resulted in 4179 records having a valid GMS GP code

Table 2.13 summaries the data available following completion of Phase 2.

Table 2.13: Summary of de-duplication within each data source

	MCP	LTI	PAS	HIPE	LIS
Total Clients Extracted	7083	4306	7348	6171	14624
Incomplete Records Removed	1009	219	0	0	17
Duplicates	3	11	132	1783	2923
Records after De-Duplication (Phase 2)	**6071**	**4076**	**7216**	**4388**	**11684**

Phase 3: Data Source Hierarchy

The third phase of the feasibility study was to select the data source hierarchy. This was to decide on which data sources were to take priority when merging the data sources into one single file.

In determining the data source hierarchy the following factors were considered:

- Coverage; if the data source was available on a national or local basis
- Data quality of each data source
- Availability of GP demographic/identification data within each data source to facilitate the validation process (See Phase 6 – GP Visit Verification)

The data source hierarchy selected for the purpose of the feasibility study was:

1. Medical Card Prescriptions (MCP)
2. Long Term Illness (LTI) Scheme
3. Patient Administration System (PAS)
4. Hospital In-Patient Enquiry (HIPE) Scheme
5. Laboratory Information System (LIS)

1. MCP was selected as the primary data source. The data is held centrally by PCRS and therefore is available nationally. The full client demographic data set required would be available from PCRS if the data was sourced directly from them. However, for the purpose of this feasibility study, as the full demographic data was not available directly in the data source file, the additional fields required were obtained from the Mid West local Careworks Schemes System. This system contains 12 separate modules/schemes and demographic data would therefore be updated on a regular basis by HSE staff as clients avail of the associated services. A business unit in the HSE West reviews schemes data on an ongoing basis ensuring PPS numbers are populated (where possible) and validated, and deceased clients identified using DEPS and other local information sources. PPSN is a mandatory data field on the Mid West Careworks Schemes Client Index since April 2005. The data quality of this source would therefore be deemed very good. In addition, most client records had an associated GP identifier which contained a GMS GP code which is a nationally recognised code for a GP.

2. LTI is a national scheme and was given second priority. All LTI scheme applications from the HSE areas are sent to PCRS. Similar to MCP data, as the LTI clients were extracted from the Mid West local Careworks Schemes System, client demographics would be subject to the same data quality exercises. However, LTI records would not be subject to review on a similar basis to medical cards and therefore if a client was not accessing other schemes on the Careworks Schemes System, the client address and GP identifier (where recorded) may be somewhat historic. GP demographic/identification data where available could be obtained from the LTI module (at the time the study was undertaken GP was not a mandatory field on Careworks Schemes for LTI clients) or it could be extracted from a related module e.g. medical card module. The GP demographic/identification data for the LTI clients contained the GMS GP code.

3. PAS was placed in third position. PAS and HIS applications are available in almost all Irish hospitals but the level of information captured in relation to diabetic outpatient and inpatient attendances would vary greatly depending on the hospital site and the applications available to them. There are multiple different PAS and HIS products in place around the country. In the Mid West, demographic and GP information on PAS would be validated with patients on admission and on attendance at A&E, outpatient and radiology departments etc and therefore data quality should be of a high standard. While GP demographic/identification data is recorded on PAS for clients, PAS does not capture any nationally recognised GP code. There are multiple PAS sites within the Mid West and each site maintains its own GP codes. Therefore, it is possible to have multiple instances of the same GP across the Mid West PAS sites (e.g. Dr. John Ryan, Test Street, Limerick could have a different code on the Mid Western Regional Hospital Limerick PAS site, Ennis General Hospital PAS site and Nenagh General Hospital PAS site). In addition, PAS stores the full GP name (i.e. Forename and Surname) in the one data field so the creation of duplicate GPs is a particular problem for hospitals in the Mid West.

4. HIPE was placed in fourth position. HIPE collects clinical and administrative data on discharges from, and deaths in, acute hospitals nationally. An electronic file is forwarded on a monthly basis to the HIPE Unit in the ESRI. However, very limited demographic information, other than date of birth and surname is available on HIPE. No GP demographic/identification data is held on HIPE.

5. LIS was placed in last position. HbA_{1C} test results are recorded on LIS and it was assumed that each diabetic would have the test approximately once a year. Logically laboratory information systems should therefore be a good central source for diabetic register data. However, there are a number of different Laboratory Information Systems used around the country. Local knowledge indicated that there is some doubt around the quality of data on laboratory systems as in most cases the lab only receives a specimen and request form and therefore staff cannot validate the client demographic information directly with the client. This results in poor or incomplete client data or in many cases the generation of duplicate client records. Also, in the Mid West five hospital PAS sites (Mid Western Regional Hospital Limerick, Ennis General Hospital, Nenagh General Hospital, Mid Western Regional Maternity Hospital and Mid Western Regional Orthopaedic Hospital) are linked to the LIS and add individual client records often creating duplicate clients on the database. In many cases the client's GP may not be available as the HbA_{1C} test may be requested by a Consultant. The Mid West LIS maintains its own GP code and does not store any nationally recognised GP code.

Phase 4: De-Duplication Across Data Sources

Due to the volume, nature and variation of the data within each data source, and the combination of iterative checking and manual checking of duplicates (Phase 2), it was felt that this process would become less reliable as different data sources were merged into one single file. In addition, the process was very labour intensive, time consuming and would not be practical for the volumes involved. It was recognised that the development of the register would require a greater degree of automation in order to identify and eliminate duplicates.

The identification of duplicate clients is difficult in the absence of a common unique identifier across the data sources. Where a unique identifier exists for each client, such as PPSN, then the exercise would be relatively straightforward. A unique identifier, such as PPSN, was only available in a limited number of client records i.e. client records from MCP and LTI only.

For a large number of client records the identification of duplicate clients was only possible using a combination of 'Forename', 'Surname', 'Date of Birth' and 'Address' fields. However, this was problematic as there was no consistency across the data source fields in terms of naming or address conventions. A person could have a forename of 'Joe' in one data source, 'Joseph' in another and 'John Joseph' in a third. The address could all be within one field or spread over a number of fields. Even if the address was within one field the formatting may be different e.g. separated by commas or not. An address could be '3 Main Street, Limerick' in one data source or '3 Main St., Limerick' in another.

Generally computer systems use an 'exact' match to identify records which are duplicates within or across data sources, in which case '3 Main Street, Limerick' and '3 Main St., Limerick' would appear as two different records. In Phase 2 the approach was to search for exact matches and then rely on manual checking of the potential duplicates identified. This was a very slow and time consuming process and to a certain extent unreliable. Given the number of client records (33435) in the five data sources it was decided that an alternative method of identifying duplicates would have to be considered.

Microsoft SQL Server 2005 provides a toolkit, called Fuzzy Logic, which assists with the management of unstructured data (data not in a uniform format). Fuzzy Logic provides a way of comparing records which are unlikely to be exact matches; it assumes there will be sufficient matching elements in the data to be able to determine the probability of a match.

Using an exact match de-duplication process would not identify the pairs of clients below as duplicates.

Source	Forename	Surname	Date of Birth	Full Address
1	Joseph	O'Connell	20/12/1974	8 Main Street, Ardagh, Co. Limerick
2	Joe	O Connell	20/12/1974	Main St. Ardagh Limerick
1	Anthony Tony	O Halloran	9/8/1925	Newport, Thurles, Co. Tipperary
2	John A	O Holloran	9/7/1925	Newport Thurles
1	Cornelius	Sullivan	27/10/1945	Inagh Ennis Co. Clare
2	Con	Sullivan	27/10/1945	Inagh Ennis

However, the Fuzzy Logic process would assess a likely match based on an element-by-element basis. For example, in the first pair of address it would pick out 'Main', 'Ardagh', 'Limerick' and consider it very likely that this pair was the same client.

Fuzzy Logic provides two options: Fuzzy Lookup which allows the comparison of records in two different files and Fuzzy Grouping which is used to compare records within the same file. As outlined in Phase 2 the de-duplication of one of the data sources HIPE was undertaken using Fuzzy Grouping as part of the evaluation of the toolkit and its possible use for across data source de-duplication. Fuzzy Logic, following the review, was considered an effective method which would greatly assist with de-duplication across the data sources.

The first stage in the Fuzzy Logic process was to create a uniform data structure for all data sources to ensure that tests and comparisons to identify duplicates would be as reliable as possible, i.e. comparing like with like. If one data source had the address stored in 3 address line fields and another data source had the address in 5 address line fields it would be difficult to compare these. In order to create a uniform data structure for address, all address lines were amalgamated into one full address field. The data available from each data source was considered and the following uniform data structure was agreed.

Forename,
Surname,
Date of Birth,
First Address Line,
Full Address (amalgamation of all address information within one field)
PPSN (not available in all),
Medical Card Number (not available in all).

The data from each data source was converted to the above standard format.

A Fuzzy Model was developed to manage the de-duplication process. The Model was based on comparing each data source against a master data source, and then adding the new records identified in the data source to the master before the next data source was processed. In Phase 3 it was decided on the order of priority each data source was to be given when merging the data sources. A blank master file was populated with the primary data source, i.e. Medical Card Prescriptions. Each data source was compared against the master file in the predetermined hierarchy order. At the end of the process the master file contained unique records from all the data sources.

The Fuzzy Model consisted of four parts and was defined as:

a) *the combination of data elements to be matched,*

Combinations of data elements varied depending on the data source, for example:

1. Forename + PPSN
2. Forename + Surname + Date of Birth + Full Address

b) *the minimum match score allocated to each data element,*

Fuzzy Logic allowed the setting of minimum scores for a specific data element, for example:

1. Forename (≥0.7) + PPSN (1.0)
2. Forename (≥0.5) + Surname (≥0.75) + Date of Birth (≥0.75) + Full Address (no minimum)

In example 1, Forename (≥0.7) and PPSN (1.0) meant that Forename had to be a 70% match or greater and PPSN had to be a 100% match.

c) *the overall match score,*

An overall match score was set for the combination of data elements, for example:

1. Forename (≥0.7) + PPSN (1.0) + overall match score (≥0.7)
2. Forename (≥0.5) + Surname (≥0.75) + Date of Birth (≥0.75) + Full Address (no minimum) + overall match score (≥0.7)

In example 1, in addition to meeting the minimum scores set out, the overall match score had to be a 70% match or greater before it was deemed a duplicate. In example 2, minimum scores were set for 'Forename', 'Surname' and 'Date of Birth'. No minimum score was set for 'Full Address' but the overall match score for the combined data elements had to be 70% match or greater to be deemed a duplicate.

d) the sequence in which the tests were carried out.

Tests were sequenced to ensure that the matching combinations likely to capture the greatest number of duplicates and use the least 'computing' resources (e.g. processing power) were run first. In this way the number of records each subsequent test had to process was reduced. The exact match test was carried out first. Fuzzy matches use significantly more 'computing' resources than exact matches. Fuzzy matches which had more fields and fields with greater volumes of data (such as full address) to compare were more resource intensive, particularly where minimum scores were applied.

As each data source was added, Fuzzy Grouping was used to quality assure the Fuzzy Model. By comparing records within the master file Fuzzy Grouping confirmed that the Fuzzy Model was effective in identifying duplicates. The output from the Fuzzy Grouping tests was manually reviewed by HSE staff. This exercise confirmed that the Fuzzy Model used for each data source captured the optimum level of duplicates; only 21 additional duplicates were identified across the data sources.

Refinement of Fuzzy Logic Model

During the Fuzzy Logic process the individual data elements and overall match scores were reviewed and refined as each data source was added. Deciding on the precise match scores for each data source is iterative and takes some time and needs to take account of the data available in each data source. This involved a series of test runs and manually reviewing the output until the optimum match level was determined. The Fuzzy Logic output was categorised as; matched records, possible matches and new records.

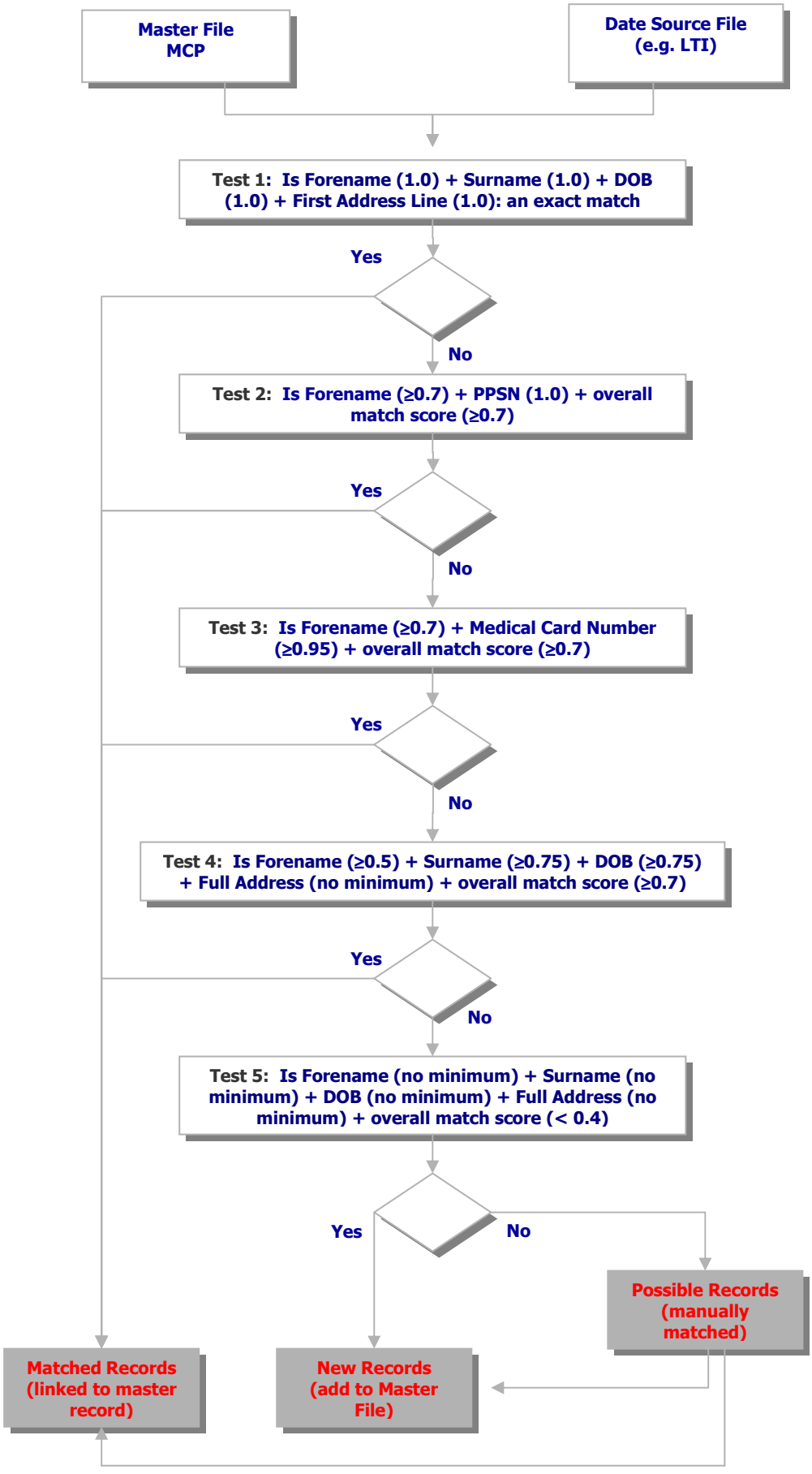
- Matched Records: These were definite duplicates and were linked to the master record as they were encountered.
- Possible Matches: These were possible duplicates and were manually matched using a manual match application developed in Microsoft Access specifically for this purpose. The application was designed to take each possible duplicate client and use Fuzzy Logic to display the top ten likely matches on the master file. The overall match probability score calculated by Fuzzy Logic was displayed for each client. Any duplicate records identified during this process were linked to the master record. Remaining records were marked as new client records.
- New Records: These were definite new clients and were added to the master file before the next data source was processed.

A sample of the output records at the margins were examined to determine if the score (tolerance) levels required adjustment or not. This manual examination was used in order to set the score levels appropriate for each test for each individual data source.

Following the addition of PAS the volume of clients on the master file was 12929 clients. A review was undertaken to ensure that the identification of duplicate clients was maximised. As part of the review it was agreed that client records from the subsequent data sources (HIPE and LIS) should be compared to the 'Matched' client records in addition to the 'Master' client records. It was expected that this additional test would increase the number of duplicate records automatically matched and reduce the possible matches for manual matching.

For example: 'John Ryan, Test Street, Limerick (DOB 18/11/1950)' was added as the 'Master' client from MCP. As the processing of LTI was undertaken 'John Joseph Ryan, Sky Road, Ennis (DOB 18/11/1950)' was identified as a duplicate record from the 'Possible Matches' during the manual match process and linked to the original John Ryan on the master file (the LTI record was now deemed a 'Matched' client record). Matching against the 'Matched' client records would ensure that the demographic information associated with both the 'Master' record and any 'Matched' records was checked. If 'Joseph Ryan, Blue Street, Ennis (DOB 18/11/1950)' was a record in HIPE it would be compared with both sets of demographics already captured for this client from MCP and LTI. Matching against the 'Master' client would only have compared the record with the MCP client demographics.

The flow chart graphically illustrates an example of the Fuzzy Logic process.



Tables 2.14-2.17 set out the test sequence used for each data source. Initially, the master file was populated with all 6071 MCP records and each of the other data sources were processed in the hierarchy order.

Table 2.14: Fuzzy Logic test sequence for LTI

Test	Pass Test	Fail Test	Notes
1. Is Forename (1.0) + Surname (1.0) + DOB (1.0) + First Address Line (1.0): an exact match	Assume matched record	Go to test 2	
2. Is Forename (≥0.7) + PPSN (1.0) + overall match score (≥0.7)	Assume matched record	Go to test 3	
3. Is Forename (≥0.7) + Medical Card Number (≥0.95) + overall match score (≥0.7)	Assume matched record	Go to test 4	A (≥0.95) Medical Card Number match allowed the trailing letter on the Medical Card to be incorrect. The trailing letter recorded may be unreliable
4. Is Forename (≥0.5) + Surname (≥0.75) + DOB (≥0.75) + Full Address (no minimum) + overall match score (≥0.7)	Assume matched record	Go to test 5	
5. Is Forename (no minimum) + Surname (no minimum) + DOB (no minimum) + Full Address (no minimum) + overall match score (< 0.4)	Assume new client and add to master file	Manually match those above 0.4	Manual matching resulted in records categorised as matched or new records

Table 2.15: Fuzzy Logic test sequence for PAS

Test	Pass Test	Fail Test	Notes
1. Is Forename (1.0) + Surname (1.0) + DOB (1.0) + First Address Line (1.0): an exact match	Assume matched record	Go to test 2	
2. Is Forename (≥0.4) + Surname (≥0.75) + DOB (≥0.75) + Full Address (no minimum) + overall match score (≥0.6)	Assume matched record	Go to test 3	
3. Is Forename (no minimum) + Surname (no minimum) + DOB (no minimum) + Full Address (no minimum) + overall match score (< 0.4)	Assume new client and add to master file	Manually match those above 0.4	Manual matching resulted in records categorised as matched or new records

Table 2.16: Fuzzy Logic test sequence for HIPE

Test	Pass Test	Fail Test	Notes
1. Is Forename (1.0) + Surname (1.0) + DOB (1.0) + First Address Line (1.0): an exact match	Assume matched record	Go to test 2	
2. Is Forename (≥0.7) + Medical Card Number (≥0.95) + overall match score (≥0.7)	Assume matched record	Go to test 3	A (≥0.95) Medical Card Number match allowed the trailing letter on the Medical Card to be incorrect. The trailing letter recorded may be unreliable
3. Is Forename (≥0.4) + Surname (≥0.75) + DOB (≥0.75) + Full Address (no minimum) + overall match score (≥0.6)	Assume matched record	Go to test 4	
4. Is Forename (≥0.4) + Surname (≥0.75) + DOB (≥0.75) + Full Address (no minimum) + overall match score (≥0.6)	Assume matched record	Go to test 5	This was a new test added to check the client demographics against the 'Matched Records'. Test 3 checked the client demographics against the 'Master Records' as was done for the previous data sources
5. Is Forename (no minimum) + Surname (no minimum) + DOB (no minimum) + Full Address (no minimum) + overall match score (< 0.4)	Assume new client and add to master file	Manually match those above 0.4	Manual matching resulted in records categorised as matched or new records

Table 2.17: Fuzzy Logic test sequence for LIS

Test	Pass Test	Fail Test	Notes
1. Is Forename (1.0) + Surname (1.0) + DOB (1.0) + First Address Line (1.0): an exact match	Assume matched record	Go to test 2	
2. Is Forename (≥0.4) + Surname (≥0.75) + DOB (≥0.75) + Full Address (no minimum) + overall match score (≥0.6)	Assume matched record	Go to test 3	
3. Is Forename (≥0.4) + Surname (≥0.75) + DOB (≥0.75) + Full Address (no minimum) + overall match score (≥0.6)	Assume matched record	Go to test 4	As per test 4 in HIPE
4. Is Forename (no minimum) + Surname (no minimum) + DOB (no minimum) + Full Address (no minimum) + overall match score (< 0.4)	Assume new client and add to master file	Manually match those above 0.4	Manual matching resulted in records categorised as matched or new records

Table 2.18 summaries Phase 4. At the start of the phase, before the data sources were merged and de-duplicated across the data sources, there were 33435 client records. MCP (6071 client records) was selected as the primary data source and these client records were first added to the blank master file. LTI (4076 client records) was next selected and each LTI client record was compared with each record on the master file. The Fuzzy Logic process identified 570 client records (13.9% of original LTI data source) that were already in MCP, therefore 3506 client records (86.1%) were added to the master file. PAS (7216 client records) was next selected, 2486 client records (34.5%) were already in MCP and 1378 client records (19.1%) in LTI, thus 3864 (53.6%) duplicates identified and 3352 (46.4%) new client records added to the master file. This process was repeated for HIPE and LIS data sources.

This exercise shows that when using the selected hierarchy approach

- 13.9% of LTI records were already in MCP
- 53.6% of PAS records were already in MCP and LTI
- 71.2% of HIPE records were already in MCP, LTI and PAS
- 57.7% of LIS records were already in MCP, LTI, PAS and HIPE.

In summary, at the end of Phase 4 following the de-duplication process, 14298 duplicates were identified resulting in 19137 client records on the master file.

Table 2.18: Summary of Fuzzy Logic de-duplication across data sources

		MCP	LTI	PAS	HIPE	LIS	
Records after De-Duplication (Phase 2)		**6071**	**4076**	**7216**	**4388**	**11684**	
Duplicates Identified:	**MCP**	-	570 (13.9%)	2486 (34.5%)	1934 (44.1%)	3749 (32.1%)	
	LTI	-	-	1378 (19.1%)	588 (13.4%)	1845 (15.8%)	
	PAS	-	-	-	604 (13.7%)	678 (5.8%)	
	HIPE	-	-	-	-	466 (3.9%)	
	LIS	-	-	-	-	-	
Total Duplicates Identified		**-**	**570 (13.9%)**	**3864 (53.6%)**	**3126 (71.2%)**	**6738 (57.7%)**	**14298 (42.8%)**
Records Added to Master File by Source		**6071**	**3506 (86.1%)**	**3352 (46.4%)**	**1262 (28.8%)**	**4946 (42.3%)**	**19137**

Phase 5: Refinement

Following de-duplication of client records across the data sources there were 19137 client records on the master file. It was believed that this number was high when compared to estimated prevalence figures for diabetes in the Mid West region published by the Institute of Public Health in Ireland. They estimated that approximately 12495 adults aged 20 years and older in the Mid West had diabetes (diagnosed and undiagnosed) in 2007 [2]. As this estimate included undiagnosed diabetes (the estimated number of diagnosed diabetics would therefore be less than 12000 adults) there was a considerable difference between the two sources. It was decided to further analyse the client records on the master file to identify if the numbers could be further refined.

Capture of Additional Data

In order to support further refinement of the master file an exercise was first undertaken to capture additional data, i.e. date of death, PPSN if missing and county of residence if missing.

- Date of death was important as some of the clients on the master file may have died since the data was extracted.
- A number of client records on the master file were missing PPSN. Obtaining the missing PPSN was important as it would further assist with removing duplicate client records.
- Some client records on the master file did not have the county name in their address. The register was initially to contain diabetic clients resident in counties Clare, Limerick and North Tipperary. A number of clients on the register were resident outside the Mid West region, i.e. from neighbouring counties. These clients were captured as they were availing of the healthcare services in the Mid West. In order to compare to the Institute of Public Health estimated prevalence figure it was proposed to temporarily mark clients resident outside the Mid West region as inactive on the master file.

A process similar to the fuzzy matching process used for across data source de-duplication was carried out to capture the additional data. The master file was matched to the Careworks Schemes System to capture 'Date of Death', 'PPSN' and 'County of Residence' where missing on the master file. The matching criteria used and the order of the tests was:

- Forename (1.0) + Surname (1.0) + Date of Birth (1.0) + First Address Line (1.0): an exact match
- Medical Card Number (≥0.95) + Forename (≥0.7) + overall match score (≥0.7)
- Forename (≥0.4) + Surname (no minimum) + Date of Birth (≥0.75) + Full Address (no minimum) + overall match score (≥0.4397)

Deceased

Some clients on the master file may have died since the data was extracted. The match to Careworks Schemes System, to capture 'Date of Death' if available, identified 1275 deceased clients. These clients were removed[†] from the master file.

Duplicate PPSN

The match to Careworks Schemes System captured an additional 261 PPSNs. The master file was then de-duped using PPSN. 62 potential duplicates were identified from this exercise but following manual review 53 were confirmed as duplicates. 9 clients had the incorrect PPSN recorded for them.

Resident outside the Mid West Region

The Mid West region includes Clare, Limerick and North Tipperary. A number of client records on the master file had an address outside the region, i.e. in bordering counties and South Tipperary. Two issues complicated identification of the non-region client records. Firstly, a significant number of the records (918) did not have any county name in the address and secondly, addresses did not differentiate between North and South Tipperary. Following the capture of additional data, as

[†] marked as inactive on the master file

outlined above, 486 records remained with no county name in the address. These records were manually reviewed and any residents outside the Mid West were removed†.

To differentiate between addresses in North and South Tipperary a number of SQL queries were run to extract the town from the Tipperary records. These towns were then compared to Tipperary electoral divisions or townlands to determine whether the person lived in North or South Tipperary. Any South Tipperary addresses were removed† from the master file. The total number of clients removed from outside the region was 729.

HbA_{1c} Results

The master file included clients with a HbA_{1c} test record where the result was either High (greater than 6.0%) or within the upper part of the normal range (between 5.5% and 6.0%). Following consultation with a Consultant Endocrinologist it was decided to remove† those clients (2673) within the upper part of the normal range HbA_{1c}. It was believed that the majority of the clients in this range were not diabetic.

A summary of the refinement phase is presented in Table 2.19. 4730 client records were removed† from the master file; 1275 clients deceased since the data extraction period, 53 clients had a duplicate PPSN, 729 clients were resident outside the Mid West region and 2673 clients had a HbA_{1c} within the upper part of the normal range. Following refinement 14407 client records remained on the master file. A complete summary of the data source analysis is presented in Appendix 1.

Table 2.19: Summary of master file refinement

	MCP	LTI	PAS	HIPE	LIS	Total
Records added to Master File by Source	**6071**	**3506**	**3352**	**1262**	**4946**	**19137**
Deceased	214	29	485	402	145	1275
Duplicate PPSN	0	2	41	2	8	53
Resident outside Mid West Region	10	14	167	167	371	729
HbA_{1c} Upper Part of Normal Range	-	-	-	-	2673	2673
Total Unique Records	**5847**	**3461**	**2659**	**691**	**1749**	**14407**

Comparison with Institute of Public Health in Ireland

A report [2] published by the Institute of Public Health in Ireland estimated that 4.5% of all adults, aged 20 years and older, had diabetes (type 1 and 2 combined) in 2007. Estimates of population prevalence of adult diabetes for the Mid West region, as published by the Institute of Public Health, are presented in Table 2.20. These figures are adjusted for the population of the Mid West and include diagnosed and undiagnosed diabetes. The percentage of undiagnosed diabetes in the Republic of Ireland is not known. The Institute of Public Health estimated that almost a tenth (9.8%) of all cases of diabetes (type 1 and type 2 combined) in people aged 17 years and over in Northern Ireland are undiagnosed (Institute of Public Health, Pers. Comm., 2010). This assumption of undiagnosed diabetes was applied, by the project team, to the Institute of Public Health's Mid West estimated prevalence. Therefore, it could be assumed that an estimated 11270 people aged 20 years and over had diagnosed diabetes in the Mid West in 2007.

Table 2.20: Estimated prevalence of diagnosed and undiagnosed diabetes in adults aged 20 years and older in the Mid West in 2007

	Estimated Prevalence	Estimated Number	Undiagnosed ‡	Diagnosed
Clare	4.8%	3814	374	3440
Limerick	4.7%	5338	523	4815
North Tipperary/East Limerick	4.6%	3343	328	3015
		12495	**1225**	**11270**

‡ Undiagnosed figures based on 9.8% as per Northern Ireland average

The estimated prevalence of diagnosed diabetes in the Mid West was compared to the number of clients on the master file. This was to identify if the final number in the master file was similar to the estimate of diagnosed diabetes for the Mid West region. As the master file contained clients aged under 20 years, these clients were temporarily excluded from the final figure in order to make comparisons. This resulted in 424 clients excluded from the final figure, as shown in Table 2.21.

Table 2.21: Number of clients aged 20 years and older on the master file

	MCP	LTI	PAS	HIPE	HIS	Total
Total Unique Records	**5847**	**3461**	**2659**	**691**	**1749**	**14407**
Aged Under 20 Years	82	144	180	4	14	424
Final Total	**5765**	**3317**	**2479**	**687**	**1735**	**13983**

The comparison test showed that the number on the master file (13983) was 19.4% higher than the estimated prevalence of diagnosed diabetes (11270), assuming the Northern Ireland average of 9.8% undiagnosed diabetes, in the Mid West region.

One possible explanation for the difference could be methodological issues associated with the Institute of Public Health estimates. The Diabetes Population Prevalence model used by the Institute was developed in the UK and accounts for age, sex, ethnicity and socio-economic factors. Age, sex and ethnicity estimates of population prevalence, from UK reference population studies, were applied to the Irish population. Differences between the Irish and UK population could account for the disparities between the two sources. The IPH estimates do not include a calculation of confidence intervals.

Another possible reason could be that the percentage of undiagnosed diabetes in the Mid West could be different to the Northern Ireland average percentage (9.8%) which was applied to the Institute's estimated population figure to make an assumption of the prevalence of diagnosed diabetes in the Mid West.

A third reason for the difference could be an overestimate of people with diabetes on the master file following the merging of the five data sources in the Mid West. In order to verify that there was no further duplication and that those on the master file were true diabetics a number of verification tests were performed, as outlined in Phase 6.

Phase 6: Verification

This Phase involved verification of a sample of the client records within the master file. This was a final check to verify that there were no further duplicates on the master file and to ensure that those on the master file were true diabetics.

Three different verification methods were used:

- Surname Verification
- Laboratory Information System Verification
- GP Visit Verification

Surname Verification

Clients on the master file with common surnames, 'McNamara' and 'O'Connor' (386 client records), were manually checked to see if any duplicates existed. The main purpose of this exercise was to check that the Fuzzy Model used for the across data source de-duplication had no significant flaws and had not missed any obvious matches. By comparing the two sample surnames manually it was determined that this was not the case. Of the 386 client records examined only two sets of duplicate records were identified, which had completely different addresses. This could only be determined by manually checking the records.

Laboratory Information System Verification

The LIS data in this study was extracted for the full year 2007. Clients who had a HbA_{1c} test in another year were not captured. It was assumed that the majority of clients on MCP, LTI, PAS and HIPE would have at least one HbA_{1c} done in any year and therefore appear on LIS. A sample of 40 records which did not appear on the LIS data file but were in the other data sources were checked to see if an explanation could be found. Ten client records were taken from each of MCP, LTI, PAS and HIPE. The outcome is presented in Table 2.22. The main reason the 40 client records selected did not appear on LIS data was that they had HbA_{1c} tests in a year other than 2007 (28 clients) and the remainder (12 clients) had no record of HbA_{1c} tests being done on the laboratory information system.

Table 2.22: Reasons why clients on other data sources were not on LIS

Data Source	HbA_{1c} tests taken in year other than 2007	No HbA_{1c} records on LIS application	Total
MCP	5	5	10
LTI	6	4	10
PAS	10	0	10
HIPE	7	3	10
Total	**28**	**12**	**40**

GP Visit Verification

The GP visit verification exercise was to validate the diabetic client list for a sample of GPs on the master file with the GP's own practice list. There are 218 GPs in the Mid West region and a sample of 16 GPs was selected. Details of the selected GPs are presented in Appendix 2. GP practices were selected to capture a mix of the following:

- Location: Local Health Office area, urban/rural
- Practice Type: single handed/partnership
- Practice Nurse: yes/no
- Practice Management System: HEALTHOne, Dynamic©GP, GP Mac, Clinical Objects, Socrates, no electronic practice management system.

Visits to the GPs were arranged through the practice nurse where they existed. Two single handed GP practices did not have a practice nurse and the GPs declined visits. Subsequently, one of the GPs validated the client list sent to him and returned with corrections. However, the GP was unable to identify any new clients. One GP raised issues in relation to client confidentiality and declined to participate at the time. In advance of the GP visits the diabetic client list for each of the sample GPs

was extracted from the master file, an example of the list is in Appendix 3. A client diabetic register form (Appendix 4) with patient demographics was produced for each diabetic on the GP diabetic client list. Actual visits took place to 13 GPs by a Community Diabetes Care Facilitator (Diabetes Nurse Specialist).

The diabetic client list was checked against the GP practice list which was extracted from the electronic practice management system. In many cases the GP practice list was not available, three practices used the practice management system entirely for producing prescriptions and no clinical data was recorded. In other practices it was not possible to create a computerised list due to data quality and the way in which the data was recorded. To accurately identify diabetics in a practice management system it is essential that all clinical data is coded by disease/medical condition. Some GPs asked the Community Diabetes Care Facilitator to code diabetic clients found on the practice management system (and this was facilitated) in order to create a comprehensive list of their known diabetics.

The GP verification process was both time consuming and labour intensive, taking an experienced nurse with good knowledge of all practice management systems approximately a day per practice to complete this task. Each client on the diabetic client list was checked against the extracted GP practice management list. Clients on the diabetic client list were marked as 'validated diabetic', 'not diabetic per GP', 'unknown' 'deceased' or 'duplicates'. The client diabetic register form for validated clients was updated with any demographic changes and any additional information if available from the practice management system. Some clients on the diabetic client list were known to the GP but deemed to be non-diabetic. These were marked on the diabetic client list as 'not diabetic per GP'. Some clients on the diabetic client list were not patients of the GP and were marked as 'unknown'. Clients who had since died were marked as 'deceased' and duplicate clients identified by the GP were marked as 'duplicates'. Blank client diabetic register forms were completed for any additional clients identified by the GP practice.

The results of the GP visit verification exercise are presented in Table 2.23. Percentages are shown in brackets. Further details for each GP are presented in Appendix 5.

Table 2.23: Summary of findings from GP visit verification exercise

	MCP	LTI	PAS	HIPE	LIS	Not on Master File	Total
Diabetic Client List	452	238	214	39	183		1126
Validated Diabetic	377 (83.4)	186 (78.2)	31 (14.5)	19 (48.8)	91 (49.7)		704 (62.5)
Not Diabetic per GP	40 (8.9)	22 (9.2)	91 (42.5)	7 (17.9)	67 (36.7)		227 (20.2)
Unknown	30 (6.6)	29 (12.2)	65 (30.4)	7 (17.9)	20 (10.9)		151 (13.4)
Deceased	1 (0.2)	0	25 (11.7)	5 (12.8)	2 (1.1)		33 (2.9)
Duplicates	4 (0.9)	1 (0.4)	2 (0.9)	1 (2.6)	3 (1.6)		11 (1.0)
Clients on Master File under Different GP or no GP Recorded	52	21	4	1	9		87
New						28	28
Total	**504**	**259**	**218**	**40**	**192**	**28**	**1241**

The headings in the table are explained as follows:

- Diabetic Client List – Number of records on the diabetic client list produced for GP visits.
- Validated Diabetic – Clients on the diabetic client list the GP verified as diabetic.
- Not Diabetic per GP – Clients on the diabetic client list the GP considered should not have been classified as diabetic.
- Unknown – Clients on the diabetic client list not patients of the GP.
- Deceased – Clients on the diabetic client list the GP identified as deceased.
- Duplicates – Clients on the diabetic client list the GP identified as duplicate clients.
- Clients on Master File under Different GP or no GP Recorded – Additional clients identified by the GP practice as diabetic who were not on the diabetic client list produced for the GP visits. These clients were on the master file but had a different GP recorded or had no GP recorded.

- New – Additional clients identified by the GP practice as diabetic who were not on the diabetic client list produced for the GP visits and were not on the master file.
- Total – Number of clients on the diabetic client list produced for the GP visits plus additional clients identified by GP as 'clients on master file under different GP or no GP recorded' and 'new'.

Validated diabetics within each individual data source ranged from:
- 83.4% for MCP
- 78.2% for LTI
- 14.5% for PAS
- 48.8% for HIPE
- 49.7% for LIS

227 (20.2%) clients were identified by GPs as not diabetic. An additional 33 (2.9%) clients were deceased and 11 (1.0%) clients were duplicates. A total of 24.1% of those on the diabetic client list for the sample GPs were invalid for the reasons outlined.

Clients classified as 'not diabetic per GP' within each individual data source ranged from:
- 8.9% for MCP
- 9.2% for LTI
- 42.5% for PAS
- 17.9% for HIPE
- 36.7% for LIS

Clients identified as 'deceased' and 'duplicates' totalled in the individual data sources to:
- 1.1% for MCP
- 0.4% for LTI
- 12.6% for PAS
- 15.4% for HIPE
- 2.7% for LIS

Clients classified as 'unknown' within each individual data source ranged from:
- 6.6% for MCP
- 12.2% for LTI
- 30.4% for PAS
- 17.9% for HIPE
- 10.9% for LIS

However, as the GP verification exercise only represented a 7.8% sample of the master file clients, the unknown clients could be clients of other GPs in the Mid West area.

Of the 115 additional clients, 87 had client records on the master file but had a different GP recorded (34) or had no GP recorded (53). 28 clients identified by the GP practices did not exist on the master file and were therefore added. These clients were diagnosed within the data extraction periods but were not picked up as they were either private patients, did not have a LTI card, did not attend diabetic outpatients or have a HbA_{1c} recorded.

The 227 clients (20.2%) deemed 'not diabetic per GP' appeared across all data sources. Further analysis of these clients was carried out to check if there was an explanation of why the clients were on the data sources if the client GP identified them as not diabetic.

40 clients (8.9%) on MCP were deemed not diabetic by the GP. Of the 40 clients, 4 had impaired glucose tolerance and 6 had gestational diabetes, no data was available on the remaining 30 clients (according to the GP) as to why they were included.

There were 22 clients (9.2%) on LTI that the GP indicated were not diabetic. 20 clients were identified by the GP as having gestational diabetes and required a LTI card during pregnancy, and 2 clients had impaired glucose tolerance. Ideally these clients should not have been on the master file but only the GP verification exercise would highlight this as the clients did have a LTI card for their condition.

There were 91 clients (42.5%) on PAS that the GP indicated were not diabetic. It was known that a number of these records related to clients who had gestational diabetes or IGT. Upon investigation with staff of the outpatient departments, it was apparent that clients with general medical and ophthalmic conditions were seen in the specialised diabetic and diabetic eye clinics. This exercise showed that these non diabetic clients should not have been included. However, these clients could only have been excluded by reviewing each medical record for the clients listed.

There were 7 clients (17.9%) on HIPE that the GP indicated were not diabetic. Their hospital medical records (charts) were examined and these clients were coded by the hospital as having diabetes. However, one client was considered to have impaired glucose tolerance by the GP and the other 6 clients were all deemed not diabetic by their GP.

67 clients (36.7%) on LIS were deemed not diabetic by the GP. Clinical reasons for patients to have a raised HbA_{1c} and not to be diabetic include; raised impaired glucose tolerance, gestational diabetes, metabolic syndrome, increased alcohol intake and obesity. 17 clients were deemed by the GP to have IGT and 1 client to have gestational diabetes, with multiple explanations, as outlined above, for the remaining 49 clients.

If the percentage of clients (24.1%) identified by the sample GPs as 'not diabetic per GP', 'deceased' and 'duplicates' are applied to the complete Mid West register this would reduce the number of clients on the register down to 10935. This would compare favourably with the estimated prevalence of just over 11000 diagnosed diabetics in the Mid West, assuming 9.8% undiagnosed diabetes (average North Ireland percentage).

3. Conclusions

The following are the main conclusions from the feasibility study.

- In summary:
  - Medical Card Prescriptions (MCP) and Long Term Illness (LTI) are good data sources for identifying diabetics
  - MCP and LTI are national data sources which can be accessed from one central location i.e. Primary Care Reimbursement Service (PCRS)
  - A national GP identifier was available for the majority of MCP clients and for a number of LTI clients
  - Hospital In-Patient Enquiry (HIPE), Patient Administration System (PAS) and Laboratory Information System (LIS) were not accurate data sources for identifying diabetes
  - PAS, LIS and HIPE can only be accessed locally (HIPE while a national system must be matched to relevant local PAS/HIS systems to capture full client demographic data)
  - PAS/HIS and LIS systems vary from site to site, resulting in capturing data from multiple systems which would not be feasible if creating a regional/national diabetes register
  - PAS/HIS and LIS each maintain a different set of GP codes. No GP identifier data is held on HIPE.

- The identification of duplicate clients was difficult in the absence of a common unique health identifier across the data sources. Where a unique identifier exists for each client, such as PPSN, then the exercise is relatively straightforward. A unique identifier, such as PPSN, was only available in a limited number of data sources and not for all clients in the data source. The issue of a unique health identifier may be resolved with the planned enactment of the Health Information Bill 2010.

- There was no uniform data structure across data sources. The lack of consistency significantly complicated the de-duplication exercise. For example, one data source had address stored in 3 address line fields and another data source had the address in 5 address line fields.

- There was variation in basic demographic information recorded across the data sources, e.g. different or incomplete address, different name and different date of birth.

- There was no consistency in the way GP demographic/identification data was stored on the data sources. A recognised national GP code (GMS GP code) was only available in MCP and LTI. Each PAS site in the Mid West has a different set of GP codes, resulting in each GP having a different code in each PAS site. No GP demographic/identification data is held on HIPE. The Mid West LIS maintains its own GP code and does not store any nationally recognised GP code. PAS/HIS and LIS applications around the country maintain their own local set of GP codes.

- Identification of a single occurrence of diabetic clients with GP demographic/identification data across multiple data sources was a very time consuming and labour intensive exercise, as shown in Phases 2 and 4. The complexity and level of effort increased with the introduction of multiple data sources.

- Fuzzy Logic was identified as an efficient method for creating a single occurrence of a client across data sources.

- The creation of a GP identifier from multiple data sources required time and resources. The quality of the GP demographic/identification data available on PAS, HIPE (from PAS) and LIS was poor and the lack of a nationally recognised GP identifier made processing of these GPs very slow. MCP and LTI, with the availability of a GMS GP code for clients (where GP was recorded), required very little effort to match.

- The GP visit verification process was both time consuming and labour intensive, taking an experienced nurse with good knowledge of all practice management systems approximately a day per practice to complete this task.

- The GP visit verification exercise represented a 7.8% sample (1126/14407) of the register population. The exercise highlighted the following with regard to the accuracy of the client records on the master file.

  - Validated diabetics within each individual data source ranged from:
    - **83.4% for MCP**
    - **78.2% for LTI**
    - 14.5% for PAS
    - 48.8% for HIPE
    - 49.7% for LIS

  - Clients identified as 'not diabetic per GP' within each individual data source ranged from:
    - **8.9% for MCP**
    - **9.2% for LTI**
    - 42.5% for PAS
    - 17.9% for HIPE
    - 36.7% for LIS

  - Clients identified as 'deceased' and 'duplicates' totalled in the individual data sources to:
    - 1.1% for MCP
    - 0.4% for LTI
    - 12.6% for PAS
    - 15.4% for HIPE
    - 2.7% for LIS

  - Clients identified as 'unknown' within each individual data source ranged from:
    - 6.6% for MCP
    - 12.2% for LTI
    - 30.4% for PAS
    - 17.9% for HIPE
    - 10.9% for LIS

  As the GP verification exercise only represented a 7.8% sample of the master file clients, the unknown clients could be clients of other GPs in the Mid West area.

- The GP exercise highlighted that MCP and LTI were the most accurate data sources for identifying true diabetics. PAS, HIPE and LIS were deemed to contain a significant number of clients who were not diabetic as the analysis in Table 2.23 showed.

- PAS, HIPE and LIS data sources represent poor value for the level of effort involved. This was particularly evident when the above validated diabetic statistics from the GP visit verification exercise were applied to the full register population, as presented in Table 3.1.

Table 3.1: Findings from GP visit verification exercise when applied to the full register population*

	MCP	LTI	PAS	HIPE	LIS	Total
Total Unique Records (following refinement phase)	**5847**	**3461**	**2659**	**691**	**1749**	**14407**
% of Unique Records	40.6%	24%	18.5%	4.8%	12.1%	100%
% of Validated Diabetics (following sample GP verification exercise)	83.4%	78.2%	14.5%	48.8%	49.7%	62.5%
Estimated Validated Diabetics	4876	2706	386	337	869	9174
% of Estimated Validated Diabetics (as a % of 9174)	53.1%	29.5%	4.2%	3.7%	9.5%	100%

*Given that the sample size was small it is difficult to draw assumptions on the validity of the full register

- MCP and LTI, which are national data sources, generate approximately 65% of clients across the data sources, of which 83.4% of clients on MCP and 78.2% of clients on LTI were validated diabetics (if GP exercise findings applied). This compares to PAS which generated 18.5% of clients on the register, of which only 14.5% were validated diabetics; HIPE generated 4.8% of clients on the register, of which 48.8% were validated diabetics; and LIS generated 12.1% of clients on the register, of which 49.7% were validated diabetics.

- MCP and LTI generate approximately 65% of clients across the data sources. When the GP verification exercise is undertaken, in order to identify the 'true' or validated diabetics, this figure would rise to approximately 83% of true diabetics identified through these data sources. When all GPs in an area are included in the validation exercise this figure will rise further, as the problem of patients being identified with the incorrect GP and therefore not 'validated' by that GP is removed. The exact proportion for this could not be estimated.

- Therefore, the study estimates that over 83% of true diabetics will be identified using the above recommended methodology. This figure will rise when all GPs are included in the process as stated above. Therefore, the recommended process will identify a very high proportion of diabetics who are currently identified on HSE ICT systems.

- In addition, GPs could identify diabetics who had not registered on the LTI scheme and were managed by diet alone, these together with newly diagnosed diabetics, not yet registered on either scheme, would also be captured for the register on an ongoing basis by the GP validation process.

- Verification of the complete register is only possible after all GPs and clients have been contacted for diabetic retinopathy screening. Any discrepancies would then be identified.

4. Recommendations

a) Based on the findings from the feasibility study Medical Card Prescriptions (MCP) and Long Term Illness (LTI) are the preferred data sources to form a register of known diabetics.

They are national data sources which can be accessed from one central location i.e. Primary Care Reimbursement Service (PCRS). Community drug schemes (MCP, LTI, Drugs Payment Scheme (DPS) etc) are in the process of being transferred to PCRS. Medical cards for over 70s were transferred on the 1st of January 2009 and the transfer for all schemes is planned to take place. In the future community drug schemes data will only be available directly from PCRS. It may be an option to include DPS as a data source as it will also be centrally held by PCRS. A review of DPS data (when available) would be required to assess its value in identifying any additional diabetic clients not already captured on MCP and LTI schemes.

b) It is recommended for the purpose of creating a diabetes register that the data is obtained directly from PCRS. Ideally one file with a single occurrence of each client should be obtained from PCRS to cover MCP, LTI and DPS (if to be included as a data source following assessment). Details on the proposed client extraction criteria are in Appendix 6. The final data extraction criteria should be agreed following discussions with PCRS. Nominated data custodians will receive the data source files.

c) In the event of PCRS not providing one file with a single occurrence of each client, individual files containing a single occurrence of a client for MCP, LTI and DPS should be provided by PCRS. A tool such as Fuzzy Logic could be used by the diabetic retinopathy screening programme administration office to remove duplicates across the three data sources.

d) A standard data source file layout is required. This will assist with matching of client records from PCRS and referral of clients from GPs. A sample data source file layout is in Appendix 7, this is a minimum extraction dataset for register creation and update, the final minimum extraction dataset will be agreed with PCRS.

e) A unique health identifier would greatly assist in the creation and maintenance of a diabetes register. In the absence of a unique health identifier a suitable alternative, such as PPSN, could be used. The planned Health Information Bill 2010 may assist in solving this problem.

f) Ongoing addition and updating of clients on the diabetes register should be automated as far as possible. At regular intervals, e.g. monthly, a file would be forwarded by PCRS to include new clients, clients whose demographic details have changed and clients who have deceased since the last export. DEPS should be used to identify further clients on the register who die.

g) It is recommended that the diabetic client list for each GP is produced from these two data sources and new registrations, and then verified by each GP to form the register. Practice nurse involvement in this is critical. It is recognised that initially some support to the GP practices will be required for the GP verification exercise. This may involve:
   - providing support from the diabetic retinopathy screening programme administration office
   - liaising with the GP practice and providing instructions to the GP practice nurse on how to identify/extract diabetic clients from the electronic practice management system
   - providing support via the primary care teams.

h) GP diabetic client lists should be verified annually by each GP. The GP returns the validated list (including new clients, deceased clients etc) to the diabetic retinopathy screening programme administration office.

i) Ideally, in order to accurately identify the practice diabetic clients it is essential that clinical data is coded by disease/medical condition on the electronic practice management system.

j) A national index for GPs would assist in matching patients to their GP.

k) The information received from General Practice for the formation and updating of the register is extremely important. The establishment of a structured care system for diabetics with their GPs, and the computerised returns of quality indicators to the National Diabetes Programme would greatly enhance both register accuracy and quality of care for people with diabetes.

l) Ideally the registration of newly diagnosed diabetics and changes to client demographics should be automated through the sending of structured messages from GP practice management systems to the diabetes register. This should be facilitated by the accredited GP practice management system vendors.

m) GPs should be encouraged to register new diabetic clients as soon as they are diagnosed.

n) It is recommended that a facility for self registration be put in place for people diagnosed with diabetes. Self registration should be publicised and organised through the diabetic retinopathy screening programme.

o) A discussion took place with the Data Protection Commissioners (DPC) Office concerning the formation of the diabetic retinopathy register as proposed by this report, and its use for diabetic retinopathy screening as outlined in the National 'Framework for the Development of a Diabetic Retinopathy Screening Programme for Ireland'[7]. The DPC advised that the proposed methodology for register formation and patient invitation to screening and consent was acceptable. The Commissioners preferred option was that patients would be invited for screening on behalf of their own GP in the first instance. Formal written consent for screening and data use will be obtained at their first appointment. The DPC recommended that the register be managed at regional level.

p) The resources recommended to create and maintain a diabetes register, in a HSE region, are estimated as between 1 and 2 whole time equivalents (WTEs), for a register size of approximately 30,000 people with diabetes. Initially a considerable amount of work will be required in validating with some GPs the initial diabetic client list. In carrying out the study, it was found that General Practice Systems and availability of diabetic lists vary considerably by practice. Some practices will require significant support in establishing the necessary systems and improving their ability to validate the diabetic client lists. As GP systems improve, and if systematic structured care systems are introduced in Primary Care, the need for a database support will diminish. Assuming that the combined data source file is received from PCRS, only a limited amount of IT support will be required. However, if files need to be merged etc. this will require more resources.

   The task involved will include:
   - creation of client database from extract file(s) received from PCRS
   - creation of a single GP file
   - generation of GP diabetic client lists for the GP verification exercise
   - liaison with GP practices
   - update of client records following feedback from GP verification exercise
   - ongoing register maintenance to include:
     - the upload of new clients, clients whose demographic details have changed and clients who have died
     - quality assurance of the register, to include de-duplication exercises.

5. References

1. International Diabetes Federation. (2009). *International Diabetes Federation Diabetes Atlas*, 4th Ed. Brussels: International Diabetes Federation.

2. Balanda, K. P., Barron, S., Fahy, L., McLaughlin, A. (2010). Making Chronic Condition Count: Hypertension, Stroke, Coronary Heart Disease, Diabetes. A systematic approach to estimating and forecasting population prevalence on the island of Ireland. Dublin: Institute of Public Health in Ireland.

3. Department of Health and Children. (2006). *Diabetes: Prevention and Model for Patient Care.* Dublin: Department of Health and Children.

4. Balanda, K. P., Fahy, L., Jordan, A., McArdle, E. (2006). *Making Diabetes Count. A systematic approach to estimating population prevalence on the island of Ireland in 2005.* Dublin: The Institute of Public Health in Ireland.

5. Mittman, R. (2004). *Using Clinical Information Technology in Chronic Disease Care: Expert Workshop Summary*. California: California HealthCare Foundation.

6. Health Service Executive Diabetes Expert Advisory Group. (2008). *First Report from Diabetes Expert Advisory Group*. Kildare: Health Service Executive.

7. Health Service Executive Diabetes Expert Advisory Group National Retinopathy Screening Committee. (2008). *Framework for the Development of a Diabetic Retinopathy Screening Programme for Ireland*. Kildare: Health Service Executive.

Appendix 1: Summary of Data Source Analysis

	MCP	LTI	PAS	HIPE	LIS	
Total Clients Extracted	7083	4306	7348	6171	14624	
Incomplete Records Removed	1009	219	0	0	17	
Duplicates	3	11	132	1783	2923	
Records after De-Duplication (Phase 2)	**6071**	**4076**	**7216**	**4388**	**11684**	
Duplicates Identified (Phase 4)						
MCP	-	570	2486	1934	3749	
LTI	-	-	1378	588	1845	
PAS	-	-	-	604	678	
HIPE	-	-	-	-	466	
LIS	-	-	-	-	-	
Total Duplicates Identified	**-**	**570**	**3864**	**3126**	**6738**	**14298**
Records Added to Master File by Source	**6071**	**3506**	**3352**	**1262**	**4946**	**19137**
Deceased	214	29	485	402	145	1275
Duplicate PPSN	0	2	41	2	8	53
Resident outside Mid West Region	10	14	167	167	371	729
HbA_{1c} Upper Part of Normal Range	-	-	-	-	2673	2673
Total Unique Records	**5847**	**3461**	**2659**	**691**	**1749**	**14407**

Appendix 2: Sample GP Details

	Location	Practice Type	Practice Nurse	Practice Management System	Notes
1	Limerick City	Single GP	Yes	Yes	Able to pull list from system easily. All patients coded and system contains all letters scanned onto computer. Very thorough notes, data very likely to be accurate.
2	Limerick City	Partnership	Yes	Yes	Above average notes, most patients coded and information inputted into correct fields. Data likely to be accurate.
3	Limerick City	Single GP	Yes	Yes	Above average notes, most patients coded and information inputted into correct fields. Data likely to be accurate.
4	Limerick City	Single GP	Yes	Yes	Able to pull list from system easily. All patients coded and system contains all letters scanned onto computer. Very thorough notes, data very likely to be accurate.
5	Limerick City	Single GP	Yes	Yes	Above average notes, most patients coded and information inputted into correct fields. Data likely to be accurate.
6	Limerick County	Partnership	Yes	Yes	Above average notes, most patients coded and information inputted into correct fields. Data likely to be accurate.
7	North Tipperary	Single GP	Yes	Yes	Computer system used mainly for prescription data with some data entered. No coding, paper list of diabetics.
8	North Tipperary	Single GP	Yes	Yes	Computer system used mainly for prescription data with some data entered. No coding, paper list of diabetics.
9	Clare	Single GP	No	Yes	Computer system used mainly for prescription data with some data entered. No coding, paper list of diabetics.
10	Clare	Partnership	Yes	Yes	Above average notes, most patients coded and information inputted into correct fields. Data likely to be accurate.
11	Clare	Partnership	Yes	Yes	Average notes, some patients coded and some information inputted into correct fields. Data likely to be accurate.
12	Clare	Partnership	Yes	Yes	Above average notes, most patients coded and information inputted into correct fields. Data likely to be accurate.
13	Clare	Partnership	Yes	Yes	Average notes, some patients coded and some information inputted into correct fields. Data likely to be accurate.
14	Limerick County	Single GP	No	No	Unable to gain access, GP returned list with corrections made. Unable to identify diabetic patients without delving through manual records.
15	Limerick City	Single GP	No	No	Unable to gain access
16	North Tipperary	Partnership	Yes	Yes	GP did not give consent

Appendix 3: Sample Diabetic Client List

GP Client List

Client Forename	Surname	Address	DOB	PPSN	Medical Card	Surname at Birth
GP: TEST SURNAME		TEST FORENAME TEST ADDRESS				
20588 MARY	TEST	TEST ADDRESS LINE 1 TEST ADDRESS LINE 2 TEST ADDRESS LINE 3	01/02/1940			

Appendix 4: Client Diabetic Register Form

HSE West Area, Limerick, Clare and North Tipperary
Client Diabetic Register Form

Client ID:

Patient Full Name:

Surname at Birth: Data Source:

Address:

Telephone No.: Mobile No.:

Hospital No.: PPSN: DOB:

Medical Card: Medical Card No.: LTI No.:

GP Name:

GP Address:

Ophthalmologist:

Diabetologist:

Insulin Commenced:

Complications:

Class of Diabetes:

Note: Classification is not determined by medication e.g. Insulin pt not necessarily Type 1 – See definitions.

Date of Diagnosis: Consent Form Signed:

Mobility: Ethnic Origin:

Outcome: Validated Diabetic Validated – Gestational Validated – IGT

Not Diabetic per GP Unknown RIP Duplicate New

Modified Reason:

Modified Notes:

Completed forms to be returned to: _____

Appendix 5: Sample GP Statistics

	Diabetic Client List	Validated Diabetic	Not Diabetic per GP	Unknown	Deceased	Duplicates	New	Total
1	182	111 (61.0%)	43 (23.6%)	16 (8.8%)	10 (5.5%)	2 (1.1%)	6	188
2	154	97 (63.0%)	21 (13.6%)	26 (16.9%)	9 (5.8%)	1 (0.6%)	13	167
3	132	97 (73.5%)	20 (15.2%)	11 (8.3%)	1 (0.8%)	3 (2.3%)	1	133
4	122	76 (62.3%)	28 (23.0%)	17 (13.9%)	0	1 (0.8%)	7	129
5	78	37 (47.4%)	18 (23.1%)	22 (28.2%)	1 (1.3%)	0	14	92
6	64	48 (75.0%)	10 (15.6%)	6 (9.4%)	0	0	3	67
7	88	59 (67.0%)	17 (19.3%)	10 (11.4%)	0	2 (2.3%)	3	91
8	89	42 (47.2%)	36 (40.4%)	6 (6.7%)	5 (5.6%)	0	0	89
9	58	39 (67.2%)	0	19 (32.8%)	0	0	6	64
10	2	2 (100%)	0	0	0	0	58	60
11	31	19 (61.3%)	8 (25.8%)	3 (9.7%)	1 (3.2%)	0	2	33
12	33	22 (66.7%)	7 (21.2%)	2 (6.1%)	1 (3.0%)	1 (3.0%)	0	33
13	27	18 (66.7%)	6 (22.2%)	3 (11.1%)	0	0	2	29
14	66	37 (56.1%)	13 (19.7%)	10 (15.2%)	5 (7.6%)	1 (1.5%)	0	66
Total	**1126**	**704 (62.5%)**	**227 (20.2%)**	**151 (13.4%)**	**33 (2.9%)**	**11 (1.0%)**	**115**	**1241**

Appendix 6: Proposed Client Extraction Criteria

The final data extraction criteria should be agreed following discussions with PCRS.

Register Creation

A single occurrence of a client to cover Medical Card Prescriptions (MCP), Long Term Illness (LTI) Scheme and Drugs Payment Scheme (DPS) (if used) for the data extraction period.

Clients to be extracted on the basis of:

- Clients prescribed diabetic drugs and/or blood glucose testing strips with ATC level codes
  - A10A: Insulins and analogues
  - A10B: Blood glucose lowering drugs, excluding insulins
  - V04CA91: Blood glucose test strips

- Clients with a LTI card diagnosed with diabetes mellitus

Register Update

At regular intervals, a file to include new clients (based on extraction as per register creation), clients whose demographic details have changed and clients who have deceased since the last extract.

Appendix 7: Sample Data Source File Layout

The data source file layout should at a minimum include the following data fields.

Data Field	Field Size
Forename	25
Surname	25
Address Line 1	50
Address Lines 2 to 5	30 each
Full Address	170
Local Health Office Area	2
Date of Birth	10 [dd-mm-yyyy]
PPSN	10
Unique Health Identifier	15
Unique IDs (any other unique identifier available for a client from the source system e.g. Medical Card Number, LTI Number, DPS Number)	15
Phone Number, if available	20
GMS GP Code	5
GP Forename	25
GP Surname	25
GP Address Line 1	50
GP Address Lines 2 to 5	30 each
GP Full Address	170
GP Phone Number, if available	20
Unique IDs (any other unique identifier available for a GP from the source system e.g. Medical Council Registration Number)	15

Glossary of Terms

A10A	Drugs used in Diabetes: Insulins and analogues
A10B	Drugs used in Diabetes: Blood glucose lowering drugs, excluding insulins
ATC	Anatomical Therapeutic Chemical Classification System
Careworks Schemes System	Software application used in the Mid West area (and a number of other HSE areas) to manage a range of community schemes and services
CCA	Community Care Area
DEPS	Death Event Publication Service This service is a product of the Inter-Agency Messaging Service (IAMS). The service transfers life event registrations, i.e. births and deaths, from the General Register Office (GRO) to the Department of Social and Family Affairs. Death events are validated for the accuracy of the given PPSN and then returned to GRO; additionally the validated death events are then made available to a Publication Service (DEPS) and also forwarded to the Central Statistics Office
DPS	Drugs Payment Scheme The scheme applies to Irish residents who do not have a medical card and normally have to pay the full cost of their medication. Under the scheme no individual or family will be required to pay more than a fixed amount in any calendar month for approved drugs, medicines and appliances for use by that person or his/her family in that month
EAG	Expert Advisory Group
ESRI	Economic and Social Research Institute
GMS	General Medical Services Persons who are unable without undue hardship to arrange general practitioner medical and surgical services for themselves and their dependants receive a free general medical service. Drugs, medicines and appliances supplied under the Scheme are provided through retail pharmacies. Generally the Doctor gives a completed prescription form to a person, who takes it to any pharmacy that has an agreement with the Health Service Executive to dispense GMS prescription forms
GP	General Practitioner
HbA_{1c}	Test that measures the amount of glycosolated haemoglobin in the blood. The test is used to assess the adequacy of blood glucose control over the preceding two to three months and is universally used to guide diabetes management. HbA_{1c} measurements are classified as either normal or elevated, based on whether it is below or above the upper limit of the reference ranges. The normal range for a HbA_{1c} in the Mid West laboratory information system is between 3.8% and 6.0%
HIPE	Hospital In-Patient Enquiry A computer based health information system designed to collect clinical and administrative data on discharges from, and deaths in, acute hospitals in Ireland

HIS	Hospital Information System A comprehensive integrated information system which contains administrative and clinical information on patients who avail of hospital services
HSE	Health Service Executive
LIS	Laboratory Information System A software system used in laboratories for the management of samples, lab users, instruments, standards and other lab functions
LTI	Long Term Illness Scheme Persons who suffer from one or more of a schedule of illnesses are entitled to obtain, without charge, irrespective of income, necessary drugs/medicines and/or appliances under the LTI scheme
MCP	Medical Card Prescriptions Drugs, medicines and appliances supplied under the GMS Scheme are provided through pharmacies. Generally the Doctor gives a completed prescription form to a person, who takes it to any pharmacy that has an agreement with the Health Service Executive to dispense GMS prescription forms
PAS	Patient Administration System A hospital system which contains administrative information on patients who avail of hospital services, including inpatient, day case and outpatient services
PCRS	Primary Care Reimbursement Service The HSE, Finance Shared Service, Primary Care Reimbursement Service (PCRS) supports the delivery of primary healthcare by providing reimbursement services to primary care contractors (GPs, Dentists, Pharmacists and other professionals) for the provision of health services to members of the public in their own community
PPSN	Personal Public Service Number
SQL	Structured Query Language
V04CA91	Diagnostic agents, test for Diabetes: Blood glucose test strips
WTE	Whole Time Equivalent